future internet

MDPI

Article

PQ-EID: Provenance-Quotient Evidential Interaction Decomposition and Certification in Web Question Answering

Khrystyna Zub [1], Ivanna Dronyuk [2,*] and Oksana Mulesa [3,4]

[1] Department of Social Communication and Information Activities, Lviv Polytechnic National University, 12 Bandery St., 79013 Lviv, Ukraine; khrystyna.v.zub@lpnu.ua
[2] Department of Mathematics and Informatics, Jan Dlugosz University in Czestochowa, 42-217 Czestochowa, Poland
[3] Division of Mathematics, Department of Physics, Mathematics and Technologies, Faculty of Humanities and Natural Sciences, University of Prešov, 17. Novembra St. 1, 080 01 Prešov, Slovakia; oksana.mulesa@unipo.sk
[4] Department of Software System, Uzhhorod National University, 88000 Uzhhorod, Ukraine
* Correspondence: i.dronyuk@ujd.edu.pl

Abstract

Web question answering often relies on passages that repeat the same source, offer independent support, or affect an answer only when considered together. We present PQ-EID, a method for examining how evidence interacts while accounting for its provenance. The method groups passages by claim and source information, evaluates evidence coalitions, and uses Möbius decomposition to identify individual and joint contributions. It also assesses whether an interaction keeps the same sign across plausible provenance groupings, subject to explicit bounds on evaluation error. Controlled experiments recovered known interactions and showed how mistaken source grouping can change their interpretation. A 10-question Question Answering with Conflicting Contexts (QACC) pilot revealed sensitivity to duplicated passages, although raw-context counting gave better predictive scores. We also evaluated two instruction-tuned language models on 60 QACC questions. Their coalition scores showed substantial higher-order interactions and sensitivity to the order of evidence. A separate controlled study of 80,000 trials examined when the conditional sign certificate succeeds and how it can fail if the true provenance structure is excluded. In the simulation with uncalibrated partition weights, 3.3–9.6% of the issued signs were wrong in the regimes where signs were issued. The real-data experiments use domain- and claim-based source proxies, without verified derivation chains. PQ-EID is therefore presented as a method for auditing evidence structure; its statistical guarantees apply under the stated structural and evaluation assumptions.

Keywords: web question answering; knowledge-enhanced web intelligence; semantic web; knowledge graphs; data provenance; evidential interactions; explainable artificial intelligence; retrieval-augmented generation

Check for updates

Academic Editors: Alia Abdelmoty, Usashi Chatterjee and Michael Sheng

Received: 29 September 2026
Revised: 21 October 2026
Accepted: 23 October 2026
Published: 25 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Web question answering systems increasingly use search engines, language models, and structured knowledge. This gives them access to recent information, but it also exposes them to duplicated and conflicting web content. The same fact may appear on many sites because those sites copy a common report. Other sources may confirm it independently. QACC found conflicting contexts for about one quarter of unambiguous open-domain

questions retrieved through Google Search [1]. Reliable evidence analysis therefore needs to consider both what a passage says and where its information comes from.

Data provenance standards provide a formal way to represent this structure. PROV-O defines classes and relations for entities, sources, and derivation links in web applications [2]. For question answering, this representation also needs a semantic link to the claims that support particular answer candidates. Two passages from the same domain may contain different facts, while the same claim from two independent primary sources has a different evidential meaning from two copies of one report. The unit of analysis should therefore combine claim semantics with provenance information.

Web evidence must be connected to its claims and sources before we can assess how passages work together. Removing obvious duplicates is helpful, but it does not remove every dependence. For example, two passages may be weak on their own yet jointly support an answer. Other passages may repeat a claim or support competing answers. Shapley values assign a contribution to each source [3,4]. To inspect joint effects, we also need to identify which combinations of passages produce them.

A further difficulty comes from errors in the provenance description itself. In real web data, source identity, semantic equivalence between claims, and derivation chains are often only partly known. An incorrect merge of two independent pieces of evidence can hide an interaction, while an incorrect split of repeated evidence creates artificial game elements and artificial interactions. A statistically stable coalition estimate under a fixed canonicalization can therefore have a valid numerical error bound while still describing the wrong evidence structure.

We therefore examine three questions: how to define evidence units when sources may be dependent, how to identify joint effects between those units, and how to assess whether those effects remain stable when provenance information is uncertain.

This study develops PQ-EID (Provenance-Quotient Evidential Interaction Decomposition) to audit dependent evidence in web question answering. PQ-EID constructs evidence units from claims and provenance, evaluates changes in support for competing answers, identifies non-additive interactions, and assesses the stability of their signs under stated uncertainty assumptions.

The main contributions are as follows:

1. A contrastive evidence game is defined over canonical claim–provenance units. This avoids treating the number of retrieved passages as the number of independent pieces of evidence and provides a common representation for content and provenance.
2. Möbius decomposition is used to retain irreducible joint evidential effects on the exact subsets where they arise. This separates individual support from synergy, saturation, substitutability, and conflict between evidence units.
3. A conditional dual sign certificate is formulated to combine coalition-evaluation error with uncertainty in provenance structure under explicit coverage assumptions. It separates an effect that is statistically stable under one fixed partition from an interaction whose sign remains stable across admissible provenance structures.

The remainder of the paper is organized as follows. Section 2 reviews work on knowledge-enhanced web intelligence, provenance-aware knowledge representation, dependent and conflicting evidence, and attribution and interaction analysis. Section 3 formalizes the problem and introduces the concepts used to construct the evidence game. Section 4 presents PQ-EID, including evidence canonicalization, interaction decomposition, provenance uncertainty, and dual sign certification. Section 5 describes the experimental methodology, Section 6 reports the results, Section 7 discusses their implications and future research directions, and Section 8 concludes the paper.

2. Related Work

Research on web evidence spans knowledge graphs, provenance, source dependence, and interaction analysis. These areas answer related questions: how information is represented, how sources are connected, and how groups of evidence affect an output. We review them in that order to identify what is still needed for an audit of joint evidence effects.

2.1. Knowledge-Enhanced Web Intelligence and Semantic Provenance Representation

Knowledge graphs provide explicit representations of entities, relations, identity, and context and are widely used to integrate heterogeneous sources [5]. Question answering over knowledge bases combines natural language processing, retrieval, and semantic technologies to map a user query to structured knowledge [6]. Recent work on integrating large language models with knowledge graphs treats structured facts as external support for generation and reasoning [7].

For open web sources, semantic content should be accompanied by provenance information. PROV-O provides a standardized vocabulary for entities, agents, activities, and derivation relations, and offers a basis for explicit provenance representation in web applications [2]. These mechanisms can record sources, provenance chains, and links between information objects, which is important for integrating heterogeneous web data and preserving traceability.

PROV-O records the origins and derivations of information, but those records alone do not tell us how several passages jointly affect an answer. Source links can also be incomplete. An incorrect merge may conceal an interaction, while an incorrect split can create artificial evidence units. A useful audit must therefore combine source information with an analysis of joint effects and of uncertainty in the grouping.

2.2. Dependent and Conflicting Evidence

Dependent sources were studied well before modern RAG systems. Dong et al. modeled copying relationships among dynamic sources and showed that simple voting can over-count repeated information [8]. In current language-model systems, a similar problem appears when several dependent documents create the impression of independent support. Rahadi studies settings in which copies of one report artificially increase apparent answer support [9]. GraphEcho distinguishes the number of graph-retrieval paths from the number of independent evidence provenances and shows that these quantities can differ substantially [10].

In multi-agent systems, CAMA groups agent memories by latent sources, combines dependence estimation with provenance priors, and uses an effective number of independent evidence items for arbitration [11]. CF-RAG analyzes correlated contexts through counterfactual reasoning and isolates causally relevant differences under a correlated majority [12]. Together, these studies show that the number of retrieved passages is not a sufficient description of an evidence base and that source dependence should be modeled explicitly.

These studies show why dependent passages should be treated carefully. Most focus on source dependence or the effective amount of independent evidence. They give less attention to the specific combinations of passages that change support for one answer over another. We also need to know whether such interactions survive plausible changes in source grouping.

