MATCHED PARENT-REMOVAL AUDITS EXPOSE WHEN AUGMENTATION PROVENANCE CHANGES DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Failures in augmented training corpora propagate from original assets through rewriting, filtering, distillation, and merge operations, making the root assets that require inspection or removal difficult to identify. We introduce a matched parent-removal audit that isolates the decision value of provenance by fixing the target model, candidate set, intervention, and top-k action while varying the information available to the ranker. Across four logged text and code workflows, graph structure improves top-5 recovery over flat attribution, while a learned path ranker adds 4.3–8.3 points over structure-only Metadata Provenance. The ordering persists under hard-negative and reachable-only controls built from the same candidate manifests. Controlled operator-label corruption further maps where learned structured rankers degrade relative to structure-only provenance. Together, these results establish fixed-decision auditing as a rigorous way to measure the decision value of recorded provenance for parent ranking.

1 INTRODUCTION

Augmentation pipelines can transform one source into hundreds of training examples. When a root contains copyrighted material, contamination, private information, or a policy violation, deleting an arbitrary descendant is ineffective: the remediation target is the original asset and every downstream derivative. Logged lineage appears to solve this tracing problem, yet its actual decision value remains unidentified when evaluations simultaneously change the model, candidate roots, or removal rule. A gain under such a moving decision can come from an easier candidate set rather than from provenance.

We make the remediation decision itself the controlled scientific object. For each terminal example, the audit fixes a target model, a candidate set containing the oracle top five, a leave-one-parent-out replay intervention, and a top-k removal action. It then compares flat example attribution, structure-only provenance, learned path ranking, and whole-DAG graph ranking on that same decision. This matched design separates ancestry filtering from within-ancestor ranking and measures the marginal value of each recorded provenance signal without changing what must be decided.

The evaluation compares TracIn, TRAK, Metadata Provenance, GIS, DPT, a whole-DAG graph transformer, and an R-GCN family under matched pilot labels, candidate pools, and replay targets. These families distinguish the value of recorded ancestry, operator labels, path representations, and whole-DAG structure without changing the removal decision.

The evaluation spans four logged text and code pipelines together with controlled text, vision, and cross-modal DAGs. On A–D, DPT improves top-5 parent recovery over structure-only Metadata Provenance by 4.3–8.3 points. Hard-negative and reachable-only controls confirm that the structured advantage includes within-ancestor ordering rather than easy rejection of unreachable roots. The random-corruption sweep provides an inspectable metadata-integrity stress test for rankers that consume operator labels.

The paper contributes:

1. a matched fixed-decision estimand and audit protocol that isolates provenance value under one target, candidate construction, replay intervention, and removal action;
2. a multi-family provenance ladder that compares flat, structure-only, learned-path, and whole-DAG rankers under the same candidate manifests and replay targets;

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

Table 1: Closest comparison classes and the controls used to assess recorded-provenance value. T/P/A marks a fixed target, candidate pool, and action rule; "–" indicates that the family has no removal-decision object.

Work family	T/P/A	DAG	Operators	Evaluation object
Atomic attribution (Koh & Liang, 2017; Pruthi et al., 2020; Park et al., 2023)	T/P	no	no	point influence on a trained model
Data valuation (Ghorbani & Zou, 2019; Kwon & Zou, 2022)	T/P	no	no	utility contribution of a selected set
Lineage systems (Herschel et al., 2017; Schelter et al., 2017)	–	yes	recorded	dependency capture and query
Unlearning (Bourtoule et al., 2021)	A	no	no	model update after a supplied request
Curation benchmarks (Gadre et al., 2023)	varies	partial	filtering	end-model utility across pipelines
Matched parent-removal audit	T/P/A	yes	evaluated	root intervention and decision conversion

3. a controlled provenance-integrity stress map supported by hard-negative controls, graph-native baselines, and operator-label corruption.

Together these contributions convert provenance from passive lineage metadata into a testable input to a fixed parent-ranking decision.

2 Related Work and Positioning

Per-example influence functions, TracIn, representer points, TRAK, and Shapley-style valuation score atomic training records (Koh & Liang, 2017; Pruthi et al., 2020; Yeh et al., 2018; Park et al., 2023; Ghorbani & Zou, 2019; Kwon & Zou, 2022; Jia et al., 2019). They provide strong flat baselines for the same leave-one-parent-out target, but their data model discards the transformation paths that connect a source to its derivatives. Graph explanation methods use topology to explain predictions (Duval & Malliaros, 2021); their object is a trained graph model rather than removal of source assets from an augmentation workflow.

Database provenance and ML lineage systems record dependencies and operators (Buneman et al., 2001; Cheney et al., 2009; Green et al., 2007; Herschel et al., 2017; Schelter et al., 2017). Machine unlearning then executes deletion requests efficiently (Bourtoule et al., 2021). The missing link is selection: lineage says which derivatives are connected, and unlearning says how to update a model after a request, while parent-removal triage must determine which roots warrant that request. The matched audit evaluates root selection before the downstream unlearning mechanism is chosen.

Augmentation and curation studies establish realistic transformation pipelines (Eldan & Li, 2023; Gunasekar et al., 2023; Liu et al., 2024; Longpre et al., 2023; Gadre et al., 2023; Soldaini et al., 2024). Their evaluations appropriately optimize end performance, often changing datasets, models, filters, and compute together. That design answers which complete pipeline works best. Our question is complementary and decision-centric: with the target, candidates, and action held constant, how much does each increment of provenance change the same remediation decision?

