FROM DETECTION TO INTERVENTION: COUNTERFACTUAL DECONTAMINATION FOR LANGUAGE-MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmark decontamination is commonly treated as an overlap-detection problem, even though evaluation signal can enter training data through paraphrases, answer-bearing context, and reusable solution templates. We introduce Counterfactual Decontamination, an intervention-first framework that represents benchmark-to-training links as typed edges and maps them to auditable data actions. The framework combines lexical, semantic, answer, and structural evidence, then constructs matched contaminated, exact-pruned, graph-pruned, reweighted, and oracle-pruned training arms. This design separates three questions that are often conflated: whether a candidate edge can be retrieved, whether it constitutes actionable leakage, and whether removing it changes the evaluation conclusion. In controlled real-text mathematics contamination, exact matching recovers copies but misses non-exact channels, whereas the typed graph recovers paraphrase, answer, and template exposures while retaining hard benign neighbors. Across contamination levels and random seeds, exact pruning reduces but does not eliminate the target-to-clean training-effect gap; graph pruning closes 99.1% of the gap and approaches the oracle intervention under matched evaluation. Natural-pool blind validation and matched clean controls further test whether high-severity edges remain actionable beyond controlled injection. The resulting protocol replaces overlap counts with a falsifiable criterion: a decontamination policy is supported only when its detected edges survive blinded validation and its intervention preserves the model comparison the benchmark is meant to measure.

1 INTRODUCTION

Benchmark contamination is usually treated as duplicate removal. That definition is too narrow for language-model training data: a benchmark item can be exposed through a paraphrase, an answer-bearing note, or a reusable solution template without sharing a long surface span. At the same time, semantic similarity alone is too broad. Two reasoning problems can share an operation, answer, or topic while remaining legitimate independent examples. Counting either exact matches or semantic neighbors therefore leaves the central validity question unresolved: which training examples actually make an evaluation item easier?

We formulate decontamination as a counterfactual data intervention. Candidate links are retrieved with high recall, assigned typed evidence, and mapped to explicit actions that define matched training arms. A proposed leakage edge is supported only if removing its source data changes the target–clean comparison in the predicted direction relative to exact-only, random, attribution, and oracle controls. This formulation distinguishes retrieval, attribution, and causal consequence rather than treating them as one score.

At model scale, exact removal leaves 39–65% of the contaminated gap across four Qwen2.5 scales, whereas graph removal closes 86.0–98.9%. The pattern spans three reasoning tasks, four contamination channels, and two Llama-3.1 scales. Size-matched random and benign-edge placebo deletions do not reproduce it. A 1,500-edge blinded natural-pool audit with three independent annotators and adjudication tests whether typed edges remain actionable outside controlled injection.

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.

Our contributions are a typed benchmark-to-training graph separating exact duplication, semantic paraphrase, answer leakage, template exposure, and benign similarity; an intervention design with contaminated, exact-, graph-, random-, attribution-, and oracle-pruned controls plus reweighting under matched budgets; controlled and neural evidence that graph removal approaches the oracle across scale, contamination intensity, task, and model family where exact matching leaves non-exact effects; and a falsifiable rule using matched clean controls, placebos, and blinded natural-pool review to reject unsupported edges.

2 COUNTERFACTUAL DECONTAMINATION

TYPED EXPOSURE GRAPH

Let $E = \{e_i\}$ be evaluation items and $C = \{c_j\}$ candidate training chunks. We construct a bipartite graph $G = (E, C, R)$, where every edge r_{ij} carries a type, evidence trace, and severity. Edge types are *exact duplicate*, *semantic paraphrase*, *answer leakage*, *template exposure*, and *benign similarity*. Severity 3 denotes an exact or near-exact exposure, severity 2 an answerable non-exact exposure, severity 1 a template-level governance risk, and severity 0 insufficient evidence.

Candidate generation combines canonical matching, BM25 (Robertson & Zaragoza, 2009), semantic containment, and reciprocal-rank fusion. Retrieval is deliberately separated from action. Edge evidence includes canonical prompt or solution match, answer match with local cues, numeric and entity overlap, operation and template signatures, semantic token containment, and long n-gram overlap. Same-answer and same-operation neighbors remain hard negatives unless stronger answer-support evidence is present.

ACTIONS, OUTCOMES, AND CLAIM GATE

The graph induces five principal actions: retain all exposures (*contaminated*); remove exact duplicates (*exact-pruned*); remove edges at a declared severity threshold (*graph-pruned*); retain edges with severity-dependent weights (*graph-reweighted*); or remove every injected exposure by construction (*oracle-pruned*). We additionally compare size-matched random deletion, raw retrieval deletion, TracIn (Pruthi et al., 2020), and influence functions (Koh & Liang, 2017). These controls ask whether gains arise from deleting data, retrieving neighbors, or identifying actionable exposure.

For completion negative log likelihood (NLL), define

$$\Delta = \mathrm{NLL}_{\mathrm{clean}} - \mathrm{NLL}_{\mathrm{target}}, \qquad \mathrm{closure}(a) = 1 - \frac{\Delta_a}{\Delta_c}.$$

A positive Δ indicates that exposed targets are easier than matched clean items. A supported action should reduce Δ, preserve clean behavior, approach the oracle, outperform budget-matched controls, and retain precision under blinded review. Detection alone supports only a retrieval claim. We therefore report residual gaps and closures together: closure can exceed 100% under sampling noise and does not by itself imply a negative causal effect.

EVIDENCE-TO-ACTION AUDIT TRAIL

Every candidate begins as an unactioned retrieval edge. Canonical equality and long-span overlap support exact duplication; semantic containment together with aligned entities and operations supports paraphrase; answer occurrence is actionable only when its local context provides a cue linking it to the target; and structural evidence requires a reusable solution skeleton rather than a shared topic. These distinctions matter because the same retrieval score can describe an exact copy, an answer-bearing derivation, or a benign neighbor.