2.3. Attribution and Higher-Order Interactions

Attribution methods estimate the contribution of individual components to the output of a complex system and have recently been applied to RAG and multi-agent models. Nematov et al. used Shapley attribution for RAG documents and studied redundancy,

complementarity, and synergy among retrieved sources [3]. Semantic Cooperative Games represents agent interaction as a semantic hypergraph and uses the Semantic Shapley Value to attribute contributions in multi-agent processes [13]. Quotient Semivalues formalizes Shapley-, Banzhaf-, and Beta-like values over clusters derived from observed relations between elements, reducing the effect of duplicate and pseudo-identical representations on scalar attribution [14].

Other methods represent joint effects explicitly. Möbius decomposition, also known through Harsanyi dividends, assigns an irreducible contribution to the subset where it arises [15]. It has been used to study interactions in neural networks and language models [16,17]. Liao [18] examines how context subsets affect provenance sensitivity in agent actions. These methods help reveal effects that a single score per source can hide.

Shapley–Taylor assigns interaction attributions up to a specified order and allocates higher-order effects according to its interaction axioms [19]. Faith-Shap constructs faithful polynomial approximations to coalition games and defines interaction attributions from those approximations [20]. ContextCite estimates which parts of a supplied context influence a language model's generated statement [21]. These approaches address related attribution questions, whereas PQ-EID examines unredistributed Möbius dividends on canonical claim–provenance subsets and the structural comparability of these subsets across provenance partitions.

Previous research has made progress on duplication-resistant attribution, source dependence, and subset interactions. For web question answering, an unresolved issue is how to combine these ideas when evidence units depend on uncertain claims and provenance. A change in grouping can alter which passages belong to an interaction and even change its order.

Table 1 compares closely related methods by their evidence units, treatment of dependence, and outputs. In the studies reviewed, we found no single procedure that combines claim- and provenance-based evidence units, explicit interaction decomposition, and a check of structural stability across plausible source groupings.

Table 1. Comparison of the closest research directions and methods.

Main Output	Main Mechanism	Unit of Analysis	Method
Scalar document importance	Shapley and approximations	Document	Source Attribution in RAG [3]
Effective independent support	Grouping dependent memories and arbitration	Latent evidence factor	CAMA [11]
Contribution attribution in a multi-agent process	Semantic Shapley	Agent/semantic support	Semantic Cooperative Games [13]
Arbitration under correlated contexts	Counterfactual differentiation	Retrieved evidence	CF-RAG [12]
Duplication-resistant scalar attribution	Semivalue over a quotient set	Attribution cluster	Quotient Semivalues [14]
Localization of provenance sensitivity	Perturbation and subset interactions	Context factor	Provenance-Sensitivity Audit [18]

The review points to three practical needs: evidence units that preserve claim and source information, a representation of joint contributions, and a way to assess the effect of uncertain source grouping. The following sections address these needs.

3. Problem Formulation and the Scope of the Evidence Game

Let $R = \{r_1, \ldots, r_n\}$ be the retrieved raw passages associated with a question q and a pair of answer candidates h, g. Each passage has a normalized claim c_r and a provenance descriptor ρ_r consisting of its available source identity, primary-source and derivation links.

A candidate provenance partition P is a collection of disjoint, nonempty blocks whose union is R. Each block represents one canonical evidence unit; write $E_P = \{e_1, \ldots, e_{m_P}\}$.

Definition 1 (admissible canonicalization). *A pairwise rule $\kappa(r_i, r_j)$ may be used to propose merging passages only when their atomic claims are semantically equivalent and their provenance is compatible. Because pairwise compatibility need not be transitive, equivalence classes are not defined by naive connected components of uncertain similarity links. Instead, a valid partition P is constructed subject to explicit must-link and cannot-link constraints, with its blocks required to be semantically and provenance consistent. Where source identity cannot be resolved, alternative admissible partitions are retained rather than declared equivalent. The original strict equality rule is the special case in which the verified normalized provenance identifier and normalized claim are equal.*

Remark 1 (Remark on scope). *A normalized web domain is only an observable proxy for a provenance root. An original report reproduced on a second domain may belong to the same evidence unit if a primary-source or derivation relation is established; independent confirmations of the same fact remain separate. Neither same-domain membership alone nor semantic similarity alone proves copying or independence.*

For each P, the evaluator computes a candidate-specific evidence score $z_{a,P}(S, \pi; q)$ for $S \subseteq E_P$ and an ordering π of S. Write $\mathfrak{S}(S)$ for the permutations of S and define the order-symmetrized score

$$\bar{z}_{a,P}(S; q) = \frac{1}{|\mathfrak{S}(S)|} \sum_{\pi \in \mathfrak{S}(S)} z_{a,P}(S, \pi; q) \tag{1}$$

Equation (1) defines the ideal mean of candidate scores over evidence presentation orders. When exhaustive enumeration is infeasible, the mean is estimated from L sampled permutations, and L together with the resulting Monte Carlo error must be reported. In the real-LLM audit (Section 5.6), the coalition scores used one fixed evidence order, and the order-sensitivity experiment used $L = 3$ sampled non-original permutations of the full evidence set per model–question pair. For each candidate $a \in \{h, g\}$, we first average comparable scores as specified in Equation (1). For the score-first convention, the normalized two-candidate probability is $p_P(a \mid q, S) = \frac{\exp(\bar{z}_{a,P}(S;q))}{\exp(\bar{z}_{h,P}(S;q)) + \exp(\bar{z}_{g,P}(S;q))}$. Thus, the log ratio of these two probabilities equals the difference of the averaged candidate scores. Equation (2) subtracts the corresponding empty-coalition log ratio. Averaging probabilities before taking the logarithm generally gives a different evaluator. For two-way softmax without clipping, averaging the permutation-wise log ratios gives the same result as taking the difference between the two averaged scores.

$$v_P(S) = \log \frac{p_P(h \mid q, S)}{p_P(g \mid q, S)} - \log \frac{p_P(h \mid q, \varnothing)}{p_P(g \mid q, \varnothing)}, \qquad v_P(\varnothing) = 0 \tag{2}$$

In Equation (2), $p_P(h \mid q, S)$ and $p_P(g \mid q, S)$ are obtained from the averaged candidate scores under the score-first convention. With a fixed candidate pair, the contrast is equivalently the coalition score margin minus the empty-coalition score margin. The real-LLM coalition audit used a fixed evidence presentation order for each coalition; randomized orders were evaluated separately to measure sensitivity.

For the un-clipped two-candidate score-first protocol, the contrast can also be written as $v_P(S) = \left[\bar{z}_{h,P}(S; q) - \bar{z}_{g,P}(S; q)\right] - \left[\bar{z}_{h,P}(\varnothing; q) - \bar{z}_{g,P}(\varnothing; q)\right]$.

Operational Role of the Provenance Graph and PROV-O

The provenance graph in the conceptual formulation represents raw passages as information entities, normalized claims and answer candidates as semantic nodes, and available source identities as provenance nodes. PROV-O supplies an interoperable vocabulary for observed derivation (prov:wasDerivedFrom) and primary-source relationships (prov:hadPrimarySource). These edges constrain whether fragments may be merged into one canonical evidence unit. A verified derivation path can indicate dependence, whereas a shared website domain alone cannot establish either copying or independent origin. The graph therefore supports candidate canonicalizations and traceability; it does not assign a probability of factual correctness or independently certify the existence of missing provenance links.

In the QACC experiments, canonical evidence units were constructed using normalized source domains and claim metadata as provenance proxies. These operational groups support coalition and sensitivity analysis without assigning verified primary-source identities.

4. Provenance-Quotient Evidential Interaction Decomposition (PQ-EID)

4.1. Evidence Canonicalization

Canonicalization is implemented as a valid partition of raw passages under semantic-claim and provenance constraints. Pairwise compatibility is not presumed transitive; conflicting links are resolved before quotient units are formed. When evidence lineage is unknown, normalized domains serve only as proxies, and alternative partitions are retained for structural sensitivity analysis.

