PERTURBATION ATTRIBUTION FRAGMENTATION UNDER ALIAS-REDUNDANT EVIDENCE IN KNOWLEDGE-GRAPH RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

When several retrieved graph paths express the same answer-bearing fact, a language-model reader can lose one and still answer correctly. Perturbation attribution splits credit across these paths, making the top-ranked path a poor guide to which evidence must be removed to change the answer. We study this effect in knowledge-graph retrieval-augmented generation (KG-RAG) across three corpora and four open readers, with an eight-question confirmation on a 72B reader. Controlled sweeps vary the number of alias paths for one fact while holding the question, reader, and distractors fixed. As this number grows, the top path's share of the group's credit decreases toward an equal split, and the top-ranked path changes across perturbation seeds. A deletion study uses 25 WebQSP questions with Qwen-1.5B. With four or eight equivalent paths, removing the top-ranked path changes a correct answer to an incorrect one in only 4% of cases; removing the whole group, a larger intervention, does so in 92–96%. Held-out surrogate fidelity signals controlled redundancy but does not identify a uniquely responsible path. Jointly masking known groups yields high group-level fidelity. In natural retrieval, 669 of 1,628 WebQSP rows contain multiple answer-bearing paths, which need not be semantically interchangeable, and the tested grouping detectors differ widely in accuracy. These findings support reporting path attributions together with their grouping rule and deletion unit.

1 INTRODUCTION

A KG-RAG system may retrieve several paths expressing one fact: for Jamaica's official language, "Jamaica", "JA", and "Commonwealth of Jamaica" can each lead to English. Deleting one path leaves equivalent support, making the path dispensable even when the shared fact is essential.

Perturbation explainers mask components, measure the reader's response, and fit a surrogate, following LIME (Ribeiro et al., 2016) or SHAP (Lundberg & Lee, 2017). Under exact interchangeability, symmetric attribution divides credit equally. Sampling can change which member ranks first, without identifying a uniquely responsible path.

This familiar symmetry argument leaves four empirical questions about readers and explanation units: **(Q1)** How does credit spread as the number of alias paths grows? **(Q2)** Does surrogate fidelity signal this redundancy? **(Q3)** Can deleting a known group change an answer that survives top-path deletion? **(Q4)** How common is multiple-path support in natural retrieval, and can we group it accurately?

For Q1–Q3, alias counts vary while the question, reader, and distractors remain fixed. A control adds true, non-supporting facts about the same entity to test general context dilution. We compare single-path and whole-group removals as distinct deletion budgets. For Q4, we audit naturally retrieved WebQSP paths and evaluate three grouping detectors. A smaller study on a 72B reader checks whether fragmentation persists at a larger scale.

Across six corpus–reader configurations, the top path's share of the group's credit decreases with more equivalent paths, while irrelevant additions leave credit concentrated (Q1). Fidelity signals controlled redundancy but does not identify a uniquely responsible path (Q2). In a controlled dele-

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.

tion study with one reader, removing the top-ranked path rarely changes a correct answer once several equivalents are present, whereas removing the known group, a larger intervention, usually does (Q3). Natural grouping is error-prone, and shared answer support need not imply semantic interchangeability (Q4). For correct, evidence-dependent answers, these results motivate reporting grouping rules and deletion units alongside path attribution scores.

2 Related Work

From local fit to an explanation unit. LIME learns an interpretable approximation near a prediction and evaluates its use in trust-related tasks (Ribeiro et al., 2016). SHAP places additive feature attribution in an axiomatic framework and derives a weighted-regression estimator (Lundberg & Lee, 2017). These approaches make a specified feature representation explainable; the representation still determines what receives credit. Equal credit for interchangeable inputs is consistent with that framework. Our study therefore examines the interpretation of path scores and the consequences of changing the reporting unit.

From graph retrieval to reader attribution. RoG generates graph-grounded relation plans, retrieves valid reasoning paths, and uses them to answer questions (Luo et al., 2024). KG-SMILE goes further toward post-hoc attribution: it perturbs graph components, measures response similarities, and fits weighted linear surrogates (Moghaddam et al., 2025). Its evaluation includes fidelity, faithfulness, stability, consistency, and accuracy. This provides a direct setting for our question: how should component scores be read when several components supply the same answer support? We use the path-masking surrogate design with a deterministic answer-support target. The tested ridge and Shapley-kernel estimators isolate redundancy in this design; they do not reproduce KG-SMILE's complete similarity weighting and output-comparison pipeline.