The action record stores the benchmark identifier, source-chunk identifier, retrievers that surfaced the pair, normalized evidence features, edge type, severity, action, policy version, and review state. The record makes two otherwise hidden choices inspectable: why one neighbor was removed while another was retained, and which evidence would have to change to reverse that decision. Graph reweighting uses the same record but maps severity to a nonzero example weight; graph pruning maps the declared threshold to zero. Oracle labels are available only for injected exposures and never enter natural-pool prediction.

The claim gate is ordered. First, target and clean sets must be matched well enough that the contaminated contrast is interpretable. Second, an action must improve the gap relative to exact and size-matched random deletion. Third, the effect must approach oracle removal without harming clean endpoints. Finally, natural-pool edges must survive blinded review and a held-out intervention. Failure at a later gate does not erase an earlier retrieval or controlled result; it narrows the claim to the last supported stage.

3 Experimental Design

Controlled Real-Text Contamination

We use public GSM8K text (Cobbe et al., 2021). One hundred held-out items define the target set. Into a 1,200-example training pool we inject 25 exact copies, 25 structured paraphrases, 25 answer-bearing chunks, and 25 template-exposure chunks, yielding 1,300 contaminated rows. Clean evaluation uses both an original holdout and a matched holdout constructed from non-target items. Matching reduces the largest continuous standardized difference from 0.225 to 0.128 and operation-signature Jensen–Shannon divergence from 0.532 to 0.184.

The controlled training-style estimator is a weighted completion n-gram model whose state is fully determined by each training arm. It enables exact intervention replay over policies, contamination levels, and seeds. This tier isolates the proposed causal contrast before the substantially costlier neural matrix; it is not used as a surrogate for neural-model behavior.

Neural Intervention Matrix

The confirmatory tier crosses Qwen2.5-Instruct (Yang et al., 2024) at 1.5B, 7B, 32B, and 72B parameters; three seeds $\{42, 137, 256\}$; five intervention arms; and low, medium, and high contamination. The resulting 180 adapter-training units use matched effective tokens, optimizer updates, adapter capacity per trainable parameter, and clean/contaminated sampling schedules. Each complete scale block is analyzed separately; cross-scale means are descriptive.

Evaluation spans GSM8K, MATH (Hendrycks et al., 2021b), and ARC-Challenge (Clark et al., 2018), with cluster-disjoint contaminated targets, matched clean holdouts, and a disjoint MMLU retention set (Hendrycks et al., 2021a). Common prompts are scored for completion NLL, exact answer, expected calibration error (Guo et al., 2017), refusal, and clean retention. A factorial 7B study crosses all three tasks with exact-copy, retrieval-leak, semantic-paraphrase, and graph-structured contamination, avoiding a task-by-channel confound. Llama-3.1-Instruct 8B and 70B (Dubey et al., 2024) provide a second-family check.

Controls and Statistical Analysis

The principal neural comparison uses the five complete arms in each scale–seed–intensity block. Three additional studies isolate alternative explanations. Size-matched random pruning deletes the same number of examples as graph pruning. The placebo study deletes high-similarity edges adjudicated as benign, edges independently confirmed as actionable, or all graph-flagged edges. The attribution study ranks the same candidate pool with TracIn and influence functions and applies the same top-k budget as the graph.

Means and standard deviations are over the three prespecified seed indices. For a single scale we use two-sided Student-t intervals with $df = 2$; these are intentionally wide. Policy comparisons use paired differences at the common seed index. Six-unit control analyses contain the complete 7B and 32B blocks, not a selection of favorable runs. We do not use the descriptive four-scale average for inference. Blind-audit proportions use Wilson intervals, and agreement is reported pairwise before adjudication. No non-significant residual gap is called evidence of exact equality; instead we state whether its interval is compatible with zero.

Method	Hit@1	MRR@50	R@10	R@50
Exact match	.250	.250	.250	.250
BM25	.970	.982	1.000	1.000
Semantic	.830	.849	.900	.940
Hybrid	.840	.872	.940	1.000
Typed graph	**1.000**	**1.000**	**1.000**	**1.000**

Table 1: Controlled edge recovery. Exact matching sees only the exact-copy quarter; typed evidence recovers all injected channels.

Scale	Contaminated	Exact	Retrieval	Graph	Oracle	Graph closure	Clean NLL change
1.5B	0.8181	0.3871	0.3812	0.1147	0.0895	86.0%	-0.0175
7B	1.2057	0.7780	0.4565	0.0491	0.0327	95.9%	-0.0304
32B	1.4579	0.8540	0.5504	0.0163	-0.0025	98.9%	-0.0048
72B	1.7692	0.6960	0.6404	0.0358	0.0198	98.0%	-0.0265

Table 2: Neural medium-contamination gaps, averaged over three seeds within each complete Qwen2.5 scale block. Clean NLL change is graph-pruned minus contaminated; negative values do not indicate degradation.

Natural-Pool Blind Audits

We retrieve candidate links between official MATH items and mathematical web text from OpenWebMath (Paster et al., 2024). Reviewers are blind to severity, retrieval rank, and intervention outcome. A historical ten-edge calibration pilot yielded 1/10 and is retained as a failed pilot rather than relabeled. After recalibrating on a disjoint set, a new ten-edge gate yielded 6/10. We then evaluate a 214-edge confirmation set and an independently sampled 1,500-edge severity-≥ 2 set. The latter is labeled by three independent annotators; 622 disagreements are resolved by a fourth adjudicator. Confirmed edges are replayed in held-out interventions rather than counted as retrieval successes alone.

Annotation disclosure. Four unpaid volunteers with prior Linguistics Olympiad participation performed the audit: three independently labeled every edge and a fourth adjudicated the 622 disagreements. They used the typed-edge rubric while blind to severity, rank, method, and outcomes; participation was voluntary after disclosure of research use, and only de-identified labels were retained. As no personal data were collected, no formal ethics review or exemption determination was obtained.