4.2. Evidential Interactions and Möbius Decomposition

For each nonempty $T \subseteq E_P$ define the interaction dividend

$$\Delta_P(T) = \sum_{S \subseteq T} (-1)^{|T|-|S|} v_P(S) \tag{3}$$

The inversion formula is

$$v_P(S) = \sum_{T \subseteq S} \Delta_P(T) \tag{4}$$

This is the standard Möbius/Harsanyi decomposition applied to the specified, provenance-dependent evidence game. The transform itself is standard. Its scientific use here lies in selecting the evidence units, evaluating the contrastive game, and assessing the stability of a particular interaction under alternative admissible canonicalizations. Positive and negative dividends indicate non-additivity relative to that particular game; their interpretation as synergy, substitution, or conflict requires examining the lower-order contributions and the question-specific scoring direction.

4.3. Overall Workflow

Figure 1 shows the workflow: retrieve evidence, form claim–provenance units, evaluate coalitions, decompose interactions, and assess signs over compatible provenance groupings.

Proposition 1 (replication invariance, conditional)**.** *Suppose the canonicalizer maps an added raw passage into an existing block without changing the canonical block representation, and the evaluator depends only on the resulting canonical set of blocks. Then every coalition score and dividend is unchanged.*

Proof. The canonical universe and every coalition input are identical before and after the replicated passage is added. The coalition function is therefore identical, and the Möbius transform of an identical function is identical. □

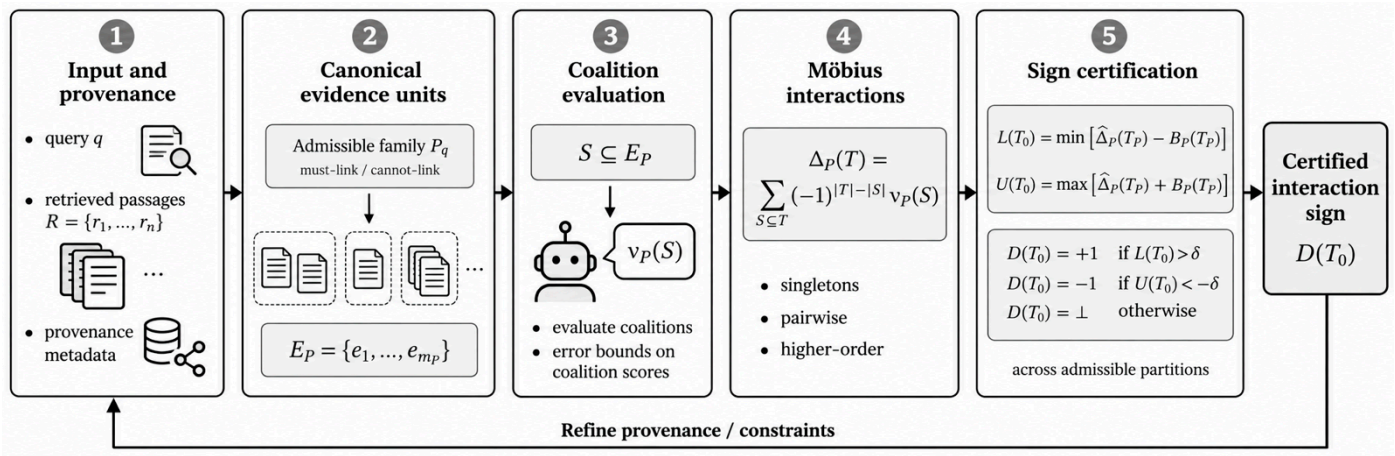


Figure 1. Overall workflow of the PQ-EID method.

4.4. Provenance Uncertainty and Comparable Interaction Targets

Let the admissible family P_q consist of complete partitions that satisfy known must-link and cannot-link constraints. One implementable candidate generator starts from the partition of individual passages, applies verified must-links, excludes every merge that violates a cannot-link, scores feasible block merges using a declared pairwise claim/provenance cost, and retains the lowest-cost B distinct partitions after each search level (beam search). It terminates when no feasible merge remains or when a stated search budget is reached. The scoring rule, B, similarity threshold, tie-breaking rule and any random seed must be fixed before evaluating interaction signs. Exhaustive enumeration scales with the Bell number, so a finite beam does not provide structural coverage by itself. For n raw passages, the search has at most $n - 1$ levels and examines at most $B \cdot \frac{n(n-1)}{2}$ candidate block merges per level, i.e., $O(B \cdot n^3)$ candidate-merge evaluations in the worst case, excluding the one-off $O(n^2)$ computation of the pairwise scores and the cost of evaluating the declared inconsistency function.

For an admissible partition P, let $C(P)$ be a stated inconsistency cost. A numerical weight can be assigned by

$$w_\tau(P) = \frac{\exp[-\tau C(P)]}{\sum_{P' \in \mathcal{P}_q} \exp[-\tau C(P')]}, \qquad \tau \geq 0 \tag{5}$$

The weight in Equation (5) is a normalized compatibility score within a candidate family, distinct from a calibrated partition probability. The Dirichlet experiment assigned weights from the specified uncertainty regimes. QACC evidence groups were defined using normalized domains and claim metadata.

To compare an interaction across partitions, fix a reference partition P_0 and target interaction T_0. An alignment map is defined only where every reference block in T_0 maps to one distinct canonical block in P under a declared raw-passage matching rule. The interaction order $|T_0|$ must be preserved. If two target blocks merge, one block splits, or the mapped coalition is ambiguous, the target is not comparable, and the procedure abstains. For any merged block B in P, its evaluator input contains the union of the raw-passage supports represented by B; the resulting game is evaluated anew on P. Its dividend must not be identified with the original dividend unless the alignment requirements hold.

$$I_{T_0}(P) = \mathbf{1}\{a_{T_0}(P) \text{ exists and preserves the target}\} \tag{6}$$

The descriptive structural-preservation mass is

$$\omega(T_0) = \sum_{P \in \mathcal{P}_q} w_\tau(P) I_{T_0}(P) \tag{7}$$

The value $\omega(T_0)$ measures weighted compatibility with preserving the target's structure within the generated family. This quantity does not provide calibrated coverage without additional assumptions. When $I_{T_0}(P) = 0$, the corresponding dividend cannot be treated as the same interaction across partitions; such cases require abstention or a separately defined mapping-aware target.

4.5. Simultaneous Evaluator Error Bounds

Fix a prespecified audit family of raw coalition scores, including each empty coalition. Let $\mu_P(S)$ be the expected raw score, $\hat{\mu}_P(S)$ its estimator, and $v_P(S) = \mu_P(S) - \mu_P(\varnothing)$ the baseline-subtracted coalition game. The estimated game is $\hat{v}_P(S) = \hat{\mu}_P(S) - \hat{\mu}_P(\varnothing)$. The simultaneous sampling event for raw scores is

$$\mathcal{E} = \bigcap_{P,S \in \mathcal{C}} \{|\hat{\mu}_P(S) - \mu_P(S)| \le e_{P,S}\} \tag{8}$$

For each raw coalition score, assume $n_{P,S}$ independent observations bounded in a known interval [*a*,*b*]. If M is the number of tested raw coalition scores, a two-sided Hoeffding bound with Bonferroni allocation gives the following radius at simultaneous error level α_E. Identically repeated deterministic model calls alone do not justify this sampling model.

$$e_{P,S} = (b-a)\sqrt{\frac{\log(2M/\alpha_E)}{2n_{P,S}}} \tag{9}$$

Because the same empty-coalition estimate enters every contrast, its uncertainty must be counted in every contrast radius. On event ε, $|\hat{v}_P(S) - v_P(S)| \le e_{P,S} + e_{P,\varnothing}$; for S = ∅ both the true and estimated contrasts equal zero exactly. Dependence between contrasts caused by their shared baseline is compatible with the union-bound construction. Without-replacement sampling or unbounded LLM logit scores requires a separately justified concentration bound.

For any interaction T on partition P, define its estimated Möbius dividend $\hat{\Delta}_P(T)$ from the estimated contrastive game and let the conservative interaction radius be

$$\hat{\Delta}_P(T) = \sum_{S \subseteq T} (-1)^{|T|-|S|} \hat{v}_P(S), \qquad B_P(T) = \sum_{\varnothing \neq S \subseteq T} (e_{P,S} + e_{P,\varnothing}) \tag{10}$$

Under the simultaneous raw-score event $\boldsymbol{\varepsilon}$, every prespecified interaction satisfies the following interval bound. The empty-subset term contributes no estimation error because the contrast at the empty coalition is identically zero.

$$\Delta_P(T) \in \left[\hat{\Delta}_P(T) - B_P(T), \hat{\Delta}_P(T) + B_P(T)\right] \tag{11}$$

If each nonempty raw-score estimate and the empty-score estimate has radius e, the bound is $B_P(T) = 2(2^{|T|} - 1)e$. This is deliberately conservative. A bound based on the combined coefficients of the shared baseline can be sharper, provided it is derived and applied consistently. Corollary 1 requires global simultaneous coverage, rather than separate unadjusted intervals for individual interactions.