Stability and intended use. Li et al. show that small adversarial graph changes can alter a GNN explanation while preserving the prediction (Li et al., 2024). This establishes a reason to measure explanation stability separately from answer quality. Our intervention instead varies redundant verbalized evidence supplied to a language-model reader. Work on explanation utility in NLP likewise argues for evaluation tied to the intended task and demonstrates that helpfulness cannot be assumed from providing an explanation (Hashemi Chaleshtori et al., 2024). We examine one concrete use, deleting the reported evidence unit, alongside surrogate fit. This behavioral test measures the reader's dependence on the named evidence.

3 Explaining Evidence at the Right Granularity

The reader and its answer support. The reader receives a question q and verbalized graph paths $C = \{c_1, \ldots, c_M\}$. It is instructed to answer from the supplied facts and return `unknown` when they do not contain the answer. To measure the effect of removing evidence without generation noise, we compare the likelihood of the gold answer with that fallback. For a retained subset $S \subseteq C$, the support score is $y(S) = \sigma\big(\log p(\text{gold} \mid S) - \log p(\text{“unknown”} \mid S)\big) \in (0, 1)$. Here σ is the logistic function and each log probability is a sum of teacher-forced token log probabilities. Larger values indicate stronger relative support for the gold answer. We select questions that the reader answers correctly with evidence and whose support falls sufficiently when that evidence is removed. Greedy-answer matching accepts a normalized gold string occurring in the response.

What exact redundancy implies. Consider a group $E = \{c_{i_1}, \ldots, c_{i_k}\}$ whose members are interchangeable for the reader: with other evidence fixed, support depends only on whether at least one member remains. The relevant operation is thus the disjunction $\bigvee_{c \in E} \mathbf{1}[c \in S]$. In this ideal OR setting, deleting one member has no effect while another survives. A symmetric attribution assigns equal shares, $1/k$ of the group's credit. The allocation is well-defined; the behavior supplies no basis for choosing a uniquely responsible member. We use *non-identifiability* in this latter sense. Verbalizations of the same fact need not satisfy exact behavioral interchangeability in a real reader, so the experiments measure how closely their attributions exhibit the predicted pattern.

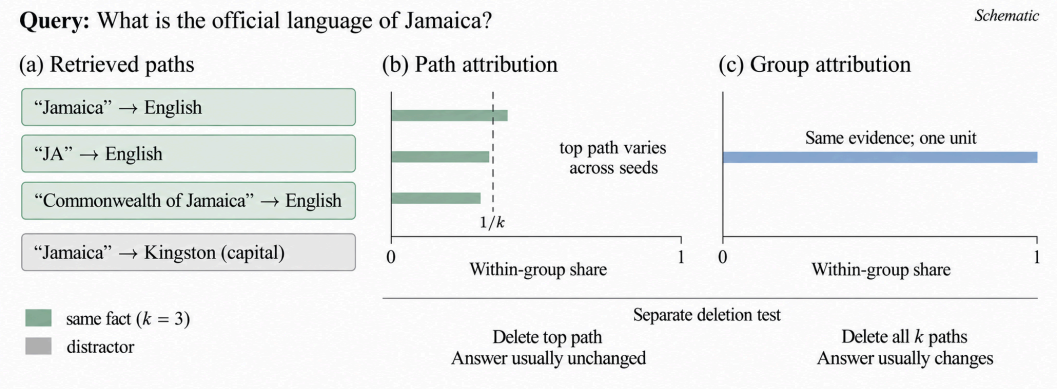

Figure 1: Conceptual illustration of alias redundancy (bar lengths are schematic). Three names express one answer-bearing fact; finite perturbation samples can rank their path-level shares differently. Joint masking changes the attribution unit without deleting the evidence, giving within-group mass 1 by construction. The separate deletion test removes one path or all equivalents; only this intervention tests answer changes, at different deletion budgets.

Estimating path contributions. For each question we repeatedly retain or remove paths, evaluate the reader's support, and fit a linear model from path-presence indicators to that score. The masks are $z \sim \text{Bernoulli}(p)^M$, so each component is retained independently with probability p. Our LIME-style estimator is ridge regression; the KernelSHAP-style estimator uses Shapley-kernel weights on the same evaluated masks. These are finite-sample implementations, and the latter's coefficients are Shapley-style estimates rather than exact Shapley values. Held-out R^2 measures the ridge surrogate's predictive fit.