Among these closest comparison classes, each differs on at least one control: atomic methods omit logged paths, lineage systems omit the matched action, and curation benchmarks vary the selected data with the pipeline. Freezing target, pool, and action prevents that variation from being attributed to provenance, but does not identify contributions of unlogged parents. The resulting estimand measures the marginal decision value of recorded provenance across progressively richer ranker families.

3 FIXED-DECISION ESTIMAND ON LOGGED DAGS

3.1 PROBLEM SETUP

We model the logged workflow as a stochastic provenance DAG $\mathcal{G} = (V, E, \mathcal{O})$ with root nodes V_0, terminal nodes V_T, and operator labels $o_{uv} \in \mathcal{O}$ on edges (u, v). Given a surrogate f_θ trained on V_T, the task is to score each root $v_0 \in V_0$ for each terminal $v_T \in V_T$ under the same leave-one-parent-out target used everywhere else in the paper.

The audit uses only the observed record: every ranked candidate is an observed root and every structural feature is computed from observed edges and recorded labels. It does not impute unlogged parents or assign them a contribution. The rankers differ in whether they consume flat features, graph topology, recorded operator labels, path representations, or whole-DAG message passing; the target, candidate pool, and intervention are identical across them.

3.2 ESTIMAND AND ANALYSIS FRAME

Our primary estimand is per-terminal parent influence. Let ξ_{rep} and ξ_{tr} denote mutually independent operator-replay and training randomness. Write $\mathcal{D}(\mathcal{G}; \xi_{\text{rep}})$ for the terminal corpus induced by the fixed operator-class DAG and $\theta(\mathcal{G}; \xi_{\text{rep}}, \xi_{\text{tr}}) = \mathcal{A}(\mathcal{D}(\mathcal{G}; \xi_{\text{rep}}); \xi_{\text{tr}})$ for the trained parameters. Matched intervention and control replays share random numbers on unaffected operators; independent replay/training seeds estimate the outer expectation. The quantity used to rank parents is

$$\phi^\star(v_0, v_T) = \mathbb{E}_{\xi_{\text{rep}}, \xi_{\text{tr}}} \left[\ell_{v_T}\left(f_{\theta(\mathcal{G} \setminus v_0; \xi_{\text{rep}}, \xi_{\text{tr}})}\right) - \ell_{v_T}\left(f_{\theta(\mathcal{G}; \xi_{\text{rep}}, \xi_{\text{tr}})}\right) \right], \tag{1}$$

the expected increase in v_T-loss when v_0 and its descendants are removed and the same operator-class chain is replayed. Larger $\phi^\star(v_0, v_T)$ means that removing v_0 harms performance on v_T more, and every ranking metric sorts candidates by descending $\phi^\star$. Leave-one-parent-out replay and retraining recover this expectation (Section 4.2).

The aggregate estimand is the corresponding corpus-level bookkeeping quantity. The corpus-level influence of v_0 is the sum over reachable terminals:

$$\Phi(v_0) \;=\; \sum_{v_T \in V_T:\, v_0 \rightsquigarrow v_T} \phi^\star(v_0, v_T), \tag{2}$$

which provides the replay-based reference for evaluating every tested ranker.

The diagnostic experiments perturb recorded operator labels and observed provenance. These interventions characterize sensitivity of the structured rankings under the matched protocol.

4 MATCHED AUDIT PROTOCOL

4.1 EVIDENCE BLOCKS AND WORKLOADS

The evidence follows the fixed ranking decision. Logged pipelines A–D establish the provenance ladder; candidate controls isolate ancestry from within-ancestor ranking, and graph-native families test learner dependence. A controlled suite spans text, vision, and cross-modal chains at depths $d \in \{2, 3, 4, 5, 6, 7, 8\}$ with exact lineage, enabling controlled depth and operator perturbations.

Text chains start from $5{,}000$ C4 paragraphs and apply GPT-3.5-turbo-0125 paraphrase, NLLB-200-3.3B back-translation (NLLB Team et al., 2022; Sennrich et al., 2016), and GPT-4o-mini merge summarization; DAGs contain 8K–45K nodes. Vision chains start from $3{,}000$ ImageNet-1K images and apply RandAugment (Cubuk et al., 2020), CutMix (Yun et al., 2019), AdaIN transfer (Huang & Belongie, 2017), and spatial merges (5K–32K nodes). Cross-modal chains start from $2{,}000$ COCO pairs and alternate BLIP-2 captioning, text rewriting, and Stable Diffusion v2.1 regeneration (4K–22K nodes).

4.2 BASELINES, METRICS, AND CANDIDATE POOLS

The ladder contains Random, Embedding Cosine, Gradient Similarity, TracIn (Pruthi et al., 2020), TRAK (Park et al., 2023), Metadata Provenance (MP; uniform merge allocation), GIS (one pilot-fitted

scalar per operator), DPT (an 18.4M path transformer), a 24.8M whole-DAG graph transformer, and a 9.7M R-GCN (Schlichtkrull et al., 2017). Text nodes use SentenceBERT-all-MiniLM-L6-v2 (Reimers & Gurevych, 2019); vision nodes use CLIP-ViT-B/32 (Radford et al., 2021); cross-modal paths retain modality-specific embeddings before a learned 256-dimensional projection. DPT and both graph-native models train on every labeled pair in the fixed 10% pilot split, while GIS fits one scalar per recorded operator. Appendix J reports parameter count, pilot access, GPU-hours, and inference cost separately.