4.6. Dual Sign Certification and Structural Stability

For each reference target T_0, prespecify a nonempty candidate family $\mathcal{A}(T_0)$ and the alignment rule before inspecting estimated signs. Every partition in $\mathcal{A}(T_0)$ must admit an unambiguous injective alignment of T_0 that preserves the interaction order. A candidate family containing a partition without such an alignment requires abstention or exclusion through a documented, prespecified rule.

$$L(T_0) = \min_{P \in \mathcal{A}(T_0)} \left[\hat{\Delta}_P(T_P) - B_P(T_P)\right], \qquad U(T_0) = \max_{P \in \mathcal{A}(T_0)} \left[\hat{\Delta}_P(T_P) + B_P(T_P)\right] \tag{12}$$

For a prespecified practical sign margin $\delta \geq 0$ define

$$D(T_0) = \begin{cases} +1, & L(T_0) > \delta, \\ -1, & U(T_0) < -\delta, \\ \perp, & \text{otherwise.} \end{cases} \tag{13}$$

The symbol $\perp$ denotes abstention; $\emptyset$ always denotes the empty coalition. A nonzero margin δ is an optional practical-sign threshold, not a statistical confidence level.

Theorem 1 (simultaneous deterministic sign preservation)**.** *Let $P^\star$ be an unknown true provenance partition for the question. Assume (i) $P^\star \in \mathcal{A}(T_0)$, (ii) the target has an unambiguous alignment $T_{P^\star} = a_{T_0}(P^\star)$, and (iii) event $\mathcal{E}$ holds. Then $D(T_0) = +1$ implies $\Delta_{P^\star}(T_{P^\star}) > \delta$; $D(T_0) = -1$ implies $\Delta_{P^\star}(T_{P^\star}) < -\delta$. Furthermore, when these conditions hold simultaneously for several targets, every issued sign is correct with respect to its target game.*

Proof. Under the alignment assumption, the interaction on the true partition has the same declared target interpretation. On the global sampling event ε, the triangle inequality for the Möbius transform, and the per-coalition contrast bounds imply $|\hat{\Delta}_P(T) - \Delta_P(T)| \leq B_P(T)$ for all audited pairs. Since the true partition belongs to $\mathcal{A}(T_0)$, its target dividend lies within the robust interval defined in Equation (12). A strictly positive lower bound above δ therefore implies a positive true interaction, and a strictly negative upper bound below $-\delta$ implies a negative true interaction. The same global event applies simultaneously to the full prespecified target family. □

Corollary 1 (unconditional probability of an erroneous issued sign)**.** *Suppose a separate structural-calibration procedure establishes a global event $\mathcal{A}$ under which each audited target's true partition lies in its candidate family and is aligned, with $Pr(\mathcal{A}) \geq 1 - \alpha_P$. If $Pr(\mathcal{E}) \geq 1 - \alpha_E$, then*

$$Pr\{\exists T_0 : D(T_0) \neq \perp \text{ and the issued sign is wrong}\} \leq \alpha_P + \alpha_E \tag{14}$$

Proof. Theorem 1 makes the event on the left impossible on $\mathcal{A} \cap \mathcal{E}$. Hence it is contained in $\mathcal{A}^c \cup \mathcal{E}^c$ and the union bound gives the stated inequality. □

The bound concerns the probability of an erroneously issued sign and does not automatically bound the conditional false-sign proportion among only certified cases. Because abstention is selective, the conditional false-sign rate can be larger than $\alpha_P + \alpha_E$ even when the unconditional event bound is valid.

Structural edit sensitivity. Define a structural edit neighborhood $\mathcal{N}_r(P_0)$ of partitions reachable by at most r admissible merge/split edits from a reference partition, with all alignment constraints fixed. The largest integer r for which the robust sign interval excludes zero over the entire neighborhood is the certified structural edit radius. This is a

deterministic sensitivity certificate, requiring no interpretation of w_τ as a probability. It does not establish that the true partition lies within radius r.

Distributional structural calibration can be formulated using question-level development and calibration samples with verified provenance labels. A prespecified nonconformity score and candidate-generation rule allow marginal coverage assessment under exchangeability; simultaneous coverage across targets requires an appropriate global construction or error allocation.

4.7. Computational Cost and Interaction Order

The number of distinct coalition values needed for an order-K truncated decomposition is

$$N_K(m) = \sum_{j=0}^{\min(K,m)} \binom{m}{j} \tag{15}$$

If all context permutations are enumerated, with R_j independent evaluator replications per permutation, the call count is

$$N_{\text{calls}} = \sum_{j=0}^{\min(K,m)} \binom{m}{j} j! R_j \tag{16}$$

For $m = 20$ and $K = 3$, there are 1351 distinct coalition values, including the empty coalition. With one evaluator replication per ordering, evaluating every permutation of all eligible subsets requires $1 + 20 + 380 + 6840 = 7241$ calls; exact unrestricted enumeration requires $2^{20} = 1{,}048{,}576$ distinct coalition values before permutation expansion. The count for a sampled-order implementation is the sum of the actual sampled orders across evaluated subsets, including any separate answer-position evaluations.

Order truncation returns exact lower-order Möbius dividends only if their required coalition values are accurately evaluated. It does not reconstruct an arbitrary full game from low-order terms unless higher-order dividends are zero (the K-additive assumption), and it does not certify that omitted higher-order effects are negligible.

Conditions for Sign Certification. Probabilistic sign certification is defined for prespecified provenance partition families and alignment rules, structurally calibrated coverage, simultaneous evaluator-error bounds that include the empty-coalition baseline, and abstention whenever target alignment or sign separation fails.

The controlled studies examine structural coverage and conditional sign guarantees, while the real-LLM analysis evaluates coalition interactions and evidence-order sensitivity.

4.8. Limitations of the Formal Guarantee

The formal guarantee is limited to the sign of an evaluator-defined interaction within a prespecified family of aligned provenance partitions. Its validity depends on structural coverage of the true provenance partition and simultaneous bounds on coalition-evaluation errors. These conditions must be independently justified in practical applications. The certificate does not establish the factual correctness of the evaluated evidence or account for systematic biases of the underlying language model.

5. Experimental Methodology

We tested PQ-EID in controlled settings and on web evidence, with each experiment addressing a different property. We first test interaction recovery when the evidence structure is known. We then study canonicalization errors and the behavior of dual certification under uncertain provenance. The method is next examined on conflicting QACC contexts, followed by a scalability analysis of exact and truncated decomposition.

5.1. Controlled Synthetic Scenarios

Controlled synthetic scenarios were used to evaluate how accurately PQ-EID recovers predefined individual and joint evidential effects under known ground truth. The experiments covered seven regimes: replication (REP), additive effects (ADD), substitution (SUB), second-order synergy (SYN2), third-order synergy (SYN3), conflicting effects (CON), and mixed interactions (MIX). Each regime was defined by a coalition game with prescribed interaction coefficients, from which the ground-truth coalition values were calculated before introducing evaluator noise. We conducted 2000 Monte Carlo repetitions per condition and assessed recovery by comparing the estimated Möbius dividends with their true values, focusing on estimation error and interaction-sign accuracy.