4 Results

Controlled Recovery and Intervention

The typed graph recovers all 100 injected edges, while 538 constructed benign traps produce no severity-≥ 2 false positives (Table 1). BM25 retrieves every gold edge by rank 10, but top-10 pruning also removes 544 clean chunks, giving precision 0.155. Thus candidate recall is not an intervention policy.

Under the deterministic estimator, the contaminated gap is 1.3989. Exact pruning reduces it to 0.6927, while graph and oracle pruning both produce 0.0121: 99.1% closure. On the matched clean holdout, the corresponding gaps are 1.3388, 0.6631, and -0.0177. Paired bootstrap intervals for these three values are $[0.9267, 1.7442]$, $[0.2561, 1.0721]$, and $[-0.3820, 0.3549]$. Across two contamination sizes and three seeds, exact pruning remains positive in all six runs; graph and oracle have identical mean gap -0.0892 and range $[-0.3257, 0.1052]$.

Neural Effects Persist Across Scale

Table 2 confirms the mechanism at all four model scales. Exact removal leaves gaps of 0.3871–0.8540; raw retrieval removal also leaves 0.3812–0.6404. Graph removal reduces the gap to 0.0163–0.1147 and remains close to the oracle at every scale. At 7B, the graph–exact closure difference is

Intensity	Contam.	Exact	Graph	Graph cl.
Low	0.4685	0.2166	0.0181	96.1%
Medium	1.3127	0.6788	0.0540	95.9%
High	2.2741	1.1159	0.1278	94.4%

Table 3: Dose response, descriptively averaged across four complete scale blocks. These averages are not used for inference.

Scale	Arm	Target NLL	Clean NLL	Gap	Gap SD	Closure	Clean change
7B	Contaminated	7.0108	8.2165	1.2057	0.1615	–	0.0000
	Exact-pruned	7.4244	8.2024	0.7780	0.0660	35.5%	-0.0142
	Retrieval-only	7.7589	8.2154	0.4565	0.1649	62.1%	-0.0011
	Graph-pruned	8.1371	8.1861	0.0491	0.0517	95.9%	-0.0304
	Oracle-pruned	8.1643	8.1969	0.0327	0.0407	97.3%	-0.0196
32B	Contaminated	6.9182	8.3761	1.4579	0.0175	–	0.0000
	Exact-pruned	7.4997	8.3536	0.8540	0.1832	41.4%	-0.0225
	Retrieval-only	7.8051	8.3556	0.5504	0.0652	62.2%	-0.0205
	Graph-pruned	8.3550	8.3713	0.0163	0.0879	98.9%	-0.0048
	Oracle-pruned	8.3679	8.3654	-0.0025	0.0752	100.2%	-0.0107

Table 4: Full medium-intensity neural NLL panel for the two principal scales, three-seed means. Gap SD is computed from paired per-run gaps, not by reusing a target or clean marginal SD.

61.0 percentage points (pp), with a 95% paired t-interval of $[32.3, 89.8]$ pp. The graph residual itself has interval $[-0.0795, 0.1776]$, compatible with complete removal.

The effect is stable across contamination dose. Descriptive graph closure is 96.1%, 95.9%, and 94.4% at low, medium, and high intensity, respectively; exact closure is 53.8%, 48.3%, and 50.9%. The absolute graph residual grows from 0.0181 to 0.1278 as the contaminated gap grows from 0.4685 to 2.2741, so we do not claim dose invariance. Rather, most of the induced advantage is removed throughout the tested range.

Table 3 clarifies why a residual-only test can mislead. The graph residual is larger under high contamination than under low contamination, yet its proportional closure remains above 94%. A fixed absolute threshold would judge the same policy differently as exposure dose changes; a closure-only threshold would hide the remaining absolute advantage. We therefore require both quantities and retain per-scale uncertainty.

Table 4 shows that closure is not obtained by worsening clean NLL. At both principal scales, clean changes are small and negative for every intervention shown. The target NLL moves toward its clean counterpart as increasingly complete evidence channels are removed. Graph and oracle means are nearby, but their seed variation prevents an exact-equivalence claim.

Matched Controls Identify the Active Edges

Deleting an equal number of examples is insufficient (Table 5). Random pruning closes only 17.4% of the gap, and its deficit to graph pruning remains positive across the six paired units. Removing benign high-similarity neighbors has no detectable effect, whereas removing independently confirmed edges reduces the gap. These controls rule out two simple explanations: that any deletion regularizes the model, or that semantic similarity alone determines the intervention.

The confirmatory calibration also contains a countervailing result. The contaminated gap clears its lower-bound threshold at both 7B and 32B. The fraction left after exact removal clears the declared 35% lower-bound gate at 7B (65.4%, interval $[37.4, 93.4]$) but not at 32B (58.6%, interval $[26.5, 90.8]$). We retain this failed gate. It narrows the statement from "exact removal always leaves a statistically bounded fraction" to the directly supported paired claim that graph removal improves closure over exact removal in the completed blocks.

Study	Action	7B gap	32B gap	Closure	Key paired contrast (95% CI)
Size-matched	Random-pruned	0.9976	1.2036	17.4%	Graph–random: $+80.9$ pp $[64.6, 97.3]$
	Exact-pruned	0.7780	0.8540	38.7%	Graph–exact: $+59.3$ pp $[45.2, 73.4]$
	Graph-pruned	0.0491	0.0163	97.5%	
Placebo	Benign-pruned	1.2608	1.5451	-0.4%	Benign reduction: -0.0055 $[-0.1713, 0.1602]$
	Confirmed-pruned	0.8394	0.9760	35.0%	Benign–confirmed: 0.4953 $[0.3283, 0.6622]$
	Graph-pruned	0.1524	0.1101	90.6%	

Table 5: Budget and placebo controls at medium contamination. Closure is descriptive across 7B and 32B; intervals use six paired scale–seed units.