For each terminal, the oracle ranks reachable roots by leave-one-parent-out loss increase, using a manifest-fixed hash to break ties. With $S_5^\star$ and $\widehat{S}_5$ denoting oracle and predicted top-five sets, recovery is $100|\widehat{S}_5 \cap S_5^\star|/5$. Spearman ρ measures ordering over the shared candidate pool, and decision tables report recovered positive loss, absolute loss change, and action disagreement. Top-5 uses five replay/retraining seeds and ρ uses three. BERT-base-uncased (Devlin et al., 2019) (110M) is the text and terminal-text target, ResNet-18 (He et al., 2016) (11.7M) is the vision target, and Pipeline A also uses DeBERTa-v3-large (He et al., 2021) (304M backbone parameters). Runs use one NVIDIA A100-SXM4-40GB with CUDA 12.4, driver 550.54.15, and PyTorch 2.3.1.

Candidate construction. Let R_t be the reachable roots for terminal t, and compute $S_5^\star$ once on the full R_t solely to guarantee oracle coverage in the evaluation pool. If $|R_t| \leq 50$, the primary pool contains all of R_t and uniformly sampled non-reachable roots up to $K = 50$. If $|R_t| > 50$, it retains $S_5^\star$ and chooses the remaining 45 reachable roots by a manifest-fixed hash set before any method is run. The hard-negative pool uses the same rule but fills with non-reachable roots matched on source degree, content similarity, and operator-profile frequency. The reachable-only control contains all reachable roots when $|R_t| \leq 50$ and the same oracle-five-plus-hash subset otherwise. Methods receive the pool without oracle labels or membership flags; every method sees the same roots, and reachable-only scores isolate within-ancestor ranking from the ancestry-filtering advantage. Appendix H reports the minimum, median, 90th percentile, and maximum $|R_t|$ for A–D, along with both controls.

Intervention and aggregation. Roots are split before augmentation. For every candidate, the evaluator (1) removes the root globally within its split and deletes all reachable descendants, (2) replays the same operator classes with fresh replay randomness, (3) retrains the fixed surrogate, (4) records terminal loss changes, (5) ranks candidates, and (6) applies the same top-k action. Split-scoped node identifiers prevent regenerated descendants from entering another train/evaluation split. Roots shared by terminals are removed once per replay, and their terminal effects are aggregated with equal terminal weight.

Seed standard deviations quantify training/replay randomness. Across-pipeline variation is summarized by split-specific effects rather than treating pipelines as exchangeable replicates. The annotation unit is an observed edge nested within a sampled chain. All agreement statistics are pre-adjudication Fleiss' κ over edge-validity labels from three independent raters; positive-label prevalence is reported with each verification sample, and adjudication occurs only after agreement is frozen.

5 Results: Provenance Value and Action Boundaries

The matched audit asks whether structure improves one fixed decision, whether operator semantics improve within-ancestor ordering, how metadata degradation changes the observed-log ranking, and whether this behavior persists across graph-native learner families.

5.1 Logged Structure Improves the Same Parent Ranking

We evaluate four logged pipelines spanning clean, noisy, and weakly observed provenance. Three raters independently label every observed edge in $120/150/80/100$ sampled A/B/C/D chains; pre-adjudication Fleiss' κ is $0.81/0.78/0.72/0.74$, with positive-edge prevalence 93.1/90.7/82.4/85.3%. Pipeline A applies paraphrase, NLLB back-translation, and two-parent summarization to 800 Wikipedia roots, producing $2{,}400$ terminals. Pipeline B applies rewriting, LLaMA-3-8B distillation, Qwen2.5 paraphrase, and three-parent summarization to 700 StackExchange pairs, producing $2{,}100$ terminals. Pipeline C applies five code-refinement stages to 600 CodeSearchNet pairs; $216/1{,}800$ Stage-1 edges lack operator labels. Pipeline D applies four dialogue-refinement stages to 500 MS MARCO conversations; $213/1{,}500$ terminals contain hidden-parent residue.

Table 2: Logged-pipeline ranking under fixed $K = 50$ pools. Panel A reports top-5 recovery (%) with five replay/retraining seeds; Panel B reports Spearman ρ with three seeds. Parentheses are seed standard deviations.

Panel A: Top-5 parent recovery (%)

Method	A (Wiki)	B (StackEx.)	C (Code)	D (Dialog.)
TracIn	58.2 ($\pm$ 3.1)	54.6 ($\pm$ 2.9)	51.3 ($\pm$ 3.2)	51.8 ($\pm$ 3.3)
TRAK	60.1 ($\pm$ 3.0)	56.8 ($\pm$ 2.8)	53.7 ($\pm$ 3.0)	54.0 ($\pm$ 3.1)
Metadata Provenance	65.4 ($\pm$ 2.7)	61.8 ($\pm$ 2.5)	58.6 ($\pm$ 2.8)	58.9 ($\pm$ 2.9)
GIS	71.2 ($\pm$ 2.5)	68.4 ($\pm$ 2.4)	63.8 ($\pm$ 2.6)	62.1 ($\pm$ 2.7)
DPT	73.0 ($\pm$ 2.3)	70.1 ($\pm$ 2.2)	65.5 ($\pm$ 2.4)	63.2 ($\pm$ 2.5)
Δ(DPT–MP)	+7.6	+8.3	+6.9	+4.3

Panel B: Spearman ρ

Method	A	B	C	D
TracIn	0.49	0.45	0.41	0.42
TRAK	0.51	0.47	0.44	0.44
Metadata Provenance	0.57	0.54	0.50	0.48
GIS	0.63	0.60	0.54	0.53
DPT	0.66	0.63	0.56	0.54

Table 3: Candidate controls, top-5 recovery averaged over A/B and C/D. Full split-level top-5/ρ results appear in Appendix H.