The synthetic games were defined by fixed nonzero Möbius coefficients, with all unspecified coefficients set to zero. REP used one evidence unit with coefficient 1.60. ADD used three units with singleton coefficients (0.80, 0.60, 0.70). SUB used two units with coefficients (1.00, 0.90) and a pairwise interaction of -0.80. SYN2 used singleton coefficients (0.05, 0.05) and a pairwise interaction of 1.50. SYN3 used three singleton coefficients of 0.02 and a third-order interaction of 2.00. CON used singleton coefficients (1.00, -0.90) and a pairwise interaction of -0.20. MIX used singleton coefficients (0.60, 0.40, -0.30), pairwise interactions of 0.70 and -0.40 for the first–second and first–third units, respectively, and a third-order interaction of 0.90. Ground-truth coalition values were obtained by summing the coefficients of all nonempty subsets of each coalition. These values were converted to probabilities using the logistic function with zero baseline log-odds. Independent uniform noise was then added to each coalition probability, with noise bounds of ± 0.0025, ± 0.005, and ± 0.010. Perturbed probabilities were clipped to the interval $[10^{-9}, 1 - 10^{-9}]$ before the log-odds transformation. The analytical sign-certification bound used a probability lower bound of 0.10. Each scenario and noise level was evaluated over 2000 Monte Carlo repetitions using NumPy's default random generator with seed 20260925. Table 2 reports the results for the noise bound of 0.005.

Table 2. Interaction recovery across controlled synthetic regimes.

False Zero Certification *	Certified-Sign Accuracy	Certification Rate	Sign Accuracy	MAE	Scenario
–	1.000	1.000	1.000	0.0194	REP
0.000	1.000	1.000	1.000	0.0215	ADD
–	1.000	1.000	1.000	0.0173	SUB
–	1.000	0.333	1.000	0.0166	SYN2
0.000	1.000	0.250	0.901	0.0184	SYN3
–	1.000	0.667	1.000	0.0171	CON
0.000	1.000	0.852	1.000	0.0202	MIX

* A value of 0.000 means that no truly zero interaction was incorrectly certified as positive or negative. The symbol "–" indicates that the metric was not applicable because the evaluated set contained no corresponding truly zero interactions.

Quotient Shapley and probabilistic Quotient Banzhaf were also computed on the same quotient set [14]. For these methods, we measured the deviation of scalar attribution from the true irreducible first-order contributions. This quantity captures the part of the joint value that is redistributed to individual units by a scalar attribution.

The comparison with quotient semivalues is therefore a comparison between two representations of evidential contribution. The reported deviation is not an error of the semivalue method; it measures how far a scalar attribution moves from the irreducible first-order effects when joint value is distributed over individual evidence units.

5.2. Provenance Canonicalization Errors

After testing interaction recovery under a correct evidence structure, we evaluated sensitivity to canonicalization quality. In real web data, provenance may be incomplete or ambiguous, and merge/split errors can change the space on which interactions are defined. This experiment measures these structural distortions and their effect on interaction identifiability.

The experiment used five canonical units, each represented by three raw replicas. Five canonicalization-error operators were considered: MERGE (merging distinct units), SPLIT (separating replicas), ROOT (incorrect provenance-root assignment), SEM (mixed semantic-equivalence errors), and ALL (combined errors).

The original configuration used a NumPy random generator seeded with 20261002 (20260925 + 77). The tested error rates were 0, 0.05, 0.10, 0.20, and 0.30, with 1000 repetitions for each error type and rate. The true game contained five atomic evidence units and 15 raw passages. Its nonzero Möbius coefficients were 0.8, 0.6, -0.4, 0.2, and 0.1 for singletons; 0.9, -0.5, and 0.4 for pairs (0,1), (1,2), and (3,4); and 1.2 and -0.7 for triples (0,1,3) and (1,2,4). MERGE combined distinct atomic units, SPLIT separated replicas, ROOT altered replica assignments, SEM applied splitting and merging at half the nominal error rate, and ALL applied root reassignment, splitting, and merging at one third of the nominal rate each. The auxiliary fixed-partition sign audit used a probability floor of 0.10 and a probability perturbation bound of 0.005.

Canonicalization quality was measured by F1 recovery of the true partition. We also measured relative distortion of interaction mass across orders, artificial overlap-interaction mass, and the fraction of runs in which the target third-order interaction remained structurally identifiable. These measures capture both partition recovery and the effect of structural errors on subsequent interaction decomposition.

5.3. Dual-Certification Test

The next experiment studied dual certification under increasing structural uncertainty. We focused on whether the certificate retains a correct interaction sign and abstains when admissible provenance structures do not support one consistent interpretation.

The experiment considered two third-order target interactions with prescribed ground-truth coefficients of +1.40 and -1.10. The candidate family comprised the true provenance partition, alternative configurations with incorrect merges, and a split configuration. Structural uncertainty was modeled by sampling partition weights from Dirichlet distributions under three regimes: low, medium, and high uncertainty. These regimes represented different degrees of concentration of the weights across candidate partitions. For each combination of uncertainty regime and target interaction, 5000 independent Monte Carlo runs were performed. The evaluation measured MAP-partition coverage and sign accuracy, dual-certificate coverage, certified-sign accuracy, false-sign frequency, and abstention rate.

The Dirichlet weights represent controlled variations in structural uncertainty rather than empirically calibrated probabilities of real-world provenance partitions.

5.4. QACC Pilot Study

We then examine the behavior of PQ-EID components on real conflicting web contexts. We used Question Answering with Conflicting Contexts (QACC), a human-annotated open-domain question answering dataset that includes cases in which different web sources support conflicting answers [1]. This dataset allows provenance-aware representation and interaction modeling to be examined under naturally heterogeneous evidence.

The exploratory QACC pilot included 10 conflicting questions, 24 candidate answers, and 28 directed correct-answer–alternative pairs. Source domains served as operational

provenance roots following manual inspection. The analysis addresses question-level interaction behavior within this pilot sample.

We first compared raw-context counting with counting unique provenance roots. To test replication sensitivity, an existing piece of evidence for an incorrect answer was duplicated without creating a new provenance root. We then used leave-one-question-out (LOQO) cross-validation with nested Platt calibration. Three nested models restricted the maximum interaction order. ADD used only the first-order support difference, PAIR added a second-order interaction, and TRIPLE also included two third-order terms. Differences in log loss were evaluated by question-level bootstrap with 10,000 resamples.

On the held-out TRIPLE games, Quotient Shapley and Quotient Banzhaf were computed over the same canonical units. Using the same quotient-set space and evaluator makes the difference in outputs attributable to the form of attribution.

5.5. Full-Corpus QACC Context from Published Results

To place the pilot in the broader QACC setting, we additionally used results published by the QACC authors for the full test set [1]. These results serve only as an external reference for how conflicting web contexts affect different language models beyond the small pilot sample.

The main reference is GPT-4o under the Context, Majority, Discern, and Explain settings. We also use published Context results for Claude-3 and two Phi-3 configurations. These values describe changes in model quality between conflict and non-conflict subsets and provide context for interpreting the pilot observations.

Published full-test-set QACC results [1] provide an external reference for contextualizing the pilot findings; PQ-EID coalition experiments are reported separately.

5.6. Real-LLM Coalition Audit and Evidence-Order Perturbation

We conducted an exploratory real-model audit on 60 QACC questions with two fixed-answer language models, Qwen2.5-0.5B-Instruct and SmolLM2-1.7B-Instruct, evaluated in CPU mode. The same questions were presented to both models. For each question, five evidence units were retained using deterministic domain-and-claim proxy grouping; the resulting coalition game contained 32 subsets (including the empty set). The evaluator compared two answer candidates using a fixed-answer, two-position scoring protocol, which reduces the impact of candidate-label position. The experiment recorded 70 forward evaluations per question and model, including additional stress-test evaluations, for 8400 evaluations overall. The coalition scores support exact Möbius decomposition and attribution comparisons. This is an audit of a fixed candidate pair, not generative exact-match accuracy.

The recorded model-forward execution time was 18,561.6 s (5.16 h) for Qwen2.5-0.5B-Instruct and 29,521.6 s (8.20 h) for SmolLM2-1.7B-Instruct. The corresponding input-token counts were 875,088 and 924,738, respectively. Across both models, the experiment comprised 8400 forward evaluations and 1,799,826 input tokens, with a combined model-forward execution time of approximately 13.36 h on the CPU-based system. These timings describe model evaluation and exclude model loading and dataset preparation.

The 60 questions were selected from QACC records containing two distinct annotated answer candidates and sufficient fragment metadata for five domain-and-claim evidence units. A seeded shuffle (seed 20261012) preceded deterministic candidate-pair and evidence-group selection. The experiment used a fixed-choice scoring protocol and documented question identifiers, evidence groups, and model configurations.