Family/scale	Task	Contaminated	Exact	Graph	Graph closure
Qwen2.5 7B	GSM8K	1.2616	0.6206	0.0841	93.3%
	MATH	1.4127	0.8971	0.1671	88.2%
	ARC-C	0.8275	0.6713	0.0991	88.0%
Qwen2.5 32B	GSM8K	1.5980	0.6292	0.0854	94.7%
	MATH	1.7776	1.1303	0.0314	98.2%
	ARC-C	1.2486	0.6170	0.0800	93.6%
Llama-3.1 8B	GSM8K	1.1358	0.6276	0.1092	90.4%
Llama-3.1 70B	GSM8K	1.5421	0.7571	0.0019	99.9%

Table 6: Task and model-family replication at medium contamination, three seeds per row. Llama values use its native tokenizer and are compared by within-family gap and closure, not absolute NLL.

The independent calibration split reproduces the direction before the confirmatory comparison: contaminated versus oracle target NLL differs by 1.2430 at 7B and 1.5074 at 32B, positive in all three seeds at both scales. Because this split uses an independent noise realization, its absolute NLL is not substituted into the main matrix. It establishes the direction of the contrast while reserving all reported policy differences for the confirmatory blocks.

Task, Channel, and Family Robustness

Graph removal closes 88.0–98.2% across the six Qwen task–scale cells (Table 6), while exact removal leaves materially larger residuals. In the fully crossed 7B task-by-channel experiment, graph gaps range from 0.0427 to 0.1330 over all 12 cells. Exact gaps range from 0.1365 for exact-copy exposure to 0.6763 for semantic paraphrase, and graph closure relative to exact ranges from 68.7% to 91.7%. The between-cell standard deviation of graph gaps (0.0258) is smaller than the average within-cell standard deviation (0.0359). With three seeds per cell this is directional, not proof of no interaction, but it shows that the result is not produced by binding one contamination type to one task.

Task	Channel	Graph	Exact	Cl.	Task	Channel	Graph	Exact	Cl.
GSM8K	Exact copy	.0560	.1804	69.0	MATH	Exact copy	.0593	.1902	68.8
GSM8K	Retrieval leak	.0625	.5475	88.6	MATH	Retrieval leak	.1120	.6336	82.3
GSM8K	Paraphrase	.0560	.6763	91.7	MATH	Paraphrase	.1330	.6624	79.9
GSM8K	Graph structure	.0736	.4631	84.1	MATH	Graph structure	.0731	.5676	87.1
ARC-C	Exact copy	.0427	.1365	68.7	ARC-C	Retrieval leak	.0675	.3967	83.0
ARC-C	Paraphrase	.0745	.4405	83.1	ARC-C	Graph structure	.0539	.3806	85.8

Table 7: Fully crossed Qwen2.5-7B task-by-contamination study. Entries are three-seed mean gaps; Cl. is graph closure relative to exact, in percent.

Method	7B gap	7B cl.	32B gap	32B cl.
Typed graph	**.1261**	**89.5**	**.0696**	**95.2**
TracIn	.3127	74.1	.2746	81.2
Influence fn.	.2580	78.6	.3266	77.6
Random	.8245	31.6	1.3067	10.4

Table 8: Equal-budget attribution controls. Closure (cl.) is in percent.

Metric	Contam.	Exact	Graph	Oracle	Clean
Answer acc. (%)	79.4	71.9	70.6	69.8	67.2
ECE	.039	.063	.064	.075	.067
Refusal (%)	2.2	9.3	8.4	11.2	8.9
MMLU (%)	65.2	63.8	64.4	64.4	64.4

Table 9: Qwen2.5-7B behavioral endpoints, three-seed means.

Table 7 reveals the mechanism hidden by aggregate accuracy. Exact pruning is competitive only when the exposure itself is an exact copy. Its residual more than triples for semantic paraphrase on every task. Graph removal remains comparatively stable because its action is conditioned on the edge evidence, not on which benchmark supplies the target. This within-task manipulation is stronger evidence for channel robustness than comparing GSM8K, MATH, and ARC-C under different contamination mixtures.

Replication in Llama-3.1 yields the same ordering. Exact removal closes 44.7% and 50.9% at 8B and 70B; graph removal closes 90.4% and 99.9%. Absolute NLLs are tokenizer-dependent, so the cross-family conclusion concerns the intervention contrast rather than direct scale comparisons.

Attribution and Behavioral Endpoints

At the same pruning budget, the graph closes more gap than TracIn or influence functions (Table 8). Across the six paired scale–seed units, graph closure exceeds TracIn by 14.7 pp ($[3.0, 26.3]$) and influence functions by 14.5 pp ($[3.8, 25.3]$). Per-scale intervals remain wide and include zero, so we report the complete-block contrast without claiming separate significance at each scale.

The NLL effect is reflected in other endpoints at 7B. Contaminated exact-answer accuracy is 79.4%, versus 71.9% after exact pruning, 70.6% after graph pruning, 69.8% for oracle pruning, and 67.2% on the clean baseline. Refusal rates are 2.2%, 9.3%, 8.4%, 11.2%, and 8.9%, respectively. MMLU retention is 65.2%, 63.8%, 64.4%, 64.4%, and 64.4%; graph minus clean is -0.08 pp with interval $[-2.51, 2.36]$. Expected calibration error does not monotonically improve after pruning, cautioning against interpreting the contaminated model's low target-set ECE as evidence of valid generalization.

Table 9 prevents a misleading interpretation of decontamination as ordinary capability loss. The contaminated arm's target accuracy is highest and refusal rate lowest, precisely the pattern expected when evaluation items are exposed. Graph pruning moves both toward the clean and oracle arms while preserving disjoint MMLU performance. Calibration is more nuanced: memorized confidence can be well aligned with memorized target answers, so lower ECE on the contaminated target set is not necessarily healthier calibration.