Method	Hard clean	Hard weak	Reachable clean	Reachable weak
TRAK	53.5	48.7	51.6	47.5
MP	60.5	56.0	57.6	53.6
DPT	68.5	61.6	65.6	59.3
R-GCN	71.0	63.2	67.1	60.4
Graph Transformer	72.0	64.5	68.2	61.4

MP exceeds TRAK in top-5 recovery by 4.9–5.3 points, and DPT exceeds MP by 4.3–8.3 points across the four logged pipelines. The DPT margin is $7.6/8.3$ points on A/B and $6.9/4.3$ points on C/D, while the structure-only advantage remains positive throughout.

Hard negatives reduce absolute recovery, and reachable-only ranking removes the easy ancestry signal. On both controls, the observed-DAG methods retain their advantage over flat attribution among plausible ancestors.

5.2 Rankings Produce Stable Removal Decisions

We apply the joint top-k intervention on Pipeline A. Table 4 reports oracle-relative recovered positive loss for $k = 1/5/10$. For each terminal, the numerator is the nonnegative loss change from jointly removing the method's top-k roots and the denominator is the corresponding change from jointly removing the oracle top-k roots. For each k, terminals with a zero oracle denominator are excluded from the ratio mean; signed and absolute-loss summaries are not used in this table.

At $k = 5$, the graph transformer recovers 88.9% of the Pipeline A oracle loss under the same joint intervention. Its result illustrates that ranking overlap and joint loss recovery weight errors differently.

5.3 Metadata Integrity Characterizes Ranking Sensitivity

The graph transformer and R-GCN cross the MP floor between 15% and 20% random corruption, while DPT crosses between 10% and 15%. Targeting influential edges moves the graph-transformer crossover to [10%,15%]. The random sweep corrupts 360/720/1080/1440/1800 of 7200 edges at the five nonzero rates. Appendix K gives the corresponding top-5 and Spearman trends. The stress test

Table 4: Pipeline A oracle-relative recovered loss (%) for $k = 1/5/10$, reported in that order. Appendix I defines the joint top-k numerator and denominator.

Method	A
TracIn	44.8/51.7/55.9
TRAK	47.1/54.1/58.6
MP	60.8/68.9/73.4
GIS	68.5/76.2/80.1
DPT	72.6/80.4/84.2
R-GCN	79.4/87.8/91.0
Graph Transformer	80.1/88.9/91.8

Table 5: Pipeline A top-5 under random and targeted operator-label corruption at $0/5/10/15/20/25\%$. Targeted attacks corrupt high-centrality or high-influence edges held out from calibration.

Method	Random corruption	Targeted corruption
MP	65.4/65.4/65.4/65.4/65.4/65.4	65.4/65.4/65.4/65.4/65.4/65.4
DPT	73.0/70.7/68.2/64.9/61.7/58.0	73.0/68.9/64.2/59.8/55.1/51.0
R-GCN	75.8/73.0/70.4/66.7/62.8/59.1	75.8/70.6/65.8/60.9/56.0/51.8
Graph Transformer	77.0/74.4/71.8/68.1/64.2/60.5	77.0/72.1/67.0/62.2/57.4/53.0

shows that recorded operator integrity is a key condition for learned structured ranking, while the MP row isolates the more durable topology-only signal.

6 Discussion

The audit evaluates ranking quality conditional on a fixed candidate construction: oracle top-five roots are retained when the reachable set exceeds the budget. The results therefore isolate the decision value of recorded provenance inside the fixed parent-ranking decision. The present evidence covers logged augmentation and refinement DAGs with fan-in at most 12 and surrogate targets up to DeBERTa-v3-large. Its most direct implication is a practical evaluation standard: compare provenance views only after target, candidate pool, replay intervention, and ranking action have been fixed.

7 Conclusion

Parent-removal evaluation is well-specified when target, candidate pool, replay intervention, and action remain fixed. Under that control, graph structure improves root ranking, learned path and graph rankers improve on structure-only provenance in the reported logs, and operator-label corruption exposes the sensitivity of rankers that consume those labels. The matched audit thus supplies a practical standard for evaluating provenance-aware parent ranking.

Ethics Statement

The annotation studies used consenting paid adult annotators, institution-neutral procedures, and public or licensed data. The protocol treats false-positive root removal as a material harm because descendants are deleted with the root.

Reproducibility Statement

Sections 3.2 and 4.2 define the estimand, candidate construction, replay, aggregation, and uncertainty units. Appendix E specifies hyperparameters, seeds, models, licenses, annotation prevalence, and the paper-evidence map for the retained tables and figures.

AI USE STATEMENT

Generative AI was used only for manuscript writing, editing, and presentation. The authors reviewed all edits and remain responsible for the final content.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy*, pp. 141–159, 2021. doi: 10.1109/sp40001.2021.00019. URL https://doi.org/10.1109/sp40001.2021.00019.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *International Conference on Database Theory*, pp. 316–330, 2001. doi: 10.1007/3-540-44503-x_20. URL https://doi.org/10.1007/3-540-44503-x_20.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. doi: 10.48550/arXiv.1504.00325. URL https://arxiv.org/abs/1504.00325.

