GRASP: Geometry-aware Residual Alignment for Scalable Pretraining Data Attribution

Anonymous ACL submission

Abstract

Scalable data attribution methods typically assign isolated utility scores to individual training examples. This prevalent additive assumption fundamentally fails to capture critical subset dynamics, including data redundancy and complementary coverage. In this work, we reframe attribution as subset-level counterfactual utility prediction and introduce GRASP, an interaction-aware surrogate. Grounded in a theoretical smoothness lower bound, GRASP explicitly models subset interactions through a quadratic geometric penalty. To achieve pretraining-scale efficiency without relying on hidden oracle tuning, we couple low-dimensional feature sketches with a strictly finite lower-confidence bound selection protocol. Extensive subset-retraining evaluations demonstrate that GRASP decisively outperforms existing scalable baselines. It more than doubles the task-level rank correlation for counterfactual subset fidelity while reducing upfront artifact construction costs by nearly an order of magnitude. Downstream diagnostics further show that this scoring mechanism transfers to language model curation and cross-domain vision selection, establishing a robust foundation for optimizing massive pretraining corpora.

1 Introduction

The capabilities and failure modes of large language models are fundamentally determined by their pretraining corpora (Raffel et al., 2020; Penedo et al., 2024). Consequently, data curation has become the central driver of model performance. This process encompasses extraction, filtering, deduplication, and domain mixing. Yet, these structural decisions typically rely on coarse heuristics or prohibitively expensive retraining trials. Data attribution bridges this methodological gap by connecting model behavior back to the data pipeline. This connection enables targeted auditing, contamination analysis, and principled data selection under fixed compute budgets.

Despite these promises, scaling attribution to guide practical curation reveals a fundamental flaw in current paradigms. Existing scalable methods, including influence functions (Koh and Liang, 2017), TracIn (Pruthi et al., 2020), TRAK (Park et al., 2023), DataInf (Kwon et al., 2024), and LESS (Xia et al., 2024), operate by assigning a scalar relevance score to each individual training example. However, practical data curation involves retaining or removing coherent data slices rather than isolated sentences. When practitioners use pointwise scores to evaluate a data subset, they implicitly treat the utility of that subset as a simple sum of individual contributions. This additive assumption is mathematically restrictive. In real-world corpora containing near-duplicates and overlapping domains, training examples interact. Selecting highly aligned but identical examples yields diminishing returns, whereas diverse examples provide complementary coverage.

While coalition-based valuations such as Data Shapley (Ghorbani and Zou, 2019) and datamodeling (Ilyas et al., 2022; Wang et al., 2025) formally capture these subset dependencies, exact coalition evaluation is computationally prohibitive for routine pretraining pipelines. This creates a critical methodological gap. Scalable attribution requires a reusable surrogate that explicitly models train-train interactions without requiring combinatorial retraining. Consequently, the evaluation of attribution methods must structurally shift from individual example ranking to *subset-level counterfactual utility prediction*. A rigorous attribution method must accurately rank candidate data interventions by their empirical utility after retraining.

To address this challenge, we propose GRASP, a geometry-aware residual-alignment surrogate for scalable pretraining data attribution. Our approach is grounded in a one-step smoothness lower bound on utility improvement for weighted subset interventions. This theoretical foundation naturally

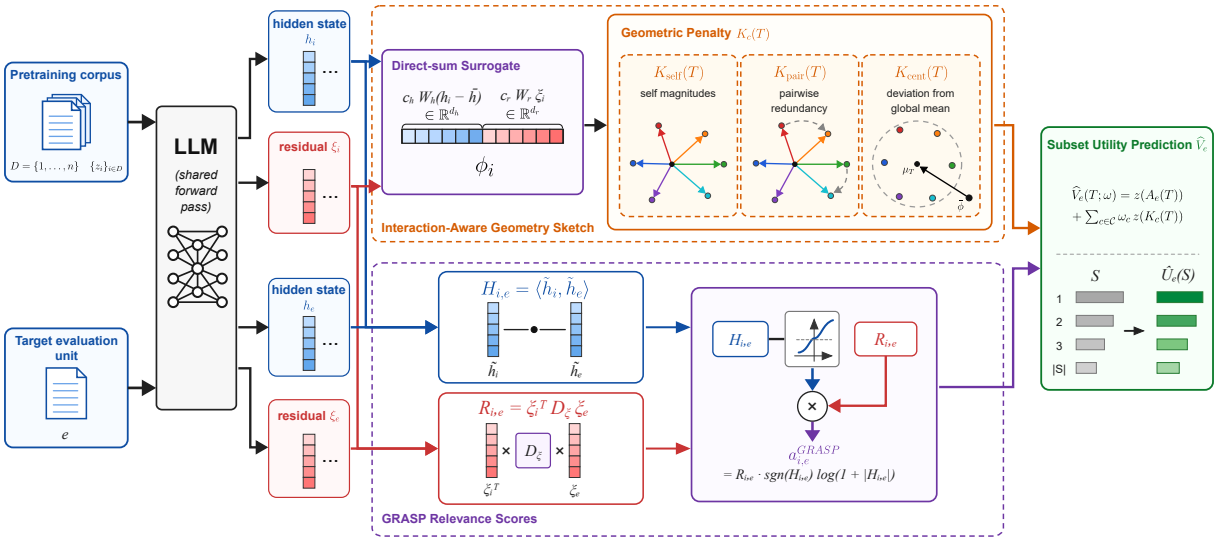

Figure 1: **Overview of the GRASP architecture.** GRASP predicts the counterfactual utility of data interventions by combining an additive relevance core (bottom) with an interaction-aware geometric penalty (top). By processing LLM hidden states and residuals into compact direct-sum sketches (ϕ_i) and stabilized relevance scores, the method efficiently captures subset redundancy and complementary coverage. The standardized components are then linearly aggregated to evaluate entire subsets without combinatorial retraining.

yields a linear target relevance term alongside a quadratic geometric penalty. Expanding this quadratic expression formally captures the mutual dependencies omitted by additive scoring. For pretraining-scale deployment, GRASP replaces exact high-dimensional update directions with low-dimensional feature sketches. For experimental discipline, all component weights and retain/delete modes are selected on development environments before test evaluation, separating scalable subset prediction from hidden per-task tuning.

In summary, our main contributions are:

- We formalize pretraining data attribution as *subset-level counterfactual utility prediction*, directly aligning the objective with practical corpus interventions to overcome the limitations of additive pointwise scoring.
- We derive GRASP, an interaction-aware surrogate grounded in a one-step smoothness lower bound. Its quadratic geometric penalty explicitly models data redundancy and complementary coverage without requiring combinatorial retraining.
- We instantiate GRASP with compact feature sketches for pretraining-scale efficiency. It reduces artifact construction time by nearly an order of magnitude, scores 100,000 subsets in 5 seconds, and more than doubles the task-level LDS rank correlation over existing scalable baselines.

2 Problem Setup

We study *subset-level counterfactual utility prediction*: given a pretraining corpus, an evaluation target, and candidate data interventions, the objective is to rank the interventions by their realized utility after retraining.

2.1 Pretraining Data Attribution

Let $\mathcal{D} = \{1, \ldots, n\}$ index a pretraining corpus $\{z_i\}_{i \in \mathcal{D}}$. A data intervention retains a subset $S \subseteq \mathcal{D}$, yielding parameters θ_S after retraining. For a target e, we define the utility as $\mathcal{U}_e(\theta) = -L_e(\theta)$, where L_e is the target loss. The true subset utility is therefore

$$\mathcal{U}_e(S) = \mathcal{U}_e(\theta_S) = -L_e(\theta_S). \tag{1}$$

Pretraining attribution aims to provide a scalable surrogate $\widehat{\mathcal{U}}_e : 2^{\mathcal{D}} \to \mathbb{R}$ for this set function $S \mapsto \mathcal{U}_e(S)$.

While traditional attribution assigns a scalar score $a_{i,e} \in \mathbb{R}$ to each example i (Koh and Liang, 2017; Park et al., 2023; Xia et al., 2024), a subset-level surrogate predicts the utility of the entire intervention S.