Reading concentration and rank. To see whether credit is concentrated on one surface path or spread over its equivalents, we normalize absolute coefficients within the answer-support group. For coefficients w, the *top-path mass* is $\max_{i \in E} |w_i| / \sum_{j \in E} |w_j|$. It equals 1 when one member carries all within-group credit and $1/|E|$ for an equal split. It does not measure the group's total importance relative to distractors. Normalized attribution entropy summarizes the same distribution's diffuseness on $[0, 1]$. Cross-seed top-1 rank instability is the fraction of perturbation-seed pairs that select different top members. This distinguishes shared credit from the reliability of a particular ranking.

Reporting a group. When paths are known to express interchangeable support, we can ask for their joint contribution directly. We concatenate their verbalizations into one component and retain or remove all of them together during perturbation. The surrogate then reports one coefficient for that group. We call this operation *alias canonicalization*; it changes the attribution unit while keeping the reader and estimator. A one-member group has top-path mass 1 and no internal rank instability by construction. The empirical checks are whether the resulting surrogate predicts reader support and whether removing the group changes the answer.

The controlled study supplies the groups from construction keys: a head entity's alias set and a relation-template family specify the verbalizations of one fact. Natural retrieval requires estimating the partition. A false split leaves redundant paths in different groups; a false merge combines evidence that should remain distinct. We evaluate detectors for this step separately from the known-group experiment, so success given a partition is not confused with recovering that partition.

3.1 An analytic reference for masking redundant support

The ideal OR case also quantifies how masking changes the information available to a surrogate. Hold other evidence fixed and let $Y = 1 - \prod_{i=1}^{k}(1 - Z_i)$, with independent $Z_i \sim \text{Bernoulli}(p)$. Here Y is an ideal binary support function, distinct from the reader's measured gold-versus-unknown score. Complete removal has probability

$$q = (1-p)^k, \qquad \mathbb{E}[Y] = 1 - q, \qquad \text{Var}(Y) = q(1-q). \tag{1}$$

Thus a fixed masking budget observes fewer unsupported subsets as multiplicity increases. With the main protocol's $p = 0.5$ and $N = 140$, the expected numbers of complete-removal masks are 70, 35, 8.75, and approximately 0.547 at $k = 1, 2, 4, 8$. For the 42 held-out masks used by ridge, the probability of never observing complete removal at $k = 8$ is $(255/256)^{42} \approx 0.848$. These values follow from the ideal reference; they do not estimate the frequency of undefined fidelity in a real reader, whose support can vary even when an alias survives.

For population least squares with an intercept and no ridge penalty, independence makes the coefficient of each Z_i equal to

$$\beta_i = \frac{\mathrm{Cov}(Z_i, Y)}{p(1-p)} = (1-p)^{k-1}. \tag{2}$$

Indeed, $Z_i Y = Z_i$, so the covariance is $p(1-p)^k$. All k coefficients are equal and positive, giving normalized top-path mass $1/k$. The explained variance of their linear projection yields

$$R^2_{\mathrm{pop}} = \frac{kp(1-p)^{k-1}}{1-(1-p)^k}. \tag{3}$$

At $p = 0.5$, this becomes $k/(2^k - 1)$: perfect at $k = 1$, but only $8/255$ at $k = 8$. A symmetric additive allocation and low linear fidelity therefore coexist even with infinitely many masks. The finite ridge fits additionally face regularization, sampling variation and the reader's deviations from exact OR behavior; Equation 3 is not a prediction of their reported R^2.

Grouping changes the mask distribution as well as the reporting unit. A known alias group is removed with probability $1-p$ when treated as one component, instead of $(1-p)^k$ under independent path masking. In the ideal reference, its support is itself linear in the group indicator. This explains why grouped fidelity and per-path fidelity answer different approximation questions. A group deletion is also a larger intervention in raw paths. Reporting both the partition and its raw removal size is therefore necessary for interpreting improved deletion validity; normalization to a single coefficient is not an independent demonstration of localization accuracy.