James Cheney, Laura Chiticariu, and Wang-Chiew Tan. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474, 2009. doi: 10.1561/1900000006. URL https://doi.org/10.1561/1900000006.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. RandAugment: Practical automated data augmentation with a reduced search space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020. URL https://openaccess.thecvf.com/content_CVPRW_2020/html/w40/Cubuk_Randaugment_Practical_Automated_Data_Augmentation_With_a_Reduced_Search_Space_CVPRW_2020_paper.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Alexandre Duval and Fragkiskos D. Malliaros. GraphSVX: Shapley value explanations for graph neural networks. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 302–318, 2021. doi: 10.1007/978-3-030-86520-7_19. URL https://doi.org/10.1007/978-3-030-86520-7_19.

Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*, 2023. doi: 10.48550/arXiv.2305.07759. URL https://arxiv.org/abs/2305.07759.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27092–27112, 2023. doi: 10.52202/075280-1179. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/56332d41d55ad7ad8024aac625881be7-Abstract-Datasets_and_Benchmarks.html.

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

Todd J Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 31–40, 2007. doi: 10.1145/1265530.1265535. URL https://doi.org/10.1145/1265530.1265535.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023. doi: 10.48550/arXiv.2306.11644. URL https://arxiv.org/abs/2306.11644.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv preprint arXiv:2111.09543*, 2021. doi: 10.48550/arXiv.2111.09543. URL https://arxiv.org/abs/2111.09543.

Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. A survey on provenance: What for? what form? what from? *The VLDB Journal*, 26(6):881–906, 2017. doi: 10.1007/s00778-017-0486-1. URL https://doi.org/10.1007/s00778-017-0486-1.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Huang_Arbitrary_Style_Transfer_ICCV_2017_paper.html.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. CodeSearchNet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*, 2019. doi: 10.48550/arXiv.1909.09436. URL https://arxiv.org/abs/1909.09436.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the Shapley value. In *International Conference on Artificial Intelligence and Statistics*, pp. 1167–1176, 2019. URL https://proceedings.mlr.press/v89/jia19a.html.

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

Yongchan Kwon and James Zou. Beta Shapley: A unified and noise-reduced data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 8780–8802, 2022. URL https://proceedings.mlr.press/v151/kwon22a.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M Dai. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=OJaWBhh61C.

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

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022. doi: 10.48550/arXiv.2207.04672. URL https://arxiv.org/abs/2207.04672.

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

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Nils Reimers and Iryna Gurevych. 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)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

Sebastian Schelter, Joos-Hendrik Böse, Johannes Kirschnick, Thoralf Klein, and Stephan Seufert. Automatically tracking metadata and provenance of machine learning experiments. In *Machine Learning Systems Workshop at NIPS 2017*, 2017. URL https://www.amazon.science/publications/automatically-tracking-metadata-and-provenance-of-machine-learning-experiments.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling Relational Data with Graph Convolutional Networks. *arXiv preprint arXiv:1703.06103*, 2017. doi: 10.48550/arXiv.1703.06103. URL https://arxiv.org/abs/1703.06103.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, 2016. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009/.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/8a7129b8f3edd95b7d969dfc2c8e9d9d-Abstract.html.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Yun_CutMix_Regularization_Strategy_to_Train_Strong_Classifiers_With_Localizable_Features_ICCV_2019_paper.html.

A Controlled Fixed-Pool Depth Suite

Table 6 fixes $K = 50$ while varying chain depth under exact lineage. This isolates propagation length from candidate pressure and complements the logged workflows. Structure and recorded operator labels improve recovery at every depth.

Table 6: Controlled-suite top-5 recovery (%) under a fixed K=50 candidate pool per terminal, matching the logged-pipeline ranking budget. Each cell is averaged over the three chain types (text, vision, cross-modal) at that depth, with standard deviations over 5 seeds. Higher is better.

Method	Depth 2	Depth 4	Depth 6	Depth 8
Random (K=50)	9.7 ($\pm$ 0.6)	9.7 ($\pm$ 0.6)	9.7 ($\pm$ 0.6)	9.7 ($\pm$ 0.6)
TracIn	74.2 ($\pm$ 1.9)	52.1 ($\pm$ 2.4)	36.4 ($\pm$ 2.6)	23.7 ($\pm$ 2.2)
TRAK	76.3 ($\pm$ 1.8)	54.2 ($\pm$ 2.3)	37.7 ($\pm$ 2.6)	25.2 ($\pm$ 2.2)
Metadata Provenance	82.4 ($\pm$ 1.7)	61.4 ($\pm$ 2.1)	43.3 ($\pm$ 2.4)	31.8 ($\pm$ 2.0)
GIS	86.7 ($\pm$ 1.5)	69.7 ($\pm$ 1.9)	52.9 ($\pm$ 2.1)	40.8 ($\pm$ 2.0)
Δ(GIS–MP)	+4.3	+8.3	+9.6	+9.0

GIS improves over MP by 4.3–9.6 points across the controlled depths. The empirical Random row averages five fixed-K pool draws, explaining its small deviation from analytic $5/K$.

B Per-Modality Structured Frontier

Table 7 resolves the depth-4 result by text, vision, and cross-modal chain. Cross-modal regeneration is hardest in absolute recovery, while the structural ordering transfers across all three modalities.

Table 7: Top-5 parent recovery (%) by chain type at depth 4. Rows average within modality over five seeds.

Method	Text	Vision	Cross-Modal
TracIn	56.2 ($\pm$ 2.3)	53.8 ($\pm$ 2.6)	48.1 ($\pm$ 2.9)
TRAK	58.3 ($\pm$ 2.2)	55.9 ($\pm$ 2.5)	50.2 ($\pm$ 2.8)
Metadata Provenance	63.7 ($\pm$ 2.0)	62.1 ($\pm$ 2.3)	58.0 ($\pm$ 2.5)
GIS	71.8	70.2	67.1
DPT	74.5	73.0	69.6
R-GCN	78.9	77.7	72.5
Graph Transformer	80.7	79.5	74.4

The graph transformer remains 16.4–17.4 points above MP across modalities. Together with the path and R-GCN rows, this shows that the matched structure-versus-flat comparison is not restricted to one structured learner family.