2.2 Counterfactual Fidelity via LDS

Following datamodeling (Ilyas et al., 2022; Park et al., 2023), we evaluate counterfactual fidelity using the Linear Datamodeling Score (LDS). Given a family of B sampled interventions $\mathcal{B} = \{S_b\}_{b=1}^{B}$,

LDS measures the rank correlation between predicted and realized utilities for target e:

$$\rho_e = \mathrm{Spearman}_b \left(\widehat{\mathcal{U}}_e(S_b), \mathcal{U}_e(S_b) \right). \quad (2)$$

When relying on pointwise scores, the subset predictor degrades to an additive baseline:

$$\widehat{\mathcal{U}}_e^{\mathrm{add}}(S) = \sum_{i \in S} a_{i,e}. \quad (3)$$

This modular form assumes independence among examples, rendering it brittle for corpora containing duplicates and domain clusters. We therefore seek a scalable surrogate incorporating a structural subset interaction term to improve ρ_e.

3 Method

GRASP predicts the utility of a candidate data intervention without retraining for every query. It first computes residual-alignment relevance scores for train–target pairs, then scores a whole subset by adding geometry-aware train-train interactions, and finally selects a fixed protocol on development environments.

Figure 1 summarizes this pipeline.

3.1 GRASP Relevance Scores

To evaluate the utility of a data intervention, GRASP first requires a scalable measure of alignment between training and target examples. Rather than computing exact but intractable influence functions ($H^{-1}\nabla\ell$), we derive an efficient proxy from the last-layer geometry of causal language models. Let h_i and $\xi_i = p_i - y_i$ denote the hidden state and next-token residual for a given token, where the local gradient admits an exact outer-product form $\nabla_W \ell_{\mathrm{tok}} = h\xi^\top$. For sequences containing multiple tokens, exact gradients require summing these outer products. To maintain linear storage complexity, we instead pool the sequence into a single summary representation and employ a rank-1 surrogate: $\nabla_W \ell_i \approx h_i \xi_i^\top$. While this simplification formally drops intra-sequence covariance, it effectively captures the macroscopic alignment of the sequence.

Under a Kronecker-factored generalized Gauss–Newton (GGN) approximation (Martens and Grosse, 2015), the inverse empirical Fisher or GGN matrix is approximated as a Kronecker product $F^{-1} \approx \Sigma_h^{-1} \otimes D_\xi$. Consequently, the inner product of preconditioned gradients elegantly factorizes into separate hidden and residual matches:

$$\langle \nabla_W \ell_i, F^{-1} \nabla_W \ell_e \rangle \approx \left(h_i^\top \Sigma_h^{-1} h_e \right) \left(\xi_i^\top D_\xi \xi_e \right). \quad (4)$$

Motivated by this decomposition, for a target evaluation unit e with representations h_e, ξ_e, we whiten the hidden states using global training statistics:

$$\tilde{h}_i = \Sigma_h^{-1/2}(h_i - \bar{h}), \qquad \tilde{h}_e = \Sigma_h^{-1/2}(h_e - \bar{h}), \quad (5)$$

where $\Sigma_h^{-1/2}$ incorporates Tikhonov regularization. The factorized similarities are then defined as:

$$H_{i,e} = \langle \tilde{h}_i, \tilde{h}_e \rangle, \quad (6)$$

$$R_{i,e} = \xi_i^\top D_\xi \xi_e. \quad (7)$$

Here, $D_\xi = \mathrm{diag}(\gamma + \epsilon)^{-\alpha}$ ($\alpha \in \{1, 1/2\}$) serves as a stabilized inverse-covariance preconditioner for the residuals, with γ estimated via a sparse top-k residual sketch to minimize the memory footprint.

Strict adherence to local influence dictates a direct bilinear product $H_{i,e}R_{i,e}$. In practice, however, hidden-state inner products exhibit extreme heavy-tailed distributions that routinely destabilize subset evaluation. To robustify the formulation, we explicitly depart from the exact Taylor expansion and introduce a monotonic dampening function:

$$a_{i,e}^{\mathrm{GRASP}} = R_{i,e} \, \mathrm{sgn}(H_{i,e}) \log(1 + |H_{i,e}|). \quad (8)$$

This odd, strictly increasing log-transformation preserves both the ordinal ranking and the directional support (sign) of the representation match, while safely truncating the disproportionate leverage of geometric outliers.

Finally, the first-order relevance of a weighted intervention subset T is computed additively:

$$A_e^{\mathrm{GRASP}}(T) = \sum_{i \in T} w_i a_{i,e}^{\mathrm{GRASP}}. \quad (9)$$

GRASP adopts $A_e^{\mathrm{GRASP}}(T)$ as its foundational linear predictor, which is subsequently augmented with geometry-aware interaction penalties to model subset redundancy.

3.2 Interaction-Aware Subset Utility

Let $g_e = \nabla_\theta L_e(\theta)$ be the target gradient at a reference checkpoint. As derived in Appendix A, implicit differentiation of a weighted training objective yields the influence-directed update $u_i = H^{-1}\nabla_\theta \ell_i(\theta)$ for example i, up to a shared scale. A

standard scalar attribution method scores a subset S additively via this first-order alignment:

$$\widehat{\mathcal{U}}_e^{\text{add}}(S) = \sum_{i \in S} a_{i,e}, \qquad a_{i,e} \approx \langle g_e, u_i \rangle. \quad (10)$$

However, this modular form assumes strict independence, fundamentally failing to capture redundancy or complementarity within a subset. To formally model subset dynamics, we must examine the geometry of the combined update, which naturally incurs a curvature penalty.

Theorem 1 (One-step interaction lower bound)**.** *Fix parameters θ and target e. Let $g_e = \nabla_\theta L_e(\theta)$, step size $\eta \geq 0$, and update direction $u_i \in \mathbb{R}^{\dim(\theta)}$. For a weighted set T with weights w_i, define the combined update $D_T = \sum_{i \in T} w_i u_i$. If L_e is β_e-smooth (with respect to the Euclidean norm) between θ and $\theta - \eta D_T$, the target utility improvement $\Delta_e(T) = \mathcal{U}_e(\theta - \eta D_T) - \mathcal{U}_e(\theta)$ satisfies:*

$$\Delta_e(T) \geq \eta \sum_{i \in T} w_i \langle g_e, u_i \rangle - \frac{\beta_e \eta^2}{2} \left\| \sum_{i \in T} w_i u_i \right\|^2. \quad (11)$$

The proof (Appendix A) bounds the counterfactual utility of full retraining. Crucially, expanding the quadratic penalty isolates the pairwise interaction:

$$\left\| \sum_{i \in T} w_i u_i \right\|^2 = \sum_{i \in T} w_i^2 \|u_i\|^2 + 2 \sum_{\substack{i<j \\ i,j \in T}} w_i w_j \langle u_i, u_j \rangle. \quad (12)$$

This reveals that mutually aligned examples ($\langle u_i, u_j \rangle > 0$) incur a strictly positive off-diagonal penalty. In terms of discrete curvature, the second difference of the set function satisfies $\Delta_i \Delta_j \widehat{\mathcal{U}}_e^{\text{int}} \propto -w_i w_j \langle u_i, u_j \rangle$. Thus, aligned directions exhibit diminishing marginal returns, while structurally opposing directions can be complementary (see Proposition 1).

While Theorem 1 dictates the necessity of the quadratic penalty, calculating the exact coefficient $\beta_e \eta^2 / 2$ is intractable for pretraining-scale interventions, as β_e and η fluctuate with subset norm and model scale. We therefore treat the theorem as a structural blueprint: a linear relevance term regularized by a non-negative geometric penalty. By replacing the intractable scalar with a tunable protocol parameter $\lambda_e \geq 0$, we arrive at the idealized interaction-aware surrogate:

$$\widehat{\mathcal{U}}_e^{\text{int}}(T) = \eta A_e^u(T) - \lambda_e K^u(T), \quad (13)$$

where $A_e^u(T) = \sum_{i \in T} w_i \langle g_e, u_i \rangle$ captures the additive utility, and $K^u(T) = \|D_T\|^2$. In the implemented predictor, λ_e is absorbed into a finite set of component weights calibrated on development environments.