Blinded Natural-Pool Validation

In the 214-edge confirmation set, 123 edges are actionable (57.5%, Wilson interval $[50.8, 63.9]$). Removing only confirmed edges closes 72.7% of a 1.1529 contaminated gap; removing all graph-flagged edges closes 93.8%. This difference quantifies the cost of using conservative human confirmation alone, while the benign placebo result shows why unreviewed similarity should not automatically trigger deletion.

In the independently sampled 1,500-edge audit, adjudication labels 807 edges actionable, a yield of 53.8% ($[51.3, 56.3]$). Pairwise Cohen's κ values are 0.463, 0.449, and 0.422 (mean 0.445); 622 edges require adjudication. The moderate agreement is itself operational evidence: indirect exposure

Stratum	n	Pos.	Rate	Corpus wt.
Web crawl	300	56	18.7%	0.62
Curated	120	11	9.2%	0.25
Synthetic	80	26	32.5%	0.13

Table 10: Stratified detector-identifiable edge sample. Wilson intervals are [14.7,23.5], [5.2,15.7], and [23.2,43.4], respectively.

Audit	n	Actionable	Result
Historical pilot	10	1	Failed calibration
Disjoint go/no-go	10	6	Gate passed
Confirmation	214	123	57.5% [50.8, 63.9]
Expanded blind	1,500	807	53.8% [51.3, 56.3]

Table 11: Natural-pool progression. Brackets are 95% Wilson intervals.

is not a mechanically obvious label, and an auditable policy needs retained evidence traces and adjudication rather than a single similarity threshold.

Finally, a separate 500-document stratified sample estimates the prevalence of *detector-identifiable edges in the studied research corpus* at 18.1% ($[14.8, 21.4]$), after weighting web-crawl, curated, and synthetic strata. This is not an estimate for the private pretraining corpus of Qwen2.5, Llama-3.1, or language models in general.

The strata in Table 10 are substantively different. Synthetic documents show the highest observed rate but have the smallest corpus weight; web crawl drives the weighted estimate. Publishing only the unweighted 93/500 sample proportion would discard the sampling design and provide no basis for transfer to a corpus with different source composition.

Table 11 distinguishes recalibration from rewriting history. The failed pilot remains visible; the positive evidence comes from new disjoint samples with a larger denominator and retained adjudication. This progression is important for interpreting the main claim: high-severity edges are useful candidates, not self-authenticating proof of leakage.

5 Discussion

What the Intervention Establishes

The experiments support a claim about benchmark validity under observed, typed source exposure. They do more than show that a detector retrieves semantically related text: the detected chunks define an action, that action changes a matched target–clean contrast, and the change approaches an oracle while surviving random, placebo, and attribution controls. The result is strongest for non-exact channels, where exact pruning leaves large residuals within every task.

The experiment does not require graph scores to be perfect probability estimates. A graph can be useful for intervention even when some natural-pool edges require adjudication, provided that its policy has acceptable yield and the downstream action is validated. Conversely, perfect recall in the controlled pool is not sufficient for deployment: BM25 already retrieves all injected edges but deletes far more benign text.

Implications for Benchmark Reporting

An overlap count should be accompanied by an intervention-sensitive result. At minimum, benchmark reports should distinguish candidate recall, reviewed edge yield, clean-set matching, effect closure, and clean retention. A model comparison is fragile when its ordering or margin changes after a supported decontamination action; it is more credible when the comparison survives graph and oracle interventions. This criterion targets the scientific conclusion rather than treating every textual neighbor as equally harmful.

The protocol also changes how negative results are used. The 32B exact-fraction gate fails, per-scale attribution intervals are wide, and annotator agreement is moderate. None is hidden. Together they bound the positive claim: graph removal consistently improves the paired intervention contrast, but exact removal's residual fraction is not tightly estimated at every scale, and natural deployment still requires human adjudication.

6 Related Work

Training-data deduplication reduces memorization and can improve model quality (Lee et al., 2022), but exact and near-exact removal does not cover every way benchmark information can be exploited. Contamination studies distinguish memorization from usable exposure (Magar & Schwartz, 2022); our intervention makes that distinction operational at the source-chunk level. Performance-based tests ask whether benchmark behavior reveals prior exposure, whereas provenance and attribution methods rank candidate sources. Counterfactual Decontamination connects these views: a retrieved source edge must support a data action, and that action must change a matched evaluation comparison before the edge is called consequential.

Data attribution asks which training examples influence a prediction. Influence functions approximate leave-one-out effects through local curvature, while TracIn aggregates gradient alignment along optimization checkpoints. These methods provide valuable general influence scores, but benchmark decontamination additionally needs a semantic account of *how* evaluation information enters a source chunk. Our equal-budget comparison treats attribution as a strong alternative policy and finds that typed evidence adds closure beyond influence ranking in the completed 7B/32B block.

Dataset documentation and benchmark audits often report provenance, deduplication, or post hoc overlap. Our framework is complementary: provenance narrows the candidate universe, typed evidence proposes actions, blinded review tests natural-pool precision, and retraining or adapter intervention tests consequence. The resulting evidence chain is longer than a detector score, but each link corresponds to a distinct failure mode that can otherwise invalidate a benchmark conclusion.

7 Limitations and Responsible Use

Three seeds per neural cell leave wide intervals, and the four Qwen2.5 scales are not independent model families; LoRA effects also need not equal full pretraining effects. Llama-3.1 adds two GSM8K scales but does not establish universality across architectures, tokenizers, or recipes. Natural-pool agreement is moderate, and prevalence is detector- and sampling-frame-specific; we make no claims about private-corpus prevalence or universal decontamination. Completion NLL is sensitive whereas answer accuracy can saturate, so the 7B panel supports convergence toward clean behavior, not preservation of every capability.