The identifiers, source-domain and claim-label proxies, grouped evidence fragments, prompt structure, and fixed-choice score extraction protocol are documented in Supplementary Materials S1–S3.

To examine sensitivity to the presentation order independently of changing evidence content, we evaluated three pseudo-random non-original permutations of the full five-unit evidence set for each model–question pair. The follow-up used the same scorer, fixed answer candidates, and deterministic permutation seeds. The design produced 120 model–question records and 720 additional forward evaluations (three permutations, two candidate positions). The fixed-order full-evidence score served as the reference. We report the mean over questions of the average absolute difference from this reference and the fraction of questions where at least one sampled permutation changed the sign of the two-candidate margin. Question-level percentile bootstrap resampling (10,000 draws) was performed separately by model; paired resampling was used to compare models on identical questions. These intervals describe sampling variability for this protocol and do not constitute finite-sample certificates for evidential-interaction signs.

Throughout this paper, interaction order means the number of canonical evidence units in a Möbius dividend. Evidence presentation order means the sequence in which passages appear in an LLM prompt. The former defines the algebraic interaction index; the latter may change the evaluator score and is studied independently through prompt permutations.

The real-data audit evaluates evidence interactions and grouping sensitivity using domain-and-claim provenance proxies derived from QACC metadata.

5.7. Extended Structural-Coverage Stress Test

We additionally tested sign certification in a fully controlled Bernoulli coalition-evaluation model. The stress-test grid combines five interaction regimes, four per-coalition sample sizes, and two structural-coverage settings, for 80,000 Monte Carlo trials in total. The true partition and interaction sign are known by construction. Coverage settings distinguish inclusion and deliberate exclusion of the true partition from the admissible family. The primary endpoints are frequency of issued signs, false-sign frequency, and abstention; they are reported separately.

5.8. Computational Complexity

The practical cost of interaction decomposition depends on both its analytical properties and the number of coalitions that must be evaluated. Because exhaustive subset enumeration grows exponentially, it is important to quantify how limiting the maximum interaction order changes the computational burden and the scalability of the method.

For m canonical evidence units, exact decomposition requires 2^m distinct coalition values, including the empty coalition. Restriction to maximum interaction order K requires $N_K(m)$ values as defined in Equation (15). These values give exact dividends through order K when the required scores are exact, but do not reconstruct arbitrary omitted higher-order effects.

The cost of complete presentation-order averaging is given by Equation (16).

We compared exact Möbius transformation with truncation at $K = 3$ for m $\in$ {10, 14, 16, 18, 20}. The reported runtime measures the algebraic transform after coalition scores are available. It excludes retrieval, canonicalization, prompt construction, LLM inference, and any repeated evaluations needed for presentation-order averaging.

In an LLM deployment, the dominant cost may be scoring each requested coalition, candidate position, and evidence presentation order. The implementation-dependent model-call count is therefore distinguished from transformation time.

6. Results

This section reports the experimental results using the methodology above. The analysis covers interaction recovery in controlled scenarios, sensitivity to canonicalization errors, dual certification under different levels of structural uncertainty, the QACC pilot, and computational scalability. Each experiment is discussed through measures that directly characterize the corresponding property of PQ-EID.

6.1. Interaction Recovery

We first evaluate how accurately the predefined evidence-interaction structure is recovered under controlled conditions. We focus on agreement with ground truth, sign correctness, and certification behavior for interactions of different orders, including truly zero effects.

Under the bounded-noise synthetic generator at the reported perturbation level, all issued nonzero interaction signs agreed with the specified ground-truth game (Table 2).

SYN3 illustrates the difference between explicit decomposition and scalar attribution. The sum of the three irreducible first-order contributions was 0.06, the third-order interaction was 2.00, and the full value was 2.06. Thus, 97.1% of the total evidential effect belonged to the joint interaction of the three units. Quotient Shapley, by design, redistributes this joint value to individual units and assigns about 0.6867 to each.

The comparison with quotient semivalues describes a difference between two forms of evidence representation rather than an error of semivalue attribution (Table 3). In the additive scenario, scalar attributions agree with irreducible first-order effects. When pure or mixed interactions are present, part of the joint value is distributed across individual scores. The reported MAE should therefore be interpreted as the deviation of scalar attribution from irreducible first-order effects caused by redistribution of joint interaction value.

Table 3. Comparison with quotient semivalues on noiseless synthetic games.

PQ-EID Deviation for First-Order Effects	Quotient Banzhaf Deviation from First-Order Effects	Quotient Shapley Deviation from First-Order Effects	Absolute Higher-Order Interaction Mass	Scenario
0.000	0.000	0.000	0.000	ADD
0.000	0.400	0.400	0.296	SUB
0.000	0.750	0.750	0.938	SYN2
0.000	0.500	0.667	0.971	SYN3
0.000	0.100	0.100	0.095	CON
0.000	0.325	0.400	0.606	MIX

6.2. Sensitivity to Provenance Errors

We next examine how errors in the provenance structure change the interaction space and how statistical stability under a fixed partition differs from structural correctness.

Incorrect merging and splitting affect decomposition in different ways. MERGE reduces the number of independent game dimensions: with a 20% error rate, the target third-order interaction remained identifiable in only 45.4% of runs (Table 4). SPLIT and ROOT preserved the target triple but created substantial artificial overlap-interaction mass between replicas.

Table 4. Effect of 20% canonicalization errors.

SYN3 Identifiability	Artificial Overlap Mass	Order-Mass Distortion	Partition F1	Error Type
0.454	0.000	0.179	0.843	MERGE
1.000	5.584	2.639	0.766	SPLIT
1.000	5.553	2.603	0.775	ROOT
0.690	2.165	0.986	0.816	SEM
0.814	3.077	1.408	0.810	ALL

The MERGE result should be interpreted as a structural identifiability limit. After distinct evidence units are merged incorrectly, information about the original units is lost before decomposition begins, so a later analysis cannot recover the original interaction order without additional provenance information. This supports conservative merging and the use of alternative admissible partitions in ambiguous cases.

Under the same conditions, a certificate that accounted only for evaluator error under a fixed factorization retained sign accuracy of 1.000 relative to the game defined on that partition. Structural accuracy for third-order interactions was 0.581 for MERGE, 0.762 for SPLIT, 0.759 for ROOT, 0.629 for SEM, and 0.647 for ALL. Evaluator-error control therefore does not cover canonicalization error.

6.3. Dual Certification

We then examine dual certification under different levels of structural uncertainty. We focus on coverage, certified-sign accuracy, and abstention when admissible partitions do not support one unambiguous interpretation.

As structural uncertainty increases, dual certification reduces coverage and abstains more often when admissible partitions change the order or direction of the interaction (Table 5).

Table 5. Coverage, accuracy, and abstention of dual certification under simulated structural uncertainty.

Abstention Rate	False-Sign Rate Among Certified	Certified-Sign Accuracy	Dual-Certificate Coverage	MAP Accuracy Among DECISIONS	MAP Coverage	Regime
0.026	0.033	0.967	0.974	0.967	1.000	Low/POS
0.505	0.096	0.904	0.495	0.882	1.000	Low/NEG
0.072	0.053	0.947	0.928	0.945	1.000	Medium/POS
1.000	–	abstain	0.000	0.695	1.000	Medium/NEG
0.134	0.068	0.932	0.866	0.920	0.953	High/POS
1.000	–	abstain	0.000	0.443	0.947	High/NEG

The effect is clearest in the High/NEG regime. A certificate based only on the MAP partition returned a sign in 94.7% of runs, but its accuracy among those decisions was 0.443. Under the same conditions, the dual certificate abstained in all runs, reflecting sensitivity to structural ambiguity.

In the positive regimes, where structural uncertainty less often changed the interaction sign, dual-certificate coverage ranged from 0.866 to 0.974 and certified-sign accuracy from 0.932 to 0.967. The observed coverage–abstention trade-off follows the logic of the certification rule: a sign is returned only when the statistical estimate and the admissible provenance structures remain consistent.

6.4. Results of the QACC Pilot Study

After controlled evaluation of individual properties, we examined how they appear in real conflicting web contexts. The QACC pilot focused on two properties directly related to the purpose of PQ-EID: robustness of the provenance-aware representation to repeated evidence and the presence of non-additive signal between canonical evidence units.