3.3 Sketched Subset Predictor

Directly storing high-dimensional update directions $u_i \in \mathbb{R}^{\dim(\theta)}$ for every pretraining row is computationally intractable. Under the Kronecker-factored view, the true geometric interaction $\langle u_i, u_j \rangle$ takes the form of a product kernel between hidden-state and residual similarities. However, explicitly constructing the corresponding tensor-product feature $\psi_i = (W_h(h_i - \bar{h})) \otimes (W_r \xi_i)$ incurs a prohibitive $O(d_h d_r)$ memory footprint. To achieve linear scaling, GRASP deliberately trades exact multiplicative cross-terms for an additive direct-sum surrogate via feature concatenation:

$$\phi_i = [c_h W_h(h_i - \bar{h});\ c_r W_r \xi_i] \in \mathbb{R}^{d_h + d_r}. \quad (14)$$

Here, W_h is the centered whitening map from Equation 5, W_r is the residual scaling induced by the stabilized preconditioner D_ξ (or its sparse top-k approximation), and c_h, c_r are fixed scaling constants. While this direct-sum kernel intentionally deviates from an unbiased product kernel, it ensures fast $O(|T|d)$ subset sweeps and linear storage space. Appendix A formalizes the local approximation bound and resulting sketch error of this relaxation.

Using this compact sketch, we operationalize the blueprint of Theorem 1. The scalar relevance $A_e(T) = A_e^{\text{GRASP}}(T)$ is retrieved separately. For any candidate intervention set T, the raw geometric penalty is $K_{\text{raw}}(T) = \|\sum_{i \in T} w_i \phi_i\|^2$. To grant the predictor finer expressivity over self-regularization and cross-sample redundancy, we orthogonally decompose and center this geometry:

$$K_{\text{self}}(T) = \sum_{i \in T} w_i^2 \|\phi_i\|^2, \quad (15)$$

$$K_{\text{pair}}(T) = K_{\text{raw}}(T) - K_{\text{self}}(T), \quad (16)$$

$$K_{\text{cent}}(T) = \left\| \sum_{i \in T} w_i (\phi_i - \bar{\phi}) \right\|^2, \quad (17)$$

where $\bar{\phi} = n^{-1} \sum_{i \in \mathcal{D}} \phi_i$ is the global feature mean.

Benchmark	BM25		TracIn		TRAK		GRASP	
	ρ	Pos.	ρ	Pos.	ρ	Pos.	ρ	Pos.
BasicSkills Common Knowledge	0.062	0.520	0.114	0.656	0.059	0.551	**0.312**	**0.973**
BasicSkills Logical Reasoning	0.029	0.500	0.220	0.703	0.167	0.672	**0.457**	**0.980**
OpenBookQA	0.064	0.508	0.151	0.578	0.182	0.582	**0.350**	**0.855**
CommonsenseQA	-0.089	0.510	0.163	0.680	-0.062	0.495	**0.260**	**0.870**
SciQ	0.187	0.609	0.223	0.609	0.242	0.641	**0.424**	**0.906**
Average	0.051	0.529	0.174	0.645	0.118	0.588	**0.361**	**0.917**

Table 1: **Counterfactual subset-utility fidelity on full-dataset benchmarks.** GRASP consistently outperforms all baselines in predicting empirical retraining outcomes. We report the task-level Spearman rank correlation (ρ) and the fraction of evaluation units exhibiting positive LDS correlation (Pos.). Higher is better for both metrics. Best and second-best results are **bolded** and underlined, respectively.

Finally, constructing a unified surrogate requires reconciling the relative scales of the linear relevance and the decomposed quadratic penalties. As dictated by Theorem 1, the true interaction scale depends on a target-specific and subset-dependent smoothness constant β_e, which cannot be analytically computed. We bridge this theory-to-practice gap via empirical standardization:

$$\widehat{V}_e(T;\omega) = z(A_e(T)) + \sum_{c\in\mathcal{C}} \omega_c\, z(K_c(T)). \quad (18)$$

Here, $\mathcal{C}$ contains the selected geometric components (omitting zero-variance ones). The map $z(x) = (x-\mu)/\sigma$ standardizes the features using calibration moments μ, σ. Crucially, these moments are strictly anchored to a fixed calibration subset family within the environment—test interventions are never used to determine their own normalization. By mapping all targets into a standardized scale space, GRASP allows a single, dataset-agnostic weight vector ω (calibrated strictly on development environments; see Appendix B) to implicitly govern the theoretical λ_e interaction scale. Once scores and sketches are pre-computed, evaluating a subset costs $O(|T|d)$ time with only $O(nd)$ global storage.

4 Experiments

The core premise of subset-level attribution is counterfactual fidelity. In this section, we first evaluate this primary capability using rigorous subset-retraining protocols (Section 4.1). We then conduct ablation studies to isolate the mechanistic contributions of our interaction-aware components (Section 4.2). Finally, we demonstrate the operational utility of GRASP across downstream applications, including language model curation, cross-domain vision transfer, and semantic retrieval (Section 4.3). Setting-specific configurations for all experiments are detailed in Appendix B.

4.1 Counterfactual Subset-Utility Evaluation

Table 1 reports the Linear Datamodeling Score (LDS) evaluation, which directly tests counterfactual fidelity by measuring the rank correlation between predicted subset utilities and their empirical outcomes after retraining. To ensure a rigorous comparison, we strictly control the intervention candidate pool and the subset-retraining recipe across all methods. Our experimental testbed utilizes a 50M-parameter causal decoder-only Transformer (Vaswani et al., 2017; Radford et al., 2019) trained on a 10B-token pool. This pool is randomly sampled from a comprehensive Common Crawl corpus curated via standard scalable data pipelines (Raffel et al., 2020; Penedo et al., 2023, 2024; Li et al., 2024). Complete metric definitions and baseline details are provided in Appendix B.

The empirical results demonstrate that GRASP strictly dominates all baseline methods across both metrics on every evaluated benchmark. Notably, GRASP more than doubles the average task-level Spearman correlation compared to the strongest alternative, improving ρ from 0.174 (TracIn) to 0.361. Furthermore, it substantially elevates the positive-LDS fraction from 0.645 to 0.917, indicating highly reliable directional predictions across diverse intervention samples. This substantial performance gap provides the central evidence that explicitly capturing train-train interactions yields a strictly superior surrogate for ranking concrete pretraining interventions.

Practical pretraining attribution demands both cheap artifact construction and rapid subset evalua-

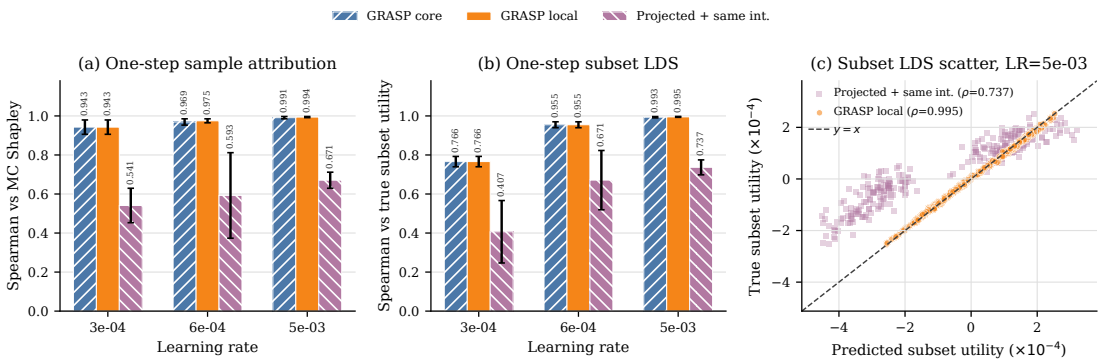

Figure 2: **One-step counterfactual fidelity on CC text.** The GRASP relevance tracks exact one-step sample and subset utilities, and the local interaction surrogate improves the more nonlinear high-learning-rate regime. The projected-gradient control receives the same interaction correction but remains weaker.

Method	Build (d)	State	Sweep (s)
TracIn	2.2	1	0.086
TRAK	2.2	1024	159.8
InRun-DS	10.0	1	0.086
Ret.-DS	12.2	0	1.05×10^6
GRASP	0.27	65	5.08

Table 2: **GRASP achieves efficient artifact construction and subset evaluation.** Build reports days, State reports cached floats per example, and Sweep reports seconds for 100,000 candidate subsets.