NLL and exact-answer accuracy are complementary, and candidate generation can miss sources; graph–oracle proximity therefore does not estimate the natural false-negative rate. The paired design controls seed, scale, dose, and budget, but three seeds remain small: we retain wide $df = 2$ intervals, avoid per-scale attribution claims or exchangeable-scale pooling, and leave hierarchical replication to future work.

8 Conclusion

Overlap detection is not yet decontamination. A useful policy must distinguish actionable exposure from benign similarity and show that its data intervention changes the benchmark comparison. In controlled real-text contamination, typed graph pruning closes 99.1% of the deterministic training-effect gap. In the neural matrix it closes 86.0–98.9% across four Qwen2.5 scales, remains effective across task, contamination channel, and Llama-3.1, and outperforms random, placebo, retrieval, and attribution controls without detectable clean retention loss. Counterfactual Decontamination thus replaces overlap counts with a falsifiable, auditable intervention on evaluation validity.

AI USE STATEMENT

AI assistance was used for code and manuscript-development support. The authors are responsible for all claims, references, numerical results, and artifact contents; no AI system is an author or cited as an evidence source.

REFERENCES

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

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

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

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8424–8445, 2022. doi: 10.18653/v1/2022.acl-long.577.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 157–165, 2022. doi: 10.18653/v1/2022.acl-short.18.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. OpenWebMath: An open dataset of high-quality mathematical web text. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jKHmjlpViu.

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

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

A Additional Policy Analysis

Natural-pool scope. The natural-pool intervention is conditional on the retriever's candidate universe. The reported 53.8% yield measures severity-≥ 2 candidates, not all corpus documents, and the 18.1% prevalence estimate refers only to detector-identifiable edges in the sampled research corpus. Missed exposure outside that support remains possible. Future audits should vary retrieval families, include gold-standard expert subsets, and estimate sensitivity as well as reviewed precision.

Removal can also discard legitimate related material. Every automated action should retain the edge type, evidence trace, policy version, and review state; high-impact corpus changes should be reversible. The method audits benchmark validity and should not be used to infer authorship, intent, or misconduct from text similarity.

Why Oracle Proximity Is Not the Only Target

Oracle proximity is informative in injected data because the true source set is known. In natural corpora, however, an oracle is unavailable and aggressive removal can trade residual contamination for collateral deletion. The confirmed-only natural intervention closes 72.7%, while full graph removal closes 93.8%. These are not contradictory policies. Confirmed-only removal prioritizes review precision; graph removal prioritizes effect closure. Reporting both makes the policy frontier explicit instead of embedding one universal operating point in the detector.

Reweighting provides a third option when deletion is irreversible or training coverage is scarce. In the controlled estimator, strong reweighting leaves gap 0.1578, between exact and graph pruning, while retaining more source data. This behavior is consistent with the framework: edge typing defines auditable actions, but governance constraints determine whether the action is deletion, downweighting, quarantine, or review.

Error Modes and Auditability

The observed errors fall into three policy-relevant classes. First, broad mathematical similarity produces false positive candidates: shared operations and vocabulary are common even when no answer-bearing information transfers. Second, short answers can appear coincidentally; an answer match becomes actionable only when neighboring text supplies target-specific cues. Third, template edges are intrinsically less determinate because a useful strategy can also be legitimate domain knowledge. These cases explain both the benign placebo result and the moderate natural-pool agreement.

Typed evidence makes disagreements localizable. An adjudicator can reject the answer-support feature while preserving a semantic-paraphrase judgment, or downgrade a structural edge without changing retrieval. A monolithic similarity score offers no corresponding explanation. This decomposition also supports future recalibration: threshold changes can be evaluated by edge type and action cost rather than by rerunning an opaque detector and comparing one aggregate precision number.

The audit trail is equally important after intervention. A benchmark result should record which policy generated each training arm, how many source chunks were removed or reweighted, whether the target set changed, and which clean comparison was used. Otherwise, a reported "decontaminated" score cannot be distinguished from a score produced by a different data budget or a selectively easier holdout. Our matched arms and retained policy records make that comparison explicit.

From Detection to Intervention: Counterfactual Decontamination for Language-Model Evaluation
Supplementary Material

Anonymous Submission

1 Scope, Estimands, and Reporting Rules

This supplement reports the completed experiment ledger underlying the main-paper tables. The primary outcome is the target–clean completion-NLL gap

$$\Delta = \mathrm{NLL}_{\mathrm{clean}} - \mathrm{NLL}_{\mathrm{target}},$$

and closure is $1 - \Delta_a/\Delta_c$, using the ratio of arm and contaminated means. Unless noted otherwise, neural entries are three-seed means for seeds 42, 137, and 256. Single-scale intervals use Student's t distribution with $df = 2$. Paired policy intervals are computed from within-seed contrasts. Cross-scale averages are descriptive and are not used as independent replicates. The anonymous evidence package includes the full supplied experimental ledger.

2 Neural Intervention Matrix

Table 1: Medium-contamination gaps by Qwen2.5 scale.

Scale	Contam.	Exact	Retrieval	Graph	Oracle	Graph cl.	Oracle cl.
1.5B	0.8181	0.3871	0.3812	0.1147	0.0895	86.0%	89.1%
7B	1.2057	0.7780	0.4565	0.0491	0.0327	95.9%	97.3%
32B	1.4579	0.8540	0.5504	0.0163	-0.0025	98.9%	100.2%
72B	1.7692	0.6960	0.6404	0.0358	0.0198	98.0%	98.9%

Table 2: Full 7B NLL panel, three-seed mean and SD. Target SD is computed from per-run target NLL, not copied from gap SD.