C Full Spearman Correlation Ladder

Table 8 compares predicted and interventional full-pool rankings under three retraining seeds per removal, averaged over chain types. It complements top-5 recovery by measuring ordering below the action cutoff.

Table 8: Spearman ρ between predicted attribution and the interventional ranking, averaged over the three chain types.

Method	Depth 2	Depth 4	Depth 6
Random	0.01 ($\pm$ 0.03)	0.00 ($\pm$ 0.02)	0.01 ($\pm$ 0.02)
Embedding Cosine	0.58 ($\pm$ 0.04)	0.35 ($\pm$ 0.05)	0.21 ($\pm$ 0.04)
Gradient Similarity	0.63 ($\pm$ 0.04)	0.41 ($\pm$ 0.04)	0.26 ($\pm$ 0.04)
TracIn	0.72 ($\pm$ 0.03)	0.51 ($\pm$ 0.04)	0.34 ($\pm$ 0.04)
TRAK	0.74 ($\pm$ 0.03)	0.53 ($\pm$ 0.04)	0.36 ($\pm$ 0.04)
Metadata Provenance	0.73 ($\pm$ 0.03)	0.55 ($\pm$ 0.04)	0.41 ($\pm$ 0.04)
GIS	0.78	0.62	0.49
DPT	0.80	0.65	0.52
R-GCN	0.83	0.70	0.57
Graph Transformer	0.84	0.71	0.59

Full-rank correlation reproduces the provenance ladder. At depth 6, MP improves over TRAK by 0.05, while DPT, R-GCN, and the graph transformer add 0.11–0.18 over MP.

D Variable-Pool Controlled Stress Test

The fixed-K companion holds $K{=}50$ candidates per terminal. As an additional stress test, we evaluate the full reachable-root pool for each terminal. Pool size therefore grows with depth, averaging $41/58/81/122$ at depths $2/4/6/8$. Table 9 shows that structure-only provenance remains above flat attribution as candidate pressure grows.

Table 9: Variable-pool controlled stress test: top-5 parent recovery (%) when each terminal ranks against its full reachable-root set. Average pool size is $41/58/81/122$ at depths $2/4/6/8$. Compare with the fixed-K companion table (Table 6).

Method	Depth 2	Depth 4	Depth 6	Depth 8
Random	12.3 ($\pm$ 1.1)	8.7 ($\pm$ 0.9)	6.2 ($\pm$ 0.8)	4.1 ($\pm$ 0.6)
Embedding Cosine	68.4 ($\pm$ 2.3)	42.1 ($\pm$ 2.8)	28.7 ($\pm$ 3.1)	19.3 ($\pm$ 2.7)
Gradient Similarity	71.2 ($\pm$ 2.1)	48.3 ($\pm$ 2.5)	31.4 ($\pm$ 2.9)	22.6 ($\pm$ 2.4)
TracIn	74.8 ($\pm$ 1.9)	52.7 ($\pm$ 2.4)	36.8 ($\pm$ 2.6)	24.1 ($\pm$ 2.2)
TRAK	76.9 ($\pm$ 1.8)	54.8 ($\pm$ 2.3)	38.1 ($\pm$ 2.6)	25.6 ($\pm$ 2.2)
Metadata Provenance	82.1 ($\pm$ 1.7)	61.3 ($\pm$ 2.1)	43.2 ($\pm$ 2.4)	31.7 ($\pm$ 2.0)

At depths $6/8$, the candidate count grows to $81/122$, directly testing structure-only ranking under deeper ancestry and broader candidate pools.

E Reproducibility: Hyperparameters, Licenses, Annotator Protocol

Fixed hyperparameters are as follows. TracIn checkpoint count $T = 10$; TRAK projection dimension $d_{\text{proj}} = 2048$; TRAK replicates $K = 8$; and DPT path truncation depth $D_{\max} = 12$. Surrogate optimizers are AdamW(lr=$2{\times}10^{-5}$, batch 32, 3 epochs) for BERT-base, AdamW(lr=$1{\times}10^{-5}$, batch 16, 3 epochs) for DeBERTa-v3-large, and SGD(lr=0.1, momentum 0.9, weight decay $1{\times}10^{-4}$, batch 256, 30 epochs) for ResNet-18. Random seeds per experiment are $\{0, 1, 2, 3, 4\}$; dense Spearman sweeps use the first three seeds.

Dataset and model licenses are as follows. C4 (Raffel et al., 2020): ODC-BY. ImageNet-1K (Russakovsky et al., 2015): ImageNet non-commercial research license. COCO Captions (Chen et al., 2015): CC BY 4.0. Wikipedia and StackExchange: CC BY-SA. CodeSearchNet (Husain et al., 2019): source files retain their repository-specific licenses; the project tooling is MIT. BERT-base-uncased

and Qwen2.5-7B-Instruct: Apache 2.0. DeBERTa-v3-large and CLIP-ViT-B/32 (Radford et al., 2021): MIT. ResNet-18: BSD-3-Clause. BLIP-2 (Li et al., 2023): BSD-3-Clause. Stable Diffusion v2.1 (Rombach et al., 2022): CreativeML Open RAIL++-M. NLLB-200-3.3B (NLLB Team et al., 2022): CC-BY-NC 4.0. LLaMA-3-8B-Instruct and DeepSeek-Coder-7B-Instruct use their respective community licenses. OpenAI, Anthropic, and Google models were accessed under their published service terms.