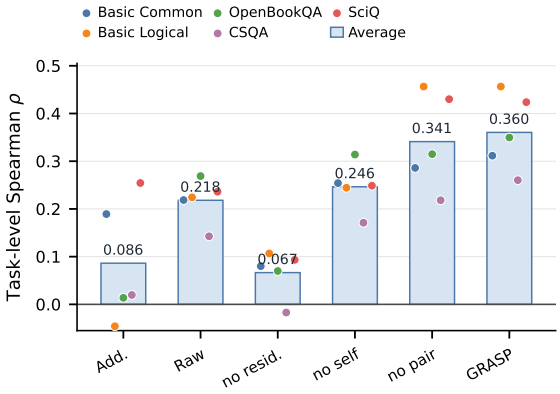

Figure 3: **Impact of interaction components on subset prediction fidelity.** We ablate GRASP components across all primary LDS benchmarks. Bars show mean task-level Spearman correlation (ρ), and colored points show individual tasks.

tion. Table 2 uncouples these costs, reporting upfront construction in days and cached 100k-subset sweeps in seconds. While scalar methods like TracIn enable fast sweeps once artifacts exist, faithful per-evaluation InRun Data Shapley requires an expensive trajectory replay, projecting to 10 days for a full SciQ run. In contrast, GRASP aggressively optimizes both phases: it requires only 0.27 days for artifact construction and executes a 100k-subset sweep in 5.08 seconds, approximately 31× faster than TRAK's 1024-dimensional feature aggregation.

4.2 Mechanistic Analysis and Ablations

To understand the source of these performance gains, we isolate the mechanistic contributions of individual GRASP components. Detailed one-step ground-truth construction and ablation protocols are given in Appendix B, with an additional cached protocol-selection diagnostic in Appendix Table 7.

Figure 2 shows that the GRASP relevance core tracks both Monte Carlo Shapley values and one-step subset utilities. Furthermore, the interaction surrogate substantially improves the high-learning-rate regime where subset updates are less linear. A projected-gradient control with an identical interaction correction remains significantly weaker, indicating that the performance gain does not stem from adding a generic pairwise correction to an arbitrary score.

The long-horizon LDS results exhibit the same pattern on complete subset retraining interventions. Relative to an additive-only predictor, incorporating the fixed interaction components raises the macro-average task-level Spearman from 0.086 to 0.360 and pairwise intervention accuracy from 0.528 to 0.623. Figure 3 further shows the dataset-level structure of this gain: residual alignment is the most critical component family, and removing it degrades every benchmark. Meanwhile, self terms provide a consistent secondary contribution, whereas hidden geometry exhibits more task-dependent behavior.

4.3 Downstream Applications and Transferability

Beyond baseline counterfactual fidelity, we evaluate how the interaction-aware scoring principle performs in practical data pipelines and entirely distinct domains.

Fixed-compute language model curation. To evaluate practical utility in pretraining pipelines, we deploy GRASP to select a high-quality training pool prior to model initialization. Table 3 reports completed held-out Common Crawl and SciQ evaluations under the same fixed-compute training schedule. GRASP achieves the lowest held-out CC loss and the best SciQ accuracy among the displayed full-budget diagnostic rows. We report early training-log snapshots and compact cross-architecture cleaning diagnostics separately in Appendix Tables 8 and 9.

Method	CC loss ↓	PPL ↓	SciQ acc. ↑
Random	2.9644	19.38	0.3203
InRun-1	2.9632	19.36	0.3086
InRun-2	2.9550	19.20	0.3047
GRASP	**2.9421**	**18.96**	**0.3398**

Table 3: **Held-out evaluation for fixed-compute CC-10B curation.** All rows use the same model, optimizer, selected-pool size, and training budget.

Method	AUC	20%	40%	60%	80%
Random	53.92	50.9	63.2	70.0	74.5
KNN	73.77	34.1	48.7	64.7	74.3
Infl.	75.59	45.8	64.4	72.2	**76.9**
TRAK	55.38	29.9	43.4	52.8	65.0
DataM.	54.34	47.6	64.1	71.4	74.1
InRun-1	80.28	55.6	68.9	71.9	75.5
InRun-2	80.04	52.8	66.4	**73.4**	75.9
GRASP	**80.32**	**56.1**	**69.6**	71.8	76.1

Table 4: **Cross-domain transferability on vision data selection.** AUC evaluates noisy-label detection; remaining columns report selected-data accuracy at each retention budget. KNN: KNN-Shapley, Infl.: Influence, DataM.: Datamodels. All entries are percentages; higher is better.

Cross-domain transfer: Vision data selection. To verify whether this scoring principle generalizes beyond language modeling, we evaluate our formulation in the standard Data Shapley vision setting (Ghorbani and Zou, 2019): ImageNet-pretrained ResNet18 (Deng et al., 2009; He et al., 2016), CIFAR-10 (Krizhevsky, 2009) 1k subset, and 10% label noise. Table 4 shows that GRASP achieves the best AUROC and strong selected-data accuracy across budgets.

Source retrieval and semantic diagnostics. Finally, we examine the semantic properties of the attribution signal through the WebOrganizer analysis (Wettig et al., 2025) in Appendix C. It evaluates whether the learned attribution signal ranks domain-labeled source rows in ways that complement lexical retrieval and whether its signed category-level mass is interpretable across checkpoints. We also include a safety-oriented misleading-information detection diagnostic in Appendix Table 10.

5 Discussion

Retained	BM25 ρ	Add. ρ	GRASP ρ
0.20	-0.153	-0.292	**0.516**
0.50	0.297	0.285	**0.503**
0.80	0.230	0.165	**0.490**
0.90	0.187	0.254	**0.447**
0.95	-0.688	0.065	**0.279**
0.99	-0.075	**0.210**	**0.210**

Table 5: **Interaction effects are regime-dependent.** SciQ retention-ratio diagnostic. BM25 is a lexical baseline; Add. is the relevance-only core of GRASP. All method columns report task-level Spearman ρ. The 0.95 row is a deadline partial stress screen.

When do subset interactions matter most? The benefit of interaction-aware attribution depends on the intervention regime. When a candidate subset differs nontrivially from the reference corpus, ranking quality is no longer determined by isolated relevance alone: redundant examples can saturate the same update direction, while complementary examples can jointly improve coverage. Table 5 summarizes this effect in the SciQ LDS environment. Across substantial and moderate retained fractions, the geometric surrogate improves task-level rank correlation over the additive core. In contrast, at 0.99 retention, candidate subsets differ by only a tiny omitted slice; the problem approaches a first-order additive limit, and GRASP correctly collapses to the additive core rather than amplifying unstable interaction noise. This behavior is consistent with the theoretical role of the quadratic term: it is most useful when subset geometry can change

the relative ordering of interventions, and least useful when all interventions are near-identical perturbations of the full corpus.

Semantic interpretability of the geometric surrogate. GRASP is designed as a counterfactual predictor, but its geometry also produces interpretable corpus diagnostics. The WebOrganizer analysis in Appendix C shows that attribution and lexical retrieval capture complementary structure. On topic labels, GRASP alone is competitive with BM25, while combining GRASP with BM25 yields the strongest normalized Recall@1000; the same complementarity appears for format labels when GRASP is combined with query overlap (Table 6). This suggests that the learned attribution signal is not merely a lexical overlap score. The checkpoint-wise category visualization (Figure 4) further shows signed source preferences: knowledge-bearing article-like regions receive positive average mass for the SciQ target, whereas boilerplate-heavy formats such as notices, support pages, and FAQ-style text fall on the negative side. We therefore view this as evidence that the geometric surrogate can support corpus auditing in addition to subset ranking, while keeping them separate from the main LDS fidelity claims.

6 Related Work

Our work is situated at the intersection of data attribution, scalable influence estimation, subset counterfactual modeling, and pretraining data auditing. We shift the core objective from assigning isolated scores to individual examples toward accurately predicting the empirical effects of concrete data interventions.

Data attribution and data valuation. Training data attribution investigates how predictions depend on specific examples. Methods like influence functions (Koh and Liang, 2017), representer points (Yeh et al., 2018), and RelatIF (Barshan et al., 2020) estimate local training contributions. Alternatively, Shapley-value methods formalize data value through marginal contributions across training coalitions (Ghorbani and Zou, 2019). While interaction-aware, Shapley aggressively averages these coalition dynamics into a static scalar per example, and exact estimation is computationally prohibitive. In-Run Data Shapley mitigates this cost along a single training trajectory (Wang et al., 2025). In contrast, we explicitly retain reusable train-train geometry to predict the utility of concrete subset interventions.

Scalable gradient-based attribution. Recent advancements make gradient attribution viable for large-scale models. TracIn integrates gradient inner products across training checkpoints (Pruthi et al., 2020). Other methods scale Hessian-based influence (Schioppa et al., 2022), analyze LLM generalization (Grosse et al., 2023), derive efficient estimators for tuned models (Kwon et al., 2024), accelerate attribution via projected features (Park et al., 2023), and adapt gradient similarity for instruction tuning (Xia et al., 2024). Closest to our domain, Chang et al. (2025) scale gradient-based influence to LLM pretraining. While these methods successfully deliver scalable first-order signals, we focus exclusively on subset-level counterfactual fidelity rather than retrieving isolated influential examples.