Arm	Target	Clean	Gap	Target SD	Gap SD	Clean SD
Contaminated	7.0108	8.2165	1.2057	0.1326	0.1615	0.0769
Exact-pruned	7.4244	8.2024	0.7780	0.0600	0.0660	0.0246
Retrieval-only	7.7589	8.2154	0.4565	0.1475	0.1649	0.0240
Graph-pruned	8.1371	8.1861	0.0491	0.1290	0.0517	0.0883
Oracle-pruned	8.1643	8.1969	0.0327	0.0378	0.0407	0.0366
Random-pruned	7.1899	8.1875	0.9976	0.1565	0.1618	0.0437

Table 3: Full 32B NLL panel, three-seed mean and SD.

Arm	Target	Clean	Gap	Target SD	Gap SD	Clean SD
Contaminated	6.9182	8.3761	1.4579	0.0357	0.0175	0.0242
Exact-pruned	7.4997	8.3536	0.8540	0.1702	0.1832	0.0134
Retrieval-only	7.8051	8.3556	0.5504	0.0632	0.0652	0.0336
Graph-pruned	8.3550	8.3713	0.0163	0.1000	0.0879	0.0129
Oracle-pruned	8.3679	8.3654	-0.0025	0.1295	0.0752	0.1294
Random-pruned	7.1493	8.3529	1.2036	0.2585	0.2458	0.0169

Table 4: Per-seed gaps at 7B and 32B, medium contamination.

Scale	Seed	Contam.	Exact	Retrieval	Graph	Oracle
7B	42	1.1984	0.7021	0.2660	0.0987	0.0715
	137	1.3707	0.8105	0.5531	-0.0046	-0.0096
	256	1.0480	0.8214	0.5503	0.0531	0.0362
32B	42	1.4486	1.0655	0.5053	-0.0504	-0.0611
	137	1.4781	0.7470	0.6252	0.1159	0.0822
	256	1.4471	0.7494	0.5208	-0.0166	-0.0287

For the low-intensity 1.5B block, contaminated, exact, graph, and oracle mean gaps are 0.2868, 0.0561, 0.0563, and 0.0337. The graph residual has 95% t-interval $[-0.0827, 0.1953]$. This interval is compatible with zero but does not prove equality; efficacy is assessed by paired reduction relative to contaminated and exact arms.

3 Dose, Task, and Channel Decompositions

Table 5: Descriptive dose response across the four complete scale blocks.

Intensity	Contam.	Exact	Graph	Graph cl.	Exact cl.
Low	0.4685	0.2166	0.0181	96.1%	53.8%
Medium	1.3127	0.6788	0.0540	95.9%	48.3%
High	2.2741	1.1159	0.1278	94.4%	50.9%

Table 6: Task-specific medium-contamination gaps.

Scale	Task	Contam.	Exact	Graph	Exact cl.	Graph cl.
7B	GSM8K	1.2616	0.6206	0.0841	50.8%	93.3%
	MATH	1.4127	0.8971	0.1671	36.5%	88.2%
	ARC-C	0.8275	0.6713	0.0991	18.9%	88.0%
32B	GSM8K	1.5980	0.6292	0.0854	60.6%	94.7%
	MATH	1.7776	1.1303	0.0314	36.4%	98.2%
	ARC-C	1.2486	0.6170	0.0800	50.6%	93.6%

Table 7: Fully crossed Qwen2.5-7B task-by-channel matrix. Entries are mean gap ± SD.

Task	Channel	Graph	Exact	Graph closure vs exact
GSM8K	Exact copy	0.0560 ± 0.032	0.1804 ± 0.052	69.0%
	Retrieval leak	0.0625 ± 0.059	0.5475 ± 0.037	88.6%
	Semantic paraphrase	0.0560 ± 0.043	0.6763 ± 0.042	91.7%
	Graph structured	0.0736 ± 0.016	0.4631 ± 0.020	84.1%
MATH	Exact copy	0.0593 ± 0.055	0.1902 ± 0.054	68.8%
	Retrieval leak	0.1120 ± 0.033	0.6336 ± 0.060	82.3%
	Semantic paraphrase	0.1330 ± 0.028	0.6624 ± 0.049	79.9%
	Graph structured	0.0731 ± 0.030	0.5676 ± 0.029	87.1%
ARC-C	Exact copy	0.0427 ± 0.032	0.1365 ± 0.059	68.7%
	Retrieval leak	0.0675 ± 0.029	0.3967 ± 0.014	83.0%
	Semantic paraphrase	0.0745 ± 0.030	0.4405 ± 0.013	83.1%
	Graph structured	0.0539 ± 0.044	0.3806 ± 0.015	85.8%

Across the 12 crossed cells, the graph mean gap is 0.0720 with SD 0.0400. Between-cell SD is 0.0258 and average within-cell SD is 0.0359. With three seeds per cell this provides directional evidence against a large task-by-channel interaction, not proof of invariance.

4 Random, Calibration, and Placebo Controls

Table 8: Size-matched pruning at medium contamination.

Arm	7B	32B	Descr. mean	Closure	Clean NLL change
Contaminated	1.2057	1.4579	1.3318	–	0.0000
Random-pruned	0.9976	1.2036	1.1006	17.4%	-0.0261
Exact-pruned	0.7780	0.8540	0.8160	38.7%	-0.0183
Graph-pruned	0.0491	0.0163	0.0327	97.5%	-0.0176

Paired closure differences are +80.9 pp for graph versus random ($[64.6, 97.3]$), +59.3 pp for graph versus exact ($[45.2, 73.4]$), and +21.6 pp for exact versus random ($[5.2, 38.0]$).

Table 9: Independent calibration and confirmatory gates.

Quantity	Estimate	95% interval	Declared gate	Decision
Calibration gap, 7B	1.2430	–	Direction positive in 3/3	Pass
Calibration gap, 32B	1.5074	–	Direction positive in 3/3	Pass
Confirmatory contam., 7B	1.2057	[0.8046,1.6068]	Lower > 0.15	Pass
Confirmatory contam., 32B	1.4579	[1.4144,1.5014]	Lower > 0.15	Pass
Exact fraction, 7B	65.4%	[37.4,93.4]	Lower $> 35\%$	Pass
Exact fraction, 32B	58.6%	[26.5,90.8]	Lower $> 35\%$	Fail

Table 10: Benign-edge placebo and confirmed-edge intervention.