The grouped coefficient is not generally the sum of coefficients from the original fit. In the ideal reference, summing the k ungrouped coefficients gives $k(1-p)^{k-1}$, whereas fitting the binary support to its group indicator gives coefficient 1. At $p = 0.5$ and $k = 8$, these are $1/16$ and 1. The difference comes from fitting under different perturbation distributions, not from recovering additional gold evidence. It prevents interpreting a larger group coefficient as an improvement on a fixed numerical attribution scale. The comparable behavioral quantity is the reader's support change under an explicitly specified removal, including how many raw paths that removal contains.

What more masks can resolve. The rare complete-removal event separates finite sampling from structural non-identifiability. To observe it at least once with probability $1-\alpha$ under independent masks requires

$$N \geq \left\lceil \frac{\log \alpha}{\log(1-(1-p)^k)} \right\rceil. \tag{4}$$

For $p = 0.5$, $k = 8$ and $\alpha = 0.05$, this lower sampling requirement is 766 masks. It is only a requirement for seeing one informative absence, not for estimating a stable surrogate or a deletion effect. Increasing the budget can reduce coefficient estimation error, but the population coefficients in Equation 2 remain equal. More observations therefore cannot select one truly privileged member of an exactly interchangeable group.

Rank variation has a corresponding symmetry reference. If independent repeats select a top member uniformly among k equivalent paths, their disagreement probability is $1 - 1/k$, or 0.875 at $k = 8$. This follows by summing the k equal probabilities of selecting the same member twice. It assumes exchangeable selection and independent repeats; fixed tie-breaking, shared masks or genuine reader preferences can change it. A deterministic tie-break can even give zero rank variation while still conveying no evidence that its selected member is uniquely responsible. Thus concentration, rank variation and deletion behavior must be interpreted together rather than assigning a universal failure threshold to any single statistic. The controlled sweep measures these quantities with the existing mask budgets; neither reference changes the reported experimental protocol or supplies additional reader measurements.

4 EXPERIMENTS

The controlled study asks whether adding alternative expressions of one fact spreads its attribution without making that fact dispensable. One trial takes a selected question, fixes its reader and distractors, and varies the number $k \in \{1, 2, 4, 8\}$ of answer-equivalent verbalizations. We estimate path credit at each k, repeat the masks to measure rank variation, and remove either one equivalent or the entire group to measure evidence dependence. Irrelevant facts provide a comparison for concentration; grouping changes the reported deletion unit. Natural-retrieval experiments then separate how often multiple paths occur from how accurately detectors group them.

Data and readers. We use WebQSP (Yih et al., 2016) and ComplexWebQuestions (CWQ; Talmor & Berant, 2018) with Freebase subgraphs from RoG (Luo et al., 2024), plus a Wikidata-native set built from real statements and aliases. On WebQSP, Qwen2.5-0.5B/1.5B/7B-Instruct and Phi-3.5-mini-Instruct retain $26/30/24/26$ questions, respectively. Qwen-1.5B retains 30 questions on each other corpus. The controlled aliases combine real head-entity names with relation surface templates for the same fact. Selection requires a correct evidence-present answer and a support decrease of at least 0.1 on removal. The reported answer-match rate at $k{=}1$ is 1.0; the main attribution sweep records support scores, rather than greedy-answer accuracy at every k.

Sampling and aggregation. Each controlled condition uses $N{=}140$ masks at $p{=}0.5$ and five perturbation seeds. The ridge surrogate fits the first 70% of masks and uses the remaining 30% for fidelity; the Shapley-kernel fit uses all evaluated masks. We average within a question over seeds, then across questions. Reported 95% confidence intervals bootstrap questions, preserving the question as the comparison unit. The $k{=}1$ condition provides the single-path reference. The 72B confirmation and deletion experiments have the separate budgets stated below.

The irrelevant-fact comparison. To distinguish repeated answer support from additional context, we also construct a condition with one answer-bearing path and four other real relations of its head entity. These facts are non-contradictory and do not support the answer; the same background distractors remain. We compare the largest coefficient share within these five candidates with top-path mass among four equivalent paths. This is a concentration comparison with different group sizes. Its maximum-share statistic does not establish that the highest-weight candidate is the gold path.