Counting unique provenance roots did not improve answer-selection accuracy in the small pilot. Pairwise accuracy was 0.786 for raw-context counting and 0.750 for unique-root counting. Raw-context counting was sensitive to replication: strict question-level accuracy fell from 0.70 to 0.30 when evidence for an incorrect answer was duplicated twice, to 0.20 at three copies, and to 0 at five copies. The unique-root representation remained unchanged because repeating an existing item did not create a new provenance root.

For the remaining analysis, we compared two baseline evidence representations and three nested interaction-model configurations. Raw-context counting uses all retrieved passages without provenance-aware aggregation, while unique-root counting keeps only distinct normalized sources. ADD contains first-order effects, PAIR adds second-order interactions, and TRIPLE extends the model with third-order interactions.

The regularized leave-one-question-out (LOQO) comparison yielded log losses of 0.678 (ADD), 0.625 (PAIR), and 0.581 (TRIPLE), as reported in Table 6. The comparison is descriptive at the question-sample scale.

Table 6. Pairwise predictive quality in the QACC pilot.

Brier Score	Log Loss	Pairwise Accuracy	Configuration/Baseline
0.172	0.529	0.786	Raw-context count
0.183	0.551	0.750	Unique provenance-root count
0.243	0.678	0.750	ADD
0.216	0.625	0.750	PAIR
0.196	0.581	0.750	TRIPLE

As shown in Table 6, raw-context counting achieved the lowest log loss and Brier score among the evaluated configurations. Within the ADD–PAIR–TRIPLE family, incorporating second- and third-order interactions progressively reduced both measures in the 10-question pilot, although raw-context counting retained the best predictive performance. The provenance-aware representation also remained unchanged when existing evidence was replicated, consistent with the replication-invariance property established in Proposition 1.

In the deliberately degree-three polynomial-logit evaluator, higher-order structure is restricted by construction. Therefore, the small observed contributions above order three are a specification check, not evidence that actual LLM interactions are generally third-order or lower.

For the held-out TRIPLE games, interactions of order two or higher accounted for a mean of 0.381 of the absolute interaction mass. The mean deviations from first-order contributions were 0.0724 for Quotient Shapley and 0.0632 for Quotient Banzhaf. Quotient Shapley changed sign relative to the first-order contribution in 2 of 51 cases. These differences reflect what each measure represents: scalar attribution distributes joint value to individual units, while Möbius decomposition keeps joint contributions on their subsets.

6.5. Full-Corpus QACC Results as an External Reference

As described in Section 5.5, the results published by the QACC authors for the complete test set [1] serve only as an external reference.

The published data show clear performance drops on the conflict subset. For GPT-4o in the Context setting, Exact Match decreases from 54.13 to 42.51, a drop of 11.62 percentage points, while F1 falls from 70.06 to 59.29. For the same model, conflict EM in the Majority, Discern, and Explain settings is 40.58, 39.13, and 38.65, respectively [1]. Table 7 reports Context results for GPT-4o, Claude-3, and two Phi-3 configurations.

Table 7. Published context-setting results on the full QACC test set [1].

F1 Non-Conflict	F1 Conflict	EM Difference	EM Non-Conflict	EM Conflict	Model
70.06	59.29	−11.62	54.13	42.51	GPT-4o
67.90	53.76	−17.39	53.14	35.75	Claude-3
67.30	56.65	−9.47	51.98	42.51	Phi-3 few-shot
64.60	56.24	−6.88	51.32	44.44	Phi-3 fine-tuned

All four model configurations in Table 7 exhibit lower performance on conflicting contexts. The mean decrease is 11.34 EM points and 10.98 F1 points. These published QACC test-set results provide an external reference for the real-context experiments.

6.6. Real-LLM Interactions and Order Robustness

Both models completed all 60 questions with five evidence units per question and all 32 coalition scores per question. Algebraic consistency checks of the Möbius expansion and additive Shapley accounting were satisfied in the independent analysis. The fraction of absolute Möbius mass on interaction orders two and above averaged 0.763 for Qwen2.5-0.5B and 0.773 for SmolLM2-1.7B under the original proxy grouping. These descriptive fractions indicate non-additivity in the specified evaluator-defined games; they are not measurements of true source-dependence recovery. The correct answer candidate received the higher fixed-choice margin on 43/60 questions for Qwen (71.7%) and 44/60 for SmolLM2 (73.3%); this is pairwise candidate preference, not open-generation Exact Match.

In the randomized-order follow-up, the mean absolute full-evidence margin change was 0.2632 for Qwen (95% question-bootstrap interval [0.2190, 0.3098]) and 0.1126 for SmolLM2 ([0.0949, 0.1319]) (Table 8). At least one of the three sampled permutations changed the decision-margin sign on 11/60 questions (18.3%) for Qwen and 13/60 (21.7%) for SmolLM2. The paired mean difference in absolute change (Qwen minus SmolLM2) was 0.1507 (paired 95% bootstrap interval [0.1087, 0.1964]); a paired Monte Carlo permutation test reported p approximately 0.00005. These results support greater average magnitude sensitivity for Qwen under this protocol, while the observed frequency of sign flips was slightly greater for SmolLM2. The two measures capture different aspects of robustness.

Table 8. Real-LLM coalition audit and evidence-order robustness.

SmolLM2-1.7B	Qwen2.5-0.5B	Real-LLM Audit Metric
0.1126	0.2632	Mean absolute order delta
[0.0949, 0.1319]	[0.2190, 0.3098]	95% bootstrap interval
13/60	11/60	Any sign change (3 permutations)
0.773	0.763	Higher-order absolute mass fraction
44/60	43/60	Fixed-choice correct preference

When we grouped evidence by domain alone rather than by domain and claim, the mean higher-order mass fraction fell from 0.7628 to 0.7236 for Qwen and from 0.7733 to 0.7290 for SmolLM2. Some evidence units merged on 18 of the 60 questions. The three sampled permutations also cover only a small part of the 120 possible orders of five evidence units.

6.7. Extended Validation of the Dual Sign Certificate Under Structural Uncertainty

We tested the sign certificate in a controlled Bernoulli game with three atomic fragments. For each subset S of {0, 1, 2}, observations had mean $q(S) = 0.5 + \theta\,\mathbf{1}[\{0, 2\} \subseteq S] + \eta\,\mathbf{1}[\{1, 2\} \subseteq S]$, with all selected parameters constrained to keep $q(S)$ within [0, 1]. Under the true three-block partition, the target interaction of blocks 0 and 2 has value θ. Under the alternative partition merging fragments 0 and 1 into a block B, the interaction between B and block 2 has value $\theta + \eta$. These are separately defined games obtained by evaluating the corresponding unions of atomic fragments. Their signs may be compared as a stress test, but the two dividends are not aligned instances of one invariant interaction when the target units have changed.

We evaluated five scenarios (strong positive, weak positive, conflicting structural signs, null true interaction, and strong negative), four sample sizes (n = 100, 500, 2000, 6000 observations per atomic coalition), and two structural-coverage regimes. Each of the 40 combinations was independently repeated 2000 times, yielding 80,000 trials. The experiment deliberately includes a misspecified regime in which the true provenance partition is omitted from the admissible family; this regime tests the necessity of the structural coverage condition rather than the proposed guarantee itself.

For each of eight atomic coalitions, independent Bernoulli observations were bounded with two-sided Hoeffding intervals and Bonferroni control at $\alpha = 0.05$. With n observations per coalition and eight coalitions, the raw-mean radius is $e = \sqrt{\frac{\log(16/\alpha)}{2n}}$. The second-order interaction is expressed as a four-term contrast of the raw coalition means, including the empty-coalition mean. If all four estimates have error radius e, the interaction error is bounded by $4e$. This controlled result is separate from the general conservative contrast bound in Equation (10), which includes a shared estimated empty baseline unless algebraic cancelation is used (Table 9).

Table 9. Controlled dual-certificate results at n = 6000 observations per atomic coalition (2000 independent trials per row). "Incorrect" means a certified sign inconsistent with the true-partition target; zero certifications imply that the conditional error rate among certificates is undefined.