Action	7B	32B	Descr. mean	Change vs contam.	Closure
Contaminated	1.1687	1.6262	1.3974	–	–
Benign-pruned	1.2608	1.5451	1.4030	+0.0055	-0.4%
Confirmed-pruned	0.8394	0.9760	0.9077	-0.4897	35.0%
Graph-pruned	0.1524	0.1101	0.1313	-1.2661	90.6%

The paired benign gap reduction is -0.0055 with interval $[-0.1713, 0.1602]$. The benign-minus-confirmed gap difference is 0.4953 with interval $[0.3283, 0.6622]$.

5 Cross-Family and Attribution Controls

Table 11: Llama-3.1 replication, three-seed mean ± SD.

Arm	8B gap	70B gap	8B closure	70B closure
Contaminated	1.1358 ± 0.163	1.5421 ± 0.084	–	–
Exact-pruned	0.6276 ± 0.034	0.7571 ± 0.087	44.7%	50.9%
Graph-pruned	0.1092 ± 0.071	0.0019 ± 0.047	90.4%	99.9%
Oracle-pruned	0.0166 ± 0.012	-0.0111 ± 0.041	98.5%	100.7%

Table 12: Equal-budget attribution baselines.

Method	7B gap	32B gap	7B closure	32B closure
Typed graph	0.1261 ± 0.084	0.0696 ± 0.013	89.5%	95.2%
TracIn	0.3127 ± 0.146	0.2746 ± 0.100	74.1%	81.2%
Influence function	0.2580 ± 0.071	0.3266 ± 0.124	78.6%	77.6%
Random	0.8245 ± 0.033	1.3067 ± 0.088	31.6%	10.4%

Across the six complete scale–seed pairs, graph closure exceeds TracIn by 14.7 pp ($[3.0, 26.3]$) and influence functions by 14.5 pp ($[3.8, 25.3]$). The corresponding per-scale intervals contain zero, so the main paper does not claim separate significance at either scale.

6 Behavioral and Natural-Pool Endpoints

Table 13: Qwen2.5-7B behavioral endpoints, mean ± SD.

Metric	Contaminated	Exact	Graph	Oracle	Clean
Answer accuracy (%)	79.4 ± 1.4	71.9 ± 2.0	70.6 ± 0.8	69.8 ± 3.0	67.2 ± 2.3
ECE	$.039 \pm .005$	$.063 \pm .007$	$.064 \pm .008$	$.075 \pm .008$	$.067 \pm .002$
Refusal (%)	2.2 ± 1.0	9.3 ± 2.4	8.4 ± 1.8	11.2 ± 1.1	8.9 ± 1.0
MMLU (%)	65.2 ± 0.5	63.8 ± 0.6	64.4 ± 0.4	64.4 ± 0.6	64.4 ± 1.2

Graph-minus-clean MMLU is -0.08 pp with paired 95% t-interval $[-2.51, 2.36]$. The interval contains zero but is wide at three seeds.

Table 14: Natural-pool audits and interventions.

Audit or action	n	Result	95% interval / closure
Historical pilot	10	1 actionable	Failed original gate
Disjoint go/no-go	10	6 actionable	Passed new gate
Confirmation labels	214	123 actionable	57.5% [50.8,63.9]
Confirmed-edge removal	–	Gap 0.3150	72.7% closure
Full graph removal	–	Gap 0.0709	93.8% closure
Expanded blind audit	1,500	807 actionable	53.8% [51.3,56.3]

For the expanded audit, pairwise Cohen's κ is 0.463, 0.449, and 0.422 (mean 0.445); 622/1,500 disagreements receive fourth-annotator adjudication.

Annotation disclosure. Four unpaid volunteers with prior Linguistics Olympiad participation performed the audit: three independently labeled every edge and a fourth adjudicated the 622 disagreements. They used the typed-edge rubric while blind to severity, rank, method, and outcomes; participation was voluntary after disclosure of research use, and only de-identified labels were retained. As no personal data were collected, no formal ethics review or exemption determination was obtained.

Table 15: Stratified detector-identifiable edge prevalence.

Stratum	n	Positive	Rate	Wilson interval	Weight
Web crawl	300	56	18.7%	[14.7,23.5]	0.62
Curated	120	11	9.2%	[5.2,15.7]	0.25
Synthetic	80	26	32.5%	[23.2,43.4]	0.13

The weighted estimate is 18.1% with interval $[14.8, 21.4]$. It applies to detector-identifiable edges in the studied research corpus and is not an estimate of any model provider's private pretraining corpus.

7 Statistical Tests and Claim Boundary

Table 16: Reported statistical tests. Cross-scale residual tests are descriptive unless the complete block is the declared unit.

Contrast	Method	Statistic	p
Graph residual = 0, 7B	One-sample t, $df = 2$	1.642	0.2423
Graph residual = 0, 32B	One-sample t, $df = 2$	0.321	0.7785
Graph > random closure	Paired t, $df = 5$	12.709	$5.36 \cdot 10^{-5}$
Benign change = 0	One-sample t, $df = 5$	-0.086	0.9348
Confirmed > benign	Paired t, $df = 5$	7.625	0.0006
Scale trend	OLS, $n = 4$	-0.00079/B	0.4025

The supported claim is that typed graph removal improves the paired target–clean intervention contrast over exact, retrieval-only, random, benign-placebo, and equal-budget attribution policies in the completed experiments, while

approaching the oracle and preserving measured clean endpoints. The results do not establish universal corpus prevalence, perfect natural-pool labels, exact graph–oracle equivalence, or invariance across all architectures and training regimes.