HOW CREDIT FRAGMENTS

Top-path mass decreases monotonically with alias multiplicity in all six corpus–reader configurations. Figure 2 shows the WebQSP curves: the largest share falls from 1.00 at $k{=}1$ to 0.20–0.27 at $k{=}8$, with Qwen-1.5B and Qwen-7B both near 0.22. The corresponding CWQ and Wikidata values are 0.25 and 0.19. The curves move toward the equal-share $1/k$ reference, although finite estimates remain above it. For the two WebQSP readers in Table 3, rank instability rises from zero to 0.48–0.52 already at $k{=}2$; Wikidata reaches 0.81 at $k{=}8$. Thus the shared fact's credit is divided, and the identity of the largest share varies across samples.

The irrelevant-fact condition retains much greater concentration than the equivalent-path condition (Table 1). Its largest share is 0.76–0.92, compared with 0.30–0.40 among four equivalents; every question-bootstrap interval for the difference excludes zero. This supports a distinction between the tested irrelevant additions and repeated answer support.

Individual removal and group dependence. For each question at $k{=}8$, we measure the support drop after removing each equivalent path and after removing them all. Table 2 reports the mean of the within-question maximum single-removal effects, alongside the mean whole-group effect. Across the six configurations, mean maximum single-removal effects range from 0.000 to 0.125, whereas group removal lowers support by 0.809–0.978. The larger single-removal effect for Qwen-0.5B indicates a departure from exact interchangeability. Overall, the results show approximate substitution: individual removals have much smaller average effects than group removal.

Cheap-signal diagnostics. Question–path embedding similarity and path frequency assign only 0.11–0.29 of their candidate-group mass to the gold path among irrelevant facts. These signals

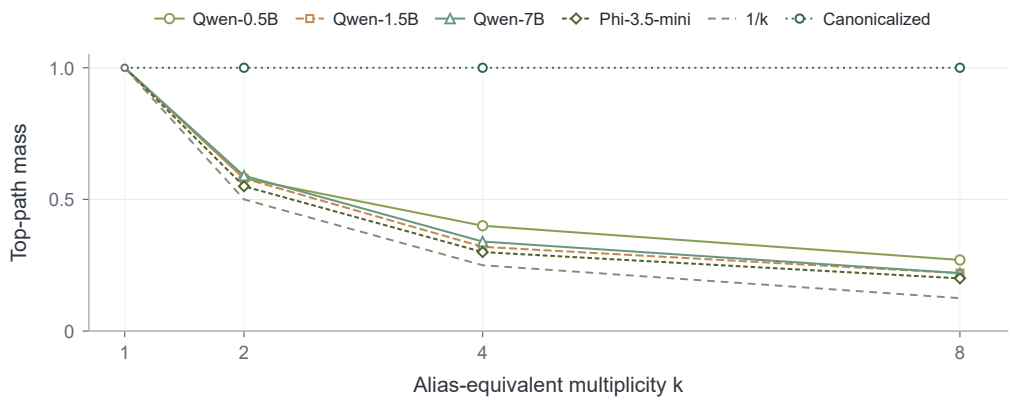

Figure 2: Credit spreads across equivalent paths as their number grows (WebQSP, LIME-style ridge attribution). Points are mean largest shares within the known answer-equivalent group. Dashed $1/k$ is the equal-share reference; the canonicalized curve is the one-unit normalization after joint masking. All four readers show fragmentation, with the same underlying answer-bearing fact across conditions.

Table 1: Concentration under equivalent and irrelevant additions. "Equiv." is the largest share among four equivalent paths; "irrel." is the largest share among one gold and four irrelevant paths. The gap is irrel. minus equiv., with a 95% question-bootstrap interval. These are group maxima, not verified shares on the gold path.

Corpus	Reader	equiv. (k=4)	irrel.	gap [95% CI]
WebQSP	Qwen-0.5B	0.40	0.87	0.47 [0.37, 0.55]
WebQSP	Qwen-1.5B	0.32	0.84	0.52 [0.43, 0.59]
WebQSP	Qwen-7B	0.34	0.92	0.59 [0.50, 0.66]
WebQSP	Phi-3.5-mini	0.30	0.81	0.50 [0.41, 0.59]
CWQ	Qwen-1.5B	0.36	0.76	0.40 [0.31, 0.49]
Wikidata	Qwen-1.5B	0.30	0.92	0.62 [0.57, 0.66]

therefore provide weak gold-path concentration in this control. Their gold-share metric differs from the perturbation maximum-share metric in Table 1, so these results do not provide a matched comparison of gold localization.