Pretraining data auditing and selection. Extensive corpus analyses reveal substantial duplication, toxic content, and benchmark contamination in pretraining datasets (Elazar et al., 2024). Existing data selection techniques utilize gradient-derived relevance (Xia et al., 2024) or pretraining-scale tracing (Chang et al., 2025) to surface beneficial examples. GRASP addresses a distinct structural challenge: predicting the relative downstream utility of candidate pretraining subsets. In this setting, redundancy, coverage, and subset interactions are first-class concerns, rendering purely lexical retrieval or independent example scoring mathematically insufficient.

7 Conclusion

This work reframes pretraining data attribution from isolated example scoring to subset-level counterfactual utility prediction. GRASP turns this view into a scalable surrogate by combining a smoothness-motivated geometric penalty, low-dimensional sketches, and development-only protocol selection. Across subset-retraining benchmarks, the method improves counterfactual ranking fidelity over scalable baselines, while the application diagnostics show that the same interaction-aware signal remains useful beyond the core LDS setting. These results suggest that subset geometry is a practical ingredient for auditing and optimizing massive pretraining corpora.

Limitations

Our theoretical formulation relies on a one-step smoothness lower bound. This is an approximation rather than a strict mathematical guarantee for full non-convex Transformer pretraining. Consequently, we rely on subset-retraining LDS as an empirical counterfactual test rather than claiming an exact analytical match.

Empirically, our validation centers on a 50M-parameter model and a 10B-token Common Crawl pool. While this scale enables strictly controlled interventions, it does not automatically guarantee identical behavior for frontier-scale models or radically different corpus mixtures. Scaling these findings remains an important empirical direction.

Furthermore, GRASP utilizes low-dimensional feature sketches to ensure computational scalability. This compression inevitably discards fine-grained attention interactions and long-range training dependencies, rendering the method a scalable surrogate rather than a lossless influence calculation.

Finally, our lower-confidence bound protocols are explicitly target-conditioned. We do not claim that GRASP provides a single universal data-quality score for all tasks. Additionally, extreme data interventions with high-overlap deletion ratios may necessitate specialized ratio-aware scaling adjustments.

Ethical Considerations

Data attribution can support responsible corpus auditing by identifying data regions that are influential for target behaviors, harmful outputs, or benchmark-like evaluations. At the same time, attribution scores are approximations and should not be treated as definitive evidence that a single document caused a model behavior. We recommend reporting uncertainty, negative controls, and failure modes whenever attribution is used for data governance.

Pretraining corpora may contain copyrighted, private, toxic, or benchmark- contaminating content. Releasing raw influential examples or high-resolution attribution tables can expose sensitive text or facilitate reconstruction of corpus membership. Any release of data examples, scores, or source-retrieval outputs should follow the underlying dataset licenses and privacy constraints.

References

Dimitris Achlioptas. 2003. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687. Special Issue on PODS 2001.

Elnaz Barshan, Marc-Etienne Brunet, and Gintare Karolina Dziugaite. 2020. Relatif: Identifying explanatory training samples via relative influence. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1899–1909. PMLR.

Tyler Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. 2025. Scalable influence and fact tracing for large language model pretraining. In *International Conference on Learning Representations*, volume 2025, pages 40976–40997.

Sanjoy Dasgupta and Anupam Gupta. 2003. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah Smith, and Jesse Dodge. 2024. What's in my big data? In *International Conference on Learning Representations*, volume 2024, pages 7735–7790.

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.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

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.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. 2025. OLMES: A standard for language model evaluations. In *Findings of the Association for Computational*

Linguistics: NAACL 2025, pages 5020–5048, Albuquerque, New Mexico. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

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.

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.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Y Zou. 2024. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. In *International Conference on Learning Representations*, volume 2024, pages 21921–21942.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024. Datacomp-lm: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 14200–14282. Curran Associates, Inc.

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.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. 1978. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294.

Team Olmo. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. Trak: Attributing model behavior at scale. *Preprint*, arXiv:2303.14186.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pages 30811–30849. Curran Associates, Inc.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, volume 36, pages 79155–79172. Curran Associates, Inc.

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, pages 19920–19930. Curran Associates, Inc.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. OpenAI technical report.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. 2022. Scaling up influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8179–8186.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jiachen (Tianhao) Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. 2025. Data shapley in one training

run. In *International Conference on Learning Representations*, volume 2025, pages 12358–12395.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pre-training data curation. *Preprint*, arXiv:2502.10341.

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*, ICML'24. JMLR.org.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. 2018. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

A Method Details

This appendix provides the formal derivations supporting the interaction-aware surrogate utilized for the counterfactual evaluation in Section 4. Our primary objective is to delineate which components of our formulation are direct mathematical consequences of objective smoothness and which originate from the approximation conditions imposed by the low-dimensional sketch.

Local influence direction. The local update direction u_i introduced in the main text is formally instantiated through classical influence function analysis. Consider the weighted regularized empirical risk

$$R(\theta, \alpha) = \frac{1}{n} \sum_{i=1}^{n} \alpha_i \ell_i(\theta) + \lambda_{\text{reg}} r(\theta), \tag{19}$$

where α_i represents the weight assigned to training example i. Assume R is twice continuously differentiable. Let $\theta^\star(\alpha)$ be a local minimizer such that $\nabla_\theta R(\theta^\star(\alpha), \alpha) = 0$. If the Hessian $H = \nabla_\theta^2 R(\theta^\star(\alpha), \alpha)$ is strictly positive definite, the Implicit Function Theorem guarantees the existence of a differentiable solution branch. Differentiating the stationarity condition with respect to α_i yields $\partial\theta^\star(\alpha)/\partial\alpha_i = -n^{-1} H^{-1} \nabla_\theta \ell_i(\theta^\star(\alpha))$.

Recall that the evaluation utility is defined as $\mathcal{U}_e(\theta) = -L_e(\theta)$. Applying the chain rule evaluated at the reference checkpoint θ gives the marginal utility change:

$$\frac{\partial \mathcal{U}_e}{\partial \alpha_i} = \frac{1}{n} \left\langle \nabla_\theta L_e(\theta), H^{-1} \nabla_\theta \ell_i(\theta) \right\rangle. \tag{20}$$

This derivation motivates defining the local influence-style update direction as $u_i = H^{-1} \nabla_\theta \ell_i(\theta)$, absorbing the shared $1/n$ scaling factor into the global learning rate.

Smoothness lower bound. Building upon this local linear approximation, we derive a certified lower bound for subset utility improvements. Assume the target loss function L_e is β_e-smooth. By definition, for any parameter update Δ, the following inequality holds:

$$L_e(\theta + \Delta) \leq L_e(\theta) + \langle \nabla_\theta L_e(\theta), \Delta \rangle + \frac{\beta_e}{2} \|\Delta\|^2. \tag{21}$$

We model the subset intervention as an aggregated gradient step defined by $\Delta = -\eta D_T$, where $\eta \geq 0$ is the step size. Substituting this into the smoothness condition and utilizing the utility definition $U_e = -L_e$ yields:

$$\begin{aligned} U_e(\theta - \eta D_T) - U_e(\theta) \geq \eta \langle \nabla_\theta L_e(\theta), D_T \rangle \\ - \frac{\beta_e \eta^2}{2} \|D_T\|^2. \end{aligned} \tag{22}$$

Finally, substituting the subset aggregate direction $D_T = \sum_{i \in T} w_i u_i$ directly recovers the one-step interaction lower bound presented in Theorem 1. Crucially, this bound relies solely on the β_e-smoothness of the target objective and does not require global convexity.

Set-function curvature and submodularity. The idealized surrogate defined in Equation 13 offers a precise set-theoretic interpretation of subset "interaction." Let $F_e(T) = \widehat{\mathcal{U}}_e^{\text{int}}(T)$. For any elements $i, j \notin T$ with $i \neq j$, the discrete second difference characterizes the marginal interaction:

$$\begin{aligned} \Delta_i \Delta_j F_e(T) = F_e(T \cup \{i, j\}) - F_e(T \cup \{i\}) \\ - F_e(T \cup \{j\}) + F_e(T). \end{aligned} \tag{23}$$

Direct algebraic expansion yields:

$$\Delta_i \Delta_j F_e(T) = -2\lambda_e w_i w_j \langle u_i, u_j \rangle. \tag{24}$$

This establishes a critical geometric intuition: aligned update directions naturally produce negative discrete curvature (redundancy penalty), whereas opposing directions can yield positive curvature (complementary coverage). Consequently,

GRASP functions as a signed pairwise-interaction set function. In the specific case where $w_i \geq 0$ and all pairs exhibit non-negative inner products ($\langle u_i, u_j \rangle \geq 0$), the surrogate becomes globally submodular, reflecting classical diminishing returns (Nemhauser et al., 1978). This conditional submodularity is formalized below.