Three paid adult annotators independently verified every observed edge in stratified samples of $120/150/80/100$ chains from A/B/C/D. The written protocol defines the three operator classes, edge-validity decision, weak-observation flag, and adjudication procedure. Pre-adjudication Fleiss' κ is $0.81/0.78/0.72/0.74$; Section 5.1 reports the corresponding positive-edge prevalence. Annotators consented in writing, were paid approximately \$28/hour for 63 person-hours, and did not provide personal data. The task was classified as dataset curation without human-subject data.

Each retained experiment record contains terminal and split identifiers, reachable roots, observed edges, operator classes, candidate-pool and replay seeds, model snapshot hash, and target table or figure. Controlled-depth records feed the controlled-depth and ablation tables; A–D records feed the logged-pipeline and baseline-ablation tables; candidate-control records feed Tables 3, 12, and 13; and corruption records feed Tables 5 and 14. These fields fix data partitions, interventions, aggregation units, and plotted values for the retained results.

F DAG Path-Transformer (DPT) Baseline: Architecture, Training, Ablations

DPT architecture is path-based rather than whole-DAG message-passing. It ingests a (v_0, v_T, path) triple, where path $= (e_1, \ldots, e_d)$ denotes one observed reverse path from v_T back to v_0 in the logged DAG. If multiple logged paths connect the same (v_0, v_T) pair, each path is scored as a separate triple and the pair-level prediction is the arithmetic mean of the resulting path scores. Each edge token carries four features: a 64-dim learned operator-class embedding (near-lossless, stochastic-rewriting, merge, unknown), a scalar fan-in $|\text{pa}(v)|$, a scalar depth-from-terminal, and a 256-dimensional projection of the child representation. Text uses SentenceBERT-all-MiniLM-L6-v2; vision uses CLIP-ViT-B/32; cross-modal paths project modality-specific representations into the shared space. The backbone is a 4-layer Pre-LN transformer (8 heads, FFN d_{ff}=1024, dropout 0.1) with a merge-attention head masked by the fan-in structure: at merge nodes, attention is restricted to observed parents only; at linear chains, the mask is trivial. A learned `[CLS]` prepends each sequence; its final hidden state is passed through a 2-layer MLP ($256 \to 64 \to 1$) to produce the predicted $\phi^\star(v_0, v_T)$. Total parameters: 18.4M, 41% of which are embedding tables. Path truncation is fixed at $D_{\max}$=12 for all DPT runs.

DPT is trained on the same 10% held-out pilot split used to fit GIS. For each pilot terminal v_T, we compute leave-one-parent-out ground-truth loss change $\phi^\star(v_0, v_T)$ over all reachable roots by the same interventional retraining protocol used in the main evaluation (5 seeds, BERT-base surrogate). Training loss is MSE between the DPT output and $\phi^\star$, averaged over (v_0, v_T) pairs; no data beyond the pilot split is used, and the evaluation split is never touched by the DPT optimizer. Data access is matched: DPT and the graph-native learners fit ranking parameters to pilot pairs, while GIS fits one scalar per recorded operator.

The training recipe is AdamW (lr=3×10^{-4}, weight decay 0.01, β_1=0.9, β_2=0.98), a cosine schedule with 500-step linear warmup, batch size 128 (gradient-accumulated from micro-batch 32), 40 epochs, and early stopping on pilot-split Spearman ρ with patience 5. Total training cost per pipeline is ≈ 2.1 A100-hours. Five seeds are used; the seed-to-seed top-5 standard deviation is ± 2.2–2.5 across pipelines.

Table 10 stress-tests the path-based comparator within the matched budget. Learned operator embeddings add 1.9 points, merge-attention masking adds 2.4, and doubling width adds 0.4. The main comparison then adds whole-DAG graph transformer and R-GCN families, separating DPT-specific capacity from the broader structured frontier.

Table 10: DPT ablations on Pipeline A, top-5 parent recovery (%).

Configuration	Top-5 (%)
DPT (18.4M, full)	73.0 (± 2.3)
− learned operator-class embed. (one-hot)	71.1 (± 2.4)
− merge-attention mask	70.6 (± 2.5)
+ d_{model}=512 (36.8M)	73.4 (± 2.6)
+ 2× pilot-label budget	73.9 (± 2.2)

Width and pilot-label increases yield modest DPT gains, while the operator and merge ablations verify consequential use of logged semantics.

G Distractor-Pool Stress Test (K = 50, 100, 250, 500)

We re-run Pipeline A with $K \in \{50, 100, 250, 500\}$, padding reachable roots with uniformly sampled non-reachable roots. Recovery decreases with shortlist pressure, while the ordering DPT>GIS>MP>TRAK persists. The DPT–MP gap is 7.6–11.5 points as decisive roots move across the cutoff.

Table 11: Distractor-pool stress test on Pipeline A: top-5 parent recovery (%) as a function of candidate-pool size K. Method ordering is preserved at every K, providing a focused candidate-pool sensitivity check.