Estimator and sampling checks. Fragmentation also appears with the Shapley-kernel estimator. On CWQ with Qwen-1.5B, its top-path masses are $1.00/0.57/0.37/0.30$ at k=$1/2/4/8$, compared with $1.00/0.57/0.36/0.25$ for ridge attribution. An ideal monotone-OR oracle reproduces the equal-share pattern independently of a language model, with finite estimates near the closed-form Shapley share $1/k$. Changing the mask-retention probability to p=0.3 gives the WebQSP/Qwen-1.5B sweep $1.00/0.55/0.32/0.22$, close to the p=0.5 result. These checks locate the effect in redundant support rather than one tested estimator or mask rate.

A larger-reader confirmation. Qwen2.5-72B-Instruct-AWQ shows the same trend on eight evidence-dependent WebQSP questions, with three seeds and N=100 masks per condition. Mean top-path mass is 1.000, 0.685, and 0.334 at k=$1, 2, 4$; the corresponding 95% question-bootstrap intervals are $[1.000, 1.000]$, $[0.590, 0.806]$, and $[0.313, 0.363]$. The k=1 to k=4 decrease is 0.666 $[0.637, 0.687]$, while mean rank instability rises from 0.000 to 0.333 and 0.583. All eight concentration and instability measurements are defined. One k=4 question has a constant support target, making its R^2 undefined; we retain that missing value. This small slice confirms fragmentation at a larger scale but has no matched irrelevant-fact control.

Table 2: Support-score decrease at k=8. Max LOO is the mean over questions of each question's maximum single-path leave-one-out effect. LAEO is the mean leave-all-equivalents-out effect. Values are differences on the $[0, 1]$ support scale, not answer-flip rates.

Corpus	Reader	max LOO	LAEO
WebQSP	Qwen-0.5B	0.125	0.809
WebQSP	Qwen-1.5B	0.011	0.889
WebQSP	Qwen-7B	0.072	0.940
WebQSP	Phi-3.5-mini	0.000	0.978
CWQ	Qwen-1.5B	0.039	0.811
Wikidata	Qwen-1.5B	0.000	0.941

Table 3: Fidelity, top-path mass, and seed-pair rank instability on WebQSP. Fragmentation is already apparent at k=2 and increases with multiplicity. At k=8, fidelity means cover only defined scores; concentration and rank statistics remain reported separately.

Reader	metric	k=1	k=2	k=4	k=8
Qwen-1.5B	fidelity R^2	0.93	0.48	0.13	−0.30
Qwen-1.5B	top-path mass	1.00	0.58	0.32	0.22
Qwen-1.5B	rank instability	0.00	0.52	0.56	0.67
Qwen-7B	fidelity R^2	0.96	0.53	0.19	−0.37
Qwen-7B	top-path mass	1.00	0.59	0.34	0.22
Qwen-7B	rank instability	0.00	0.48	0.51	0.63

What fidelity tells us

Fidelity asks whether the surrogate predicts support on held-out masks. It can therefore reveal that an additive model fits repeated, substitutable evidence less well than a single support path. Across the six configurations, mean R^2 falls from 0.89–0.97 at k=1 to 0.48–0.61 at k=2. At the question level, R^2 separates those two conditions with pooled AUC 0.99 for the WebQSP Qwen-1.5B and Qwen-7B readers. With a clean reference, it is thus a useful indicator of the controlled change in redundancy.

Fidelity and a path's deletion effect answer different questions. Table 3 shows that at k=2 the top path changes across about half of seed pairs while the surrogate still explains some held-out variation. At k=4, Qwen-1.5B has $R^2 = 0.13$ while its largest within-group share is 0.32. Predicting support with positive R^2 does not select a uniquely responsible equivalent path. The exact symmetry argument already rules out that selection even for a predictor with perfect knowledge of the support function; the empirical sweep additionally shows fit deteriorating with multiplicity.

At larger k, independently sampled masks rarely remove every equivalent, so support can become nearly constant. Fidelity is then undefined rather than evidence of a good explanation. Only 25 of the 54 pooled WebQSP question scores are defined at k=8; the negative means in Table 3 use the defined scores. The implementation treats held-out target sums of squares below 10^{-6} as undefined and clips defined R^2 to $[-1, 1]$. Moderate multiplicities therefore provide the clearest comparison of fit and fragmentation.