Abstained/2000	Incorrect/2000	Issued/2000	Scenario/True Structure
0	0	2000	aligned_strong/included
2000	0	0	aligned_weak/included
2000	0	0	discordant/included
92	1908	1908	discordant/excluded
1456	544	544	null/excluded
0	0	2000	negative/included

The experiment produced no false sign certificates in the reported correctly covered scenarios. For conflicting signs across partitions, the method abstained in all 2000 trials at n = 6000. When the true partition was excluded, 1908/2000 decisions in the conflicting scenario and 544/2000 in the null scenario were incorrectly certified relative to the target defined by the true partition. These cases expose a concrete failure mode if provenance uncertainty sets omit the true structure; they do not contradict the conditional guarantee.

The structural-coverage simulations use externally specified candidate partitions and bounded Bernoulli observations. Accordingly, their finite-sample guarantee is conditional on the specified partition family and sampling model; application to LLM logit margins requires a separately justified evaluator-error model. Verification of real-world source origins constitutes a distinct provenance assessment.

The attribution methods summarize the same coalition game in different ways. Exact Shapley values assign total contributions to individual units; Shapley–Taylor and pairwise

interaction indices summarize selected joint effects; Möbius dividends retain effects on the subsets where they occur. Agreement between these calculations checks the implementation. Determining which method best recovers real source relationships would require independently verified provenance labels.

6.8. Scalability

Finally, we evaluate the computational scalability of PQ-EID. We examine how the number of required coalitions grows under exact decomposition and how much it is reduced when the maximum interaction order is limited.

The cost of exact decomposition grows exponentially with the number of canonical evidence units because all coalitions must be evaluated. Truncating the analysis at a fixed maximum order greatly reduces this number. Table 10 compares exact decomposition with truncation at third order.

Table 10. Scaling of exact decomposition and truncation at third-order interactions.

K = 3 Time, s	Exact Time, s	Coalition Reduction	Coalitions, K = 3	Coalitions, Exact	Evidence Units, m
0.0007	0.0015	5.8×	176	1024	10
0.0021	0.0365	34.9×	470	16,384	14
0.0033	0.1720	94.0×	697	65,536	16
0.0047	0.7559	265.3×	988	262,144	18
0.0066	3.3251	776.1×	1351	1,048,576	20

As shown in Table 10, the benefit of truncation increases quickly with the number of evidence units. At $m = 20$, the number of coalition evaluations is reduced by a factor of 776.1, while the algebraic-transform time is reduced by about 504 times (computed from the rounded timings in Table 10). In systems that use an LLM evaluator, the main practical benefit is the reduction in model calls, which is usually the dominant computational cost.

Truncation also defines the resolution of the analysis. With $K = 3K = 3$, individual effects and interactions up to third order are evaluated, while higher-order effects are outside the analysis. The maximum interaction order therefore controls the trade-off between computational cost and structural detail.

The real-model experiment found non-additive effects in fixed-choice coalition scores. Those scores also changed when we reordered the evidence or altered the proxy grouping. A certificate built for one fixed evaluator and one stated family of provenance partitions applies to that setting. Certifying stability across orders would require a defined set or distribution of orders and uncertainty bounds that cover it. The present order experiment was designed to measure sensitivity, not to establish that additional guarantee.

7. Discussion

The empirical evaluation comprises fixed-choice LLM coalition and evidence-order analysis, domain-and-claim grouping sensitivity, and controlled Bernoulli experiments of conditional sign certification. Each setting evaluates a distinct component of PQ-EID.

The results distinguish statistical stability of an estimated interaction under a fixed partition and structural correctness of that partition. Provenance determines which raw passages form one evidence unit, the contrastive game measures changes in relative candidate support, and Möbius decomposition localizes irreducible joint effects on evidence subsets [15–17]. A small coalition-estimation error therefore does not by itself guarantee a structurally correct interpretation when provenance canonicalization remains uncertain.

In the controlled synthetic scenarios, PQ-EID recovered predefined interactions with small error under known factorization, and certified non-zero signs were correct in all

cases where a certificate was issued (Table 2; bounded-noise generator). In the simulation with uncalibrated Dirichlet partition weights, however, 3.3–9.6% of the issued signs were wrong (Table 5). Weaker effects were more often left uncertified, which reflects the selective nature of the procedure. The comparison with Quotient Shapley and Quotient Banzhaf illustrates a difference between scalar attribution and explicit interaction representation [14]. A semivalue distributes coalition value across individual units according to its attribution rule, while Möbius decomposition keeps an irreducible joint effect on the subset where it appears. For structural auditing, this makes the composition and order of the interaction directly visible.

The canonicalization-error experiment reveals a different source of error. Incorrect merging can change the composition and interaction order (number of evidence units) of an interaction and make the target effect unidentifiable, while incorrect splitting creates extra game elements and artificial overlap interactions. This result is consistent with the broader problem of dependent and repeated sources, for which earlier work has shown limitations of simple counting or voting [8,11,12]. PQ-EID treats this issue at the level of interaction structure. The structural-preservation mass in Equation (7) quantifies the weighted consistency of an interaction across candidate partitions, while the robust bounds in Equation (12) assess the stability of its sign over the admissible partition family.

The QACC pilot complements the controlled scenarios with observations on conflicting web contexts [1]. Unique-root counting was invariant to replication of an existing evidence item, while raw-context counting achieved lower predictive loss. Within the ADD–PAIR–TRIPLE comparison, higher interaction orders reduced the reported log loss. These findings distinguish structural auditing from answer-selection performance.

The pilot results characterize prediction and evidential auditing as distinct objectives under the selected domain-level provenance proxy.

This positioning distinguishes PQ-EID from methods centered mainly on scalar attribution or the effective number of independent sources [11,14]. Provenance-sensitivity analysis [18] is close in motivation, but PQ-EID additionally examines whether a specific interaction is preserved across alternative admissible canonicalizations and certifies its sign under combined statistical and structural uncertainty.

The comparison with Shapley–Taylor, Faith-Shap and ContextCite concerns the quantity each method explains. Their interaction or context-attribution outputs can be useful in their own right, whereas PQ-EID focuses on exact coalition dividends after provenance-aware canonicalization and tests the stability of those targets across compatible partitions.

The dual-certification results further show that increasing structural uncertainty can move the procedure from issuing a sign to abstaining. This is clearest in the High/NEG regime, where the MAP-partition certificate returned a sign in 94.7% of runs with accuracy 0.443, while the dual certificate abstained in all runs. The corresponding coverage and error rates are given in Table 5 and the accompanying text in Section 6.3.

Published full-corpus QACC results provide an external reference for the effect of conflicting contexts on answer quality [1]. The present study additionally includes a separate fixed-choice coalition audit of 60 QACC questions with Qwen2.5-0.5B-Instruct and SmolLM2-1.7B-Instruct (Section 6.6). That audit examines interaction structure and evidence-order sensitivity rather than open-ended answer generation or the full QACC test set. Extending the analysis to a complete conflicting-question corpus, additional retrieval pipelines, and larger generative models remains an important direction for future work.

The scalability analysis quantifies the cost of explicit interaction decomposition. Exact analysis requires an exponential number of coalitions, while limiting the maximum interaction order greatly reduces the number of evaluations. At 20 evidence units, truncation at

third order reduced the coalition count by a factor of 776.1. The chosen maximum order therefore controls a practical trade-off between structural detail and computational cost.

PQ-EID could be used after evidence retrieval and before an answer explanation is produced. The provenance layer records source relationships, grouping reduces repeated evidence, Möbius decomposition identifies joint effects, and the certificate assesses signs when its assumptions are met. The audit can explain the structure of the supporting evidence while leaving answer selection to the original system.

Table 11 summarizes the functional role of the main PQ-EID components and the corresponding experimental observations.

Table 11. Functional interpretation of PQ-EID components based on the experimental results.

Interpretation	Experimental Observation	Problem Controlled	Analysis Level
The number of copies from the same provenance can affect a raw-context counting decision	Strict accuracy fell from 0.70 to 0.00 as incorrect-evidence replication increased from 1× to 5×	No control of repeated-evidence multiplicity	Raw contexts
Under the adopted canonicalization rule, replication inside an existing provenance class does not create a new evidence unit	The unique-root representation did not change when an existing item was duplicated	Repeated representation of evidence from an existing provenance root	