Method	K=50	K=100	K=250	K=500
Random ($5/K$)	10.0	5.0	2.0	1.0
TracIn	58.2 (± 3.1)	52.1 (± 3.3)	44.7 (± 3.5)	38.9 (± 3.6)
TRAK	60.1 (± 3.0)	54.4 (± 3.2)	46.9 (± 3.4)	41.0 (± 3.5)
Metadata Provenance	65.4 (± 2.7)	58.1 (± 2.9)	48.9 (± 3.1)	41.2 (± 3.2)
GIS	71.2 (± 2.5)	66.3 (± 2.6)	57.8 (± 2.8)	50.1 (± 2.9)
DPT	73.0 (± 2.3)	68.3 (± 2.5)	60.4 (± 2.7)	52.6 (± 2.8)
Δ(DPT–MP)	+7.6	+10.2	+11.5	+11.4

The DPT–MP gap varies with the cutoff, while the MP row confirms the contribution of topology after ancestry is available. Hard-negative and reachable-only controls in Tables 3 and 13 extend this result beyond uniform distractors.

H Hard-Negative and Reachable-Only Candidate Controls

Table 12 reports the reachable-root distributions used by the deterministic candidate rule. The oracle five are retained before the manifest-fixed hash is applied whenever more than 50 roots are reachable. When fewer than 50 are reachable, the reachable-only control evaluates the complete ancestor set and reports its effective size.

Table 12: Reachable roots per terminal, reported as minimum / median / 90th percentile / maximum.

Split	Reachable-root distribution	Terminals
A	8/16/31/67	2,400
B	9/19/35/72	2,100
C	11/23/44/86	1,800
D	7/18/39/79	1,500

Table 13: Split-level candidate controls. Each cell is the A/B/C/D vector of top-5 recovery / Spearman ρ. Hard negatives match source degree, content similarity, and operator profile; reachable-only evaluates ancestor ranking.

Method	Hard negatives: A/B/C/D	Reachable only: A/B/C/D
TracIn	55.1/.46; 51.8/.42; 48.4/.38; 48.9/.39	53.4/.44; 49.8/.40; 47.2/.36; 47.8/.37
TRAK	57.2/.48; 53.8/.44; 50.7/.41; 51.0/.41	55.6/.46; 51.9/.42; 49.1/.39; 49.4/.39
MP	62.1/.54; 58.8/.51; 55.9/.47; 56.1/.46	59.0/.51; 56.2/.48; 53.4/.44; 53.8/.43
GIS	68.2/.60; 65.1/.56; 60.7/.51; 59.4/.50	64.2/.56; 61.8/.53; 57.7/.48; 56.4/.47
DPT	70.0/.63; 66.9/.59; 62.3/.53; 60.9/.51	67.0/.60; 64.1/.56; 60.1/.50; 58.4/.48
R-GCN	72.3/.66; 69.7/.63; 63.8/.56; 62.5/.54	68.5/.62; 65.7/.59; 61.2/.53; 59.6/.51
Graph Transformer	73.2/.67; 70.8/.64; 65.1/.57; 63.8/.55	69.4/.63; 66.9/.60; 62.0/.54; 60.7/.52

The controls lower absolute recovery. Their central result is stable: MP remains above flat attribution, and the learned path and graph rankers remain above MP on A–D.

I Joint Top-k Recovery Definition

For Pipeline A and each method and $k \in \{1, 5, 10\}$, the evaluator ranks roots, jointly removes the method's top k roots with their descendants, replays stochastic operators, retrains BERT-base for five seeds, and measures $\Delta\ell_t^{m,k}$. Define $g_t^{m,k} = \max(\Delta\ell_t^{m,k}, 0)$ and

$$\mathrm{OracleRec}(m,k) = \frac{1}{|T_+^k|} \sum_{t \in T_+^k} \frac{g_t^{m,k}}{g_t^{\mathrm{oracle},k}}, \qquad T_+^k = \{t : g_t^{\mathrm{oracle},k} > 0\}.$$

This equal-terminal-weight summary uses the joint top-k oracle denominator. Terminals outside T_+^k are excluded from the ratio mean. The table in Section 5.2 reports this quantity only; it does not combine it with signed-loss, absolute-loss, disagreement, or escalation summaries.

J Graph-Native Implementation Context

The graph transformer uses six 256-dimensional message-passing layers, eight heads, residual edge-type bias, and terminal-conditioned root pooling; R-GCN uses three relation-specific message-passing layers. Both train on the same 10% pilot labels, candidate pools, targets, and A100 host as DPT. Their inclusion tests whether the fixed-decision protocol's structure-versus-flat comparison depends on one structured ranker family.

K Random and Targeted Corruption Sweep

Random corruption relabels uniformly sampled observed edges. The targeted attack ranks edges by a fixed combination of betweenness and source influence computed on a disjoint design split, then corrupts the highest-ranked edges. Each cell in Table 14 lists top-5 recovery / Spearman ρ at $5/10/15/20/25\%$ corruption.

Table 14: Pipeline A corruption sweep at $5/10/15/20/25\%$ corruption. Each tuple gives top-5 recovery / Spearman ρ. MP is invariant because it consumes graph structure without operator labels.

Method	Random corruption	Targeted corruption
MP	65.4/.57 at every rate	65.4/.57 at every rate
DPT	70.7/.64; 68.2/.61; 64.9/.58; 61.7/.55; 58.0/.52	68.9/.61; 64.2/.56; 59.8/.51; 55.1/.47; 51.0/.43
R-GCN	73.0/.66; 70.4/.63; 66.7/.59; 62.8/.55; 59.1/.52	70.6/.62; 65.8/.57; 60.9/.52; 56.0/.47; 51.8/.43
Graph Transformer	74.4/.68; 71.8/.65; 68.1/.61; 64.2/.57; 60.5/.53	72.1/.64; 67.0/.59; 62.2/.54; 57.4/.49; 53.0/.45