Grouping and deletion behavior

Joint masking of the known alias group produces a surrogate with fidelity 0.92–0.97 in the configurations shown in Table 4. Its single group coefficient removes the arbitrary ordering among member paths. The accompanying top-path mass of 1.00 and zero within-group rank instability follow from using one unit; the fidelity result is the measured evidence that the grouped representation is easier for this surrogate to fit.

What a reported removal changes. The controlled deletion study uses 25 WebQSP questions with Qwen-1.5B. For each question and multiplicity, it averages absolute raw attribution coefficients over three seeds, removes the highest-ranked path, and checks whether a previously correct greedy

Table 4: Attribution after jointly masking the known equivalent group at k=8. Within-group mass is 1.00 by construction; held-out fidelity measures the fit on the new group-level masks.

Corpus	Reader	top-path mass	fidelity R^2
WebQSP	Qwen-1.5B	1.00	0.93
WebQSP	Qwen-7B	1.00	0.93
WebQSP	Phi-3.5-mini	1.00	0.92
CWQ	Qwen-1.5B	1.00	0.93
Wikidata	Qwen-1.5B	1.00	0.97

Table 5: Grouping detectors on controlled exact-alias pairs and natural answer-support pairs. Precision/recall/F1 use thresholds fitted on separate controlled data. FM/FS count false merges/splits among natural pairs. ARI is adjusted Rand index; V is V-measure, combining cluster homogeneity and completeness. Both are means over 200 natural questions. Natural labels group answer-bearing paths and do not certify exact semantic interchangeability.

	Controlled		Natural pairs						Natural clustering	
Detector	AUC	F1	AUC	Prec.	Rec.	F1	FM	FS	ARI	V
MiniLM cosine	0.988	0.944	0.990	0.989	0.684	0.809	10	408	0.767	0.916
Bidirectional NLI	0.924	0.852	0.573	0.427	0.283	0.340	489	926	0.179	0.470
Qwen2.5-7B judge	0.344	0.668	0.413	0.393	0.975	0.560	1944	32	0.000	0.000

answer becomes incorrect under the answer matcher. This validity rate falls from 0.96 at k=1 to 0.04 at both k=4 and k=8. The comparator directly removes the known merged answer group; it does not learn a second selector. Its validity is $0.96/0.92/0.96$ at k=$1/4/8$. The comparison demonstrates the consequence of deleting one path or all equivalent support, with different deletion budgets.

The group-removal diagnostic in Table 2 makes the underlying intervention explicit in the raw representation: remove the entire set of equivalents. Under the ideal OR condition, every surviving member can still support the answer, so removing all members is necessary to eliminate that group's support. The empirical removal effects support this mechanism approximately. They do not identify a unique smallest removal set for arbitrary retrieved evidence. Grouping is useful here because it names the shared support as the unit under examination.

Natural retrieval and automatic grouping

How often multiple paths occur. We scan both WebQSP retrieval parquet shards without injecting aliases. A path is eligible when its head is a question topic entity and its tail is a gold answer; duplicate normalized triples count once. All 1,628 rows are parseable, 1,065 contain such answer support, and 669 contain at least two distinct answer-bearing paths. This is 41.1% of scanned rows and 62.8% of answer-supported rows. A fixed cap retains the first 500 eligible examples with at least three distractors. Of these, 167 have two answer-bearing paths, 80 have three, 48 have four, and the largest count is 92. These denominators measure multiple answer-bearing paths, which can differ in relation or provenance; they do not establish exact semantic equivalence.

Reader behavior on natural paths. The earlier natural-reader studies are smaller than the mining audit. A 36-question Qwen-1.5B attribution slice has mean top-path mass 0.61 and instability 0.22. In the grouping/deletion evaluation, the Qwen-1.5B and Qwen-7B slices contain 36 and 26 questions, respectively, with five seeds and N=120 masks. Raw top-path removal has validity 0.03 and 0.08. MiniLM grouping at the evaluated cosine threshold τ=0.80 raises these rates to 0.50 and 0.77, below known-answer-group comparators of 0.86 and 1.00. These reader outcomes concern those slices, not all 500 mined examples. The detector audit below uses separate threshold calibration.

The detector comparison. The grouping task asks whether two verbalized paths belong in one support group. Controlled positive pairs are constructed exact aliases. Natural positive pairs are two topic-anchored answer-bearing paths; negatives pair answer-bearing and distractor paths. Thus

the natural labels define a broader answer-support group. We select thresholds on 2,376 controlled calibration pairs and evaluate on 1,624 balanced controlled held-out pairs and 3,291 natural pairs (1,291 positive, 2,000 negative). Stable question hashes separate controlled calibration and held-out questions. Natural labels never tune a threshold. We also score the resulting clusters on 200 natural questions.