Proposition 1 (Local submodularity over aligned clusters)**.** *Consider the unweighted discrete surrogate defined by*

$$F(T) = \eta \sum_{i \in T} a_i - \lambda \left\| \sum_{i \in T} u_i \right\|^2, \qquad \lambda > 0.$$

Let $\mathcal{D}_{\text{align}}$ be a candidate pool such that the pairwise alignment $\langle u_i, u_j \rangle \geq 0$ holds for all distinct $i, j \in \mathcal{D}_{\text{align}}$. Then, the set function F is submodular on $\mathcal{D}_{\text{align}}$. Moreover, for any $S \subseteq T \subseteq \mathcal{D}_{\text{align}}$ and $k \in \mathcal{D}_{\text{align}} \backslash T$, the marginal diminishing return is strictly positive whenever $\sum_{i \in T \backslash S} \langle u_k, u_i \rangle > 0$.

Proof. Let $\Delta_k F(S) = F(S \cup \{k\}) - F(S)$ denote the marginal gain of adding element k to set S. The linear relevance term contributes a constant ηa_k to every marginal gain. Expanding the quadratic penalty term yields:

$$\Delta_k F(S) = \eta a_k - \lambda \|u_k\|^2 - 2\lambda \sum_{i \in S} \langle u_k, u_i \rangle.$$

To verify submodularity, we examine the difference in marginal gains for $S \subseteq T$:

$$\Delta_k F(S) - \Delta_k F(T) = 2\lambda \sum_{i \in T \backslash S} \langle u_k, u_i \rangle.$$

Under the alignment assumption $\langle u_k, u_i \rangle \geq 0$, this difference is strictly non-negative since $\lambda > 0$. This guarantees $\Delta_k F(S) \geq \Delta_k F(T)$, satisfying the formal definition of diminishing returns and proving that F is submodular on $\mathcal{D}_{\text{align}}$. The strict inequality condition follows directly from the same expression. □

Sketched surrogate and error bounds. To operationalize this objective at scale, we introduce the sketched surrogate for a fixed nonnegative scale λ:

$$\widehat{B}_e^{\text{sk}}(T; \lambda) = \eta A_e(T) - \lambda K_{\text{raw}}(T). \tag{25}$$

Let $B_e^u(T; \lambda) = \eta A_e^u(T) - \lambda K^u(T)$ denote the exact local surrogate. If the sketch formulation satisfies bounded deviations on a subset family $\mathcal{T}$ such that

$$|A_e(T) - A_e^u(T)| \leq \epsilon_A(T), \tag{26}$$
$$|K_{\text{raw}}(T) - K^u(T)| \leq \epsilon_K(T), \tag{27}$$

then applying the triangle inequality guarantees that for every $\lambda \geq 0$ and $T \in \mathcal{T}$, the subset prediction error is bounded by:

$$\left| B_e^u(T; \lambda) - \widehat{B}_e^{\text{sk}}(T; \lambda) \right| \leq \eta \epsilon_A(T) + \lambda \epsilon_K(T). \tag{28}$$

In scenarios where only pointwise relevance bounds are accessible, such that $|a_{i,e} - \langle g_e, u_i \rangle| \leq \epsilon_a$, we can establish $\epsilon_A(T) = \epsilon_a \sum_{i \in T} |w_i|$. We emphasize that this constitutes a conditional approximation bound. The verification of a tightly bounded sketch error remains an empirical diagnostic rather than a structural guarantee inherent to the architecture.

Proposition 2 (Local direct-sum approximation to product geometry)**.** *Let x_i, x_j, r_i, r_j denote normalized hidden and residual sketches satisfying $\|x_i\| = \|x_j\| = \|r_i\| = \|r_j\| = 1$. Define the squared Euclidean distances as $d_h^2(i,j) = \|x_i - x_j\|^2$ and $d_r^2(i,j) = \|r_i - r_j\|^2$. Consider the product and direct-sum kernels:*

$$k_{\text{prod}}(i,j) = \langle x_i, x_j \rangle \langle r_i, r_j \rangle, \tag{29}$$
$$k_{\text{dir}}(i,j) = \langle x_i, x_j \rangle + \langle r_i, r_j \rangle. \tag{30}$$

The product kernel expands as:

$$k_{\text{prod}}(i,j) = 1 - \frac{1}{2}\left(d_h^2(i,j) + d_r^2(i,j)\right) + \frac{1}{4} d_h^2(i,j) d_r^2(i,j). \tag{31}$$

Furthermore, the direct-sum kernel satisfies:

$$k_{\text{dir}}(i,j) = k_{\text{prod}}(i,j) + 1 - \frac{1}{4} d_h^2(i,j) d_r^2(i,j). \tag{32}$$

Consequently, after omitting the constant offset, the direct-sum kernel and the product kernel agree strictly through the first order in the local squared distances. Their remaining discrepancy is confined to the fourth-order cross term.

Proof. For any unit vectors, the inner product relates to Euclidean distance via $\langle x_i, x_j \rangle = 1 - \|x_i - x_j\|^2/2$. Applying this identity identically to the residual sketches, substituting both into the definition of k_{prod}, and expanding the product yields

the first result. Summing the two inner products directly yields $k_{\text{dir}} = 2 - \frac{1}{2}(d_h^2(i,j) + d_r^2(i,j))$. Subtracting the expanded form of $k_{\text{prod}}(i,j)$ from this sum produces the second result. □

We stress that this proposition strictly characterizes a local approximation and does not imply a global equivalence between the direct-sum and product kernels. The constant offset is mathematically rank-invariant for fixed-size unweighted subset interventions. Furthermore, this discrepancy is structurally mitigated by centered components such as K_{cent}. Outside these specific configurations, the direct-sum kernel inherently operates as a scalable empirical surrogate.

Centered component interpretation. Let $W_T = \sum_{i \in T} w_i$ denote the total subset weight. When $W_T \neq 0$, we define the weighted subset mean as $\mu_T = W_T^{-1} \sum_{i \in T} w_i \phi_i$. The centered geometric component can then be strictly formulated as:

$$K_{\text{cent}}(T) = W_T^2 \|\mu_T - \bar{\phi}\|^2. \tag{33}$$

In the strictly non-negative regime where $w_i \geq 0$ and $W_T > 0$, the normalized term $K_{\text{cent}}(T)/W_T^2$ exactly corresponds to the squared Maximum Mean Discrepancy (MMD) under a linear kernel (Gretton et al., 2012). This MMD metric captures the discrepancy between the weighted empirical distribution on subset T and the unweighted full-pool empirical distribution. For arbitrary signed weights, Equation 33 retains its structural interpretation as a squared distance under signed measures but must not be strictly interpreted as a valid probability distance.

A projection sufficient condition for sketch error. The conditional sketch-error bound established in Equation 28 admits a standard sufficient condition in the idealized scenario where both target relevance and training geometry undergo an identical random projection. Let $x_T = \sum_{i \in T} w_i u_i$. Suppose the sketches are generated via a linear map Π such that $\phi_i = \Pi u_i$ and $q_e = \Pi g_e$. Consider a finite subset family $\mathcal{T}$ and construct the expanded finite vector set:

$$\mathcal{X}_e = \{x_T : T \in \mathcal{T}\} \cup \{g_e + x_T : T \in \mathcal{T}\} \cup \{g_e - x_T : T \in \mathcal{T}\}. \tag{34}$$

Assume Π preserves all squared Euclidean norms within $\mathcal{X}_e$ up to a relative error $\varepsilon \in (0,1)$. Then, for every $T \in \mathcal{T}$, the norm preservation directly guarantees:

$$\left| \|\Pi x_T\|^2 - \|x_T\|^2 \right| \leq \varepsilon \|x_T\|^2. \tag{35}$$

Furthermore, applying the standard polarization identity, $\langle a, b \rangle = (\|a+b\|^2 - \|a-b\|^2)/4$, guarantees that the inner product error is bounded by:

$$|\langle \Pi g_e, \Pi x_T \rangle - \langle g_e, x_T \rangle| \leq \frac{\varepsilon}{2} \left(\|g_e\|^2 + \|x_T\|^2 \right). \tag{36}$$

Classical Johnson-Lindenstrauss embeddings provide exactly this finite-set norm preservation (Dasgupta and Gupta, 2003; Achlioptas, 2003). Such embeddings require a sketch dimension scaling as $O(\varepsilon^{-2} \log(|\mathcal{X}_e|/\delta))$ to achieve a failure probability bounded by δ. Under this idealized projection condition, one can rigorously set $\epsilon_K(T) = \varepsilon \|x_T\|^2$ and $\epsilon_A(T) = \varepsilon(\|g_e\|^2 + \|x_T\|^2)/2$ within Equation 28. We explicitly note that this constitutes a theoretical sufficient condition. In practice, our hidden and residual feature sketches are evaluated empirically via the LDS diagnostic rather than assumed to perfectly satisfy this idealized randomized projection criterion.

Retained and omitted parameterizations. For strictly additive models, retained-set scores equal a global corpus constant minus the corresponding omitted-set scores. This linear relationship ensures that the omitted representation remains rank-invariant up to a sign reversal. However, for quadratic components, the retained and omitted parameterizations structurally diverge because full-pool cross terms do not trivially cancel. Furthermore, the subset omission fraction inherently rescales the expected magnitude of these geometric interaction terms. Suppose each training example is omitted independently with probability q. Let $M_i \in \{0,1\}$ denote the binary omission indicator and w_i represent the deterministic weight. The expected squared norm of the omitted geometry evaluates to:

$$\mathbb{E}\left[\left\| \sum_i M_i w_i \phi_i \right\|^2 \right] = q \sum_i w_i^2 \|\phi_i\|^2 + q^2 \sum_{i \neq j} w_i w_j \langle \phi_i, \phi_j \rangle. \tag{37}$$

Consequently, under a uniform omission rate q, the expected self-interaction scales linearly with q, whereas the pairwise cross-interaction decays quadratically with q^2.

Ranking stability of subset predictions. Let $\mathcal{B} = \{S_b\}_{b=1}^{B}$ define the evaluation family of LDS subsets for a specific target or task-level aggregate. Denote the true realized utility as $U_b = \mathcal{U}_e(S_b)$ and the predicted surrogate utility as $\widehat{U}_b = U_b + \varepsilon_b$, where ε_b represents the approximation error. For any sampled subset pair (b, c), the predicted ranking correctly recovers the true empirical ranking whenever the error difference is strictly bounded by the true utility gap:

$$|\varepsilon_b - \varepsilon_c| < |U_b - U_c|. \tag{38}$$

This margin condition mathematically guarantees correct pairwise ordering, as it implies $(\widehat{U}_b - \widehat{U}_c)(U_b - U_c) > 0$. Consequently, assuming a uniform error bound $\|\varepsilon\|_\infty \leq \varepsilon_{\max}$, the pairwise ranking accuracy is bounded below by:

$$\mathrm{PairAcc} \geq \frac{|\mathcal{P}_{2\varepsilon_{\max}}|}{\binom{B}{2}}, \tag{39}$$

Here, $\mathcal{P}_{2\varepsilon_{\max}}$ contains the subset pairs (b, c) with $b < c$ and realized utility gap $|U_b - U_c| > 2\varepsilon_{\max}$. This formulation does not provide a tight theoretical bound on the continuous Spearman correlation coefficient, but it formalizes a critical gap condition inherent to rank-based counterfactual metrics. Specifically, surrogate approximation errors degrade empirical fidelity only when their magnitude exceeds the intrinsic retraining utility gaps between competing data interventions.

B Experimental and Implementation Details

LDS benchmark. The full-dataset LDS experiments use a 50M-parameter causal decoder-only Transformer (Vaswani et al., 2017; Radford et al., 2019). The pretraining pool contains 10B tokens randomly sampled from our Common Crawl corpus after a DCLM/RefinedWeb-style cleaning pipeline (Raffel et al., 2020; Penedo et al., 2023, 2024; Li et al., 2024). Within each LDS environment, all methods share the same candidate pool, reference checkpoint, subset-retraining recipe, and evaluation targets. The headline tasks are OpenBookQA (Mihaylov et al., 2018), CommonsenseQA (Talmor et al., 2019), SciQ (Welbl et al., 2017), and BasicSkills tasks from the OLMo evaluation family (Gu et al., 2025; Olmo, 2026). LDS mean, positive fraction, task-level Spearman ρ, and pairwise accuracy are computed from the same retrained subset utilities; Table 1 uses task-level ρ as the primary ranking metric.

Attribution methods. BM25, TracIn (Pruthi et al., 2020), TRAK (Park et al., 2023), and GRASP all score the same CC-10B candidate rows. BM25 lowercases text, tokenizes with `[A-Za-z0-9_]+`, removes a fixed English stopword and question-word list, drops tokens shorter than two characters, and uses $k_1 = 1.2$, $b = 0.75$ with the top 50k training rows retained per evaluation query in float16. TracIn accumulates normalized gradient cosine similarity from bf16 forwards with sequence length 2048, using `model.layers.9` and `model.norm`, a 128-dimensional random projection, and learning-rate-weighted checkpoints at steps $1000, 3000, 5000, 7000, 8000$. TRAK uses the same layer set, checkpoint set, sequence length, dtype, and 128-dimensional projected features; the reported rows use the five-checkpoint TRAK variant. Completed language InRun-DS runs (Wang et al., 2025) use the Ghost replay engine with global-evaluation utility, AdamW, learning rate 5×10^{-5}, one replay epoch over a 50k-row train prefix and 256 validation examples, active layers spanning `embed_tokens`, the attention and MLP projections, and `lm_head`. We report these InRun-DS runs as global-utility diagnostics rather than Table 1 rows, because faithful per-target LDS replay is substantially more expensive in this setting. Datamodel predictors (Ilyas et al., 2022) are treated as supervised oracle diagnostics when trained on LDS outcomes.

GRASP artifacts and protocol selection. GRASP builds an additive residual-alignment cache from the same 50M checkpoint and exports 64-dimensional hidden and residual sketches. Unless otherwise stated, token features use tail-mean-32 aggregation, residual top-k is 64, and residual sketches use the $\gamma^{-1/2}$ weighting from Section 3.1. Component weights, interaction scales, and retained/omitted conventions are selected from finite calibration families using development environments only, then frozen before evaluation on reported test rows. Omitted-set scoring is used only as a fixed deletion-protocol proxy; we do not assume retained-set and omitted-set utilities are exact negatives.

Efficiency and diagnostics. Table 2 separates artifact construction from cached subset scoring. The cached sweep scores random candidate subsets from the same local CPU process and extrapolates to 100k subsets; the construction column uses rep-

Labels	Rnd.	BM25	Ovlp.	G	G+ BM25	G+ Ovlp.
Topic	1.087	5.600	5.442	5.281	**7.121**	7.039
Format	0.979	3.160	3.363	3.303	4.041	**4.100**

Table 6: **Related-domain retrieval diagnostic.** We evaluate CC-10B rows with WebOrganizer labels (Wettig et al., 2025). Values are mean normalized Recall@1000; higher is better; G=GRASP.

resentative completed runs for each method. The one-step fidelity diagnostic in Figure 2 compares GRASP against Monte Carlo Shapley and fixed-size one-step subset utilities, while Figure 3 removes component families without re-searching weights.

Application diagnostics. The DCLM-style curation study trains a 1B-class tied-embedding LLaMA model on matched-size CC pools. Full runs select two million rows, use sequence length 2048 on eight GPUs, per-device batch size 4, gradient accumulation 16, learning rate 10^{-4}, cosine decay, bf16, FlashAttention-2, and DeepSpeed ZeRO-2 when available. Table 3 reports full-budget held-out evaluations in the main text, while Appendix Table 8 reports training-log loss snapshots through 14k steps. The vision setting follows the CIFAR-10/ResNet18 data-valuation protocol of Ghorbani and Zou (2019); the WebOrganizer retrieval setting is described with its results in Appendix C.

C WebOrganizer Qualitative Attribution Diagnostics

This qualitative diagnostic groups scored CC-10B rows by WebOrganizer topic and format labels. It is designed to inspect what kinds of source regions GRASP values for a SciQ target, rather than to serve as a headline benchmark.

C.1 Related-Domain Retrieval

All retrieval methods rank the same 88,781 scored candidate rows and are evaluated with normalized Recall@1000. We report topic and format retrieval separately because lexical overlap and attribution can favor different source types.

Table 6 shows that attribution and lexical signals are complementary: GRASP alone is competitive, and reranking with BM25 or query overlap gives the strongest retrieval result for topic and format labels.

C.2 Checkpoint-Wise Category Contributions

For interpretability, we aggregate mean signed attribution scores by WebOrganizer category across several training checkpoints. Positive and negative mass are normalized separately so that category shares remain visible even when the total signed magnitude changes across checkpoints.

Figure 4 summarizes the signed category-level mean attribution scores used in the WebOrganizer diagnostic. It is intended as a qualitative probe rather than a benchmark metric. The topic view shows that the SciQ target assigns positive average value to several knowledge-bearing source regions, while broad web categories with weaker task alignment fall on the negative side. The format view similarly separates high-value article-like formats from lower-value boilerplate formats such as notices, support pages, and FAQ-style text. These patterns support the interpretation of GRASP as a model-conditioned data-value signal: it is not simply recovering the raw source distribution, and it differs from lexical matching by assigning signed value to broad source regions across training checkpoints.

D Supplementary LDS Robustness Results

This section reports cached or partially cached diagnostics that stress the LDS protocol beyond the headline Table 1. All experiments reuse existing subset-retraining outcomes and attribution artifacts; they are intended to characterize robustness rather than introduce new training runs.

D.1 Retention-Ratio Sensitivity

We vary the retained fraction in the SciQ LDS environment while holding the candidate pool, retraining recipe, and evaluation target fixed. The goal is to test whether the interaction-aware geometry remains useful when the omitted slice ranges from large deletions to near-additive high-overlap settings. Table 5 reports this diagnostic in the main text to avoid duplicating the same result in the appendix. The gains are largest at moderate deletion levels, where subset interactions are most visible. At 0.99 retention, the protocol properly approaches a near-additive limit, and GRASP ties the additive core rather than amplifying noisy interaction terms.

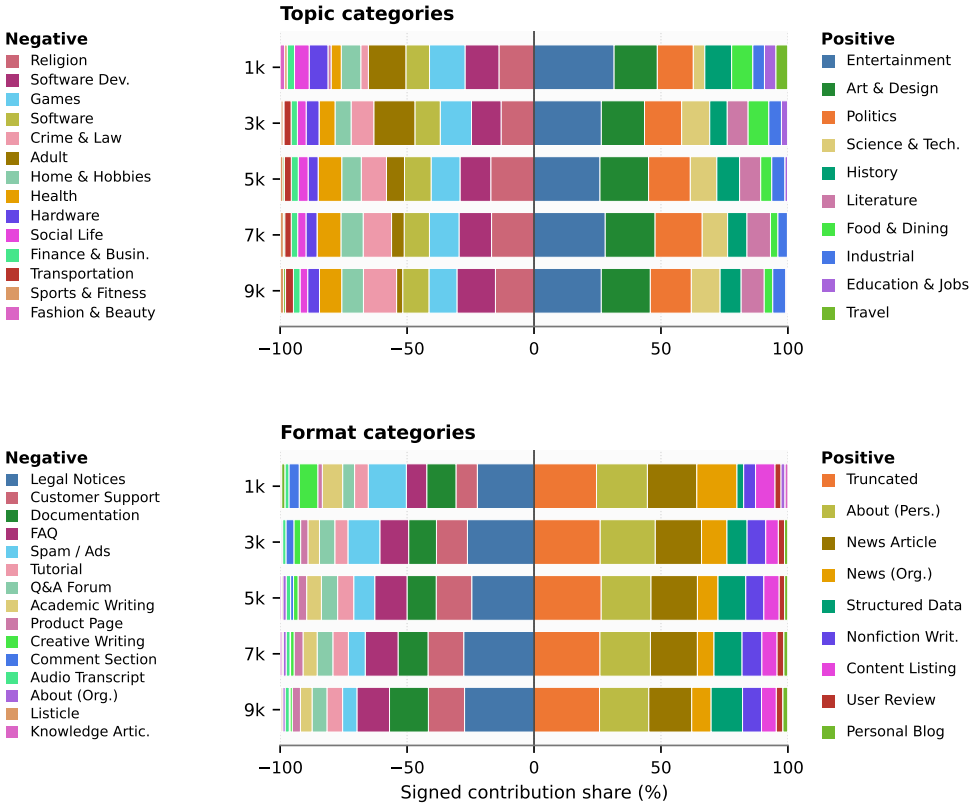

Figure 4: **Topic and format contribution shares.** We aggregate GRASP attribution across training checkpoints for a SciQ target. Positive and negative mean-score mass are normalized separately within each checkpoint.

Calib.	SciQ ρ	Logic ρ	OBQA ρ	Pos.
SciQ	**0.420**	0.298	0.309	0.876
Logical	0.333	**0.470**	0.307	0.906
SciQ + Logical	0.398	0.451	**0.314**	**0.909**

Table 7: **Cached GRASP protocol-selection diagnostic.** Each row selects fixed interaction-component weights using the listed calibration task(s), then reuses the same cached hidden/residual sketches and subset-retraining ground truth to evaluate all columns. No additional model training is performed.

D.2 Cached Protocol-Selection Diagnostic

This diagnostic studies the stability of fixed GRASP component protocols. Each row calibrates interaction-component weights on the listed development task(s), freezes them, and evaluates the same cached scores and sketches on all three strong LDS tasks.

Table 7 illustrates the expected tradeoff: single-task calibration is strongest on its own task, while joint SciQ+Logical calibration improves transfer to OpenBookQA and gives the highest average positive fraction. We therefore treat this table as a protocol-selection diagnostic, not as a model-training hyperparameter study.

E Supplementary Curation Results

The following diagnostics evaluate whether subset-level attribution can support larger data-selection workflows. These experiments use matched training budgets within each setting, but they are not used as the main LDS fidelity evidence.

E.1 CC Curation and Cleaning Diagnostics

Fixed-compute DCLM curation. We train the same 1B-class language model on matched-size CC-10B pools selected by each policy. The completed full-budget held-out evaluations are reported in Table 3. As an auxiliary optimization diagnostic, Table 8 reports training-log losses at matched steps; these are not held-out validation losses.

Method	9k ↓	12k ↓	14k ↓
GRASP	**2.6342**	**2.4965**	**2.4614**
InRun-1	2.7297	2.6637	2.6020
Random	2.7540	2.6447	2.6221
BM25	2.7920	2.6830	2.6532

Table 8: **Early training-log loss snapshots.** Columns report training loss at the indicated step for the DCLM fixed-compute scale-up diagnostic.

Model	InRun-1	InRun-2	GRASP	Gain
GPT-2	5.8417	<u>5.8237</u>	**5.7788**	0.0449
Pythia	<u>5.6937</u>	5.7020	**5.5947**	0.0990

Table 9: **Best observed full32k CC-cleaning runs.** We compare GRASP across model architectures. Each row reports final held-out loss; lower is better. Bold and underline mark the best and second-best entries, and Gain reports the loss reduction over the runner-up.

Method	F1	AUC	Prec.	Rec.
STRIP	0.177	0.645	0.097	**1.000**
ONION	0.172	0.636	0.099	0.750
Pred-Div topK	0.025	0.355	0.025	0.025
IFE strict-first-n	0.182	0.839	0.182	0.182
IFE	0.200	0.884	0.200	0.200
README-exact				
GRASP	**0.300**	**0.886**	**0.300**	0.300

Table 10: **Safety and misleading-information detection.** Rows are evaluated under the same single-run diagnostic matrix.

Cross-architecture compact CC cleaning. We also evaluate compact CC-cleaning runs under GPT-2-style and Pythia-410M architectures. This setting probes whether the same data-scoring rule can produce favorable cases when the model architecture changes. Table 9 reports the strongest observed full32k row for GRASP on each model architecture. In both GPT-2 and Pythia-410M, GRASP attains the lowest final held-out loss among the displayed methods, showing that the same scoring rule can produce positive cases under both architectures.

F Supplementary Safety Detection Results

This diagnostic studies whether attribution scores can identify suspicious or misleading fine-tuning examples. The table compares official lightweight baselines with GRASP under the same diagnostic protocol.

Table 10 indicates that the same attribution signal can be repurposed for safety-oriented ranking: GRASP improves F1 and AUC over the lightweight detection baselines in this diagnostic.