We compare MiniLM cosine similarity, the minimum of forward and reverse entailment scores from DeBERTa-v3-large, and a Qwen2.5-7B prompt judge. Every detector sees the full head–relation–tail verbalization. Thresholds maximize calibration F1; single-linkage joins paths connected by an above-threshold score. Table 5 reports pair discrimination and classification, false merges and splits, and agreement with the labeled clusters. Because single-linkage connects transitively, a false-positive pair can join larger groups. The natural test therefore includes clustering as well as pairwise accuracy.

MiniLM has the highest F1 and clustering agreement among the tested detectors, with F1 0.944 on controlled pairs and 0.809 on natural pairs. Bidirectional NLI falls from 0.852 to 0.340, missing many positive natural pairs. The prompt judge's natural recall of 0.975 comes with 1,944 false merges and below-chance AUROC; its F1 of 0.560 conceals poor grouping. These results favor the tested embedding detector for this label definition. The much lower NLI and prompt-judge performance on these same labels makes task-specific detector validation necessary before grouping paths.

5 Discussion and Limitations

An explanation's unit determines what a deletion test means. For exactly interchangeable paths, symmetric credit shares are meaningful allocations, but their ordering cannot reveal a uniquely responsible surface form. Reporting the group aligns the named unit with the shared support whose removal is tested. In practice, this calls for reporting the partition and deletion budget alongside attribution concentration, rank variation, and fidelity. A concentrated group score alone is not evidence that the partition is correct.

The exact OR argument is a conditional mechanism, while the reader experiments use surface variants of the same fact and measure their behavior. Natural paths can share an answer yet carry different qualifiers or provenance. Merging such paths may remove genuinely distinct evidence. Our natural detection labels do not resolve this partial-overlap regime, and shared tail text can make the grouping task easier. Threshold transfer from controlled aliases to natural paths is therefore an observed limitation, not a solved semantic-equivalence problem.

The results concern selected correct, evidence-dependent answers under a gold-versus-unknown support target. Hallucinations, parametric-knowledge-dominated answers, adversarial distractors, and white-box explanation methods need separate study. The natural reader slices and eight-question 72B confirmation limit external generality. More perturbation samples may reduce rank noise, although no sample budget can make a unique path identifiable under exact behavioral symmetry.

Code and data availability. The anonymous supplement provides core code for reader scoring, perturbation attribution, oracle checks, and data preparation, with dependency and execution instructions. This minimal release excludes datasets, model weights, historical result files, and separate prompt-judge and additional-control drivers.

6 Conclusion

Alias redundancy can spread one fact's attribution across several KG-RAG paths and make their top rank unstable. Equal credit is compatible with a correct symmetric allocation; it does not identify a uniquely responsible path. Our controlled results connect this fragmentation to group dependence and show why fidelity and deletion behavior must be read separately. Joint masking makes the evidence group explicit, while the natural audit shows that recovering a useful group is itself an empirical task. Explanations become easier to interpret when the reported unit, its detection errors, and the intervention it supports are stated together.

STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

Generative AI tools assisted with manuscript refinement, language clarity and readability, and LaTeX formatting. The authors are responsible for the final manuscript.

REFERENCES

Fateme Hashemi Chaleshtori, Atreya Ghosal, Alexander Gill, Purbid Bambroo, and Ana Marasovic. On evaluating explanation utility for human-ai decision making in nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Jiate Li, Meng Pang, Yun Dong, Jinyuan Jia, and Binghui Wang. Graph neural network explanations are fragile. In *ICML*, 2024.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *ICLR*, 2024.

Zahra Zehtabi Sabeti Moghaddam, Zeinab Dehghani, Maneeha Rani, Koorosh Aslansefat, Bhupesh Kumar Mishra, Rameez Raja Kureshi, and Dhavalkumar Thakker. Explainable knowledge graph retrieval-augmented generation (kg-rag) with kg-smile. *arXiv preprint arXiv:2509.03626*, 2025.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *KDD*, 2016.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In *NAACL*, 2018.

Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. The value of semantic parse labeling for knowledge base question answering. In *ACL*, 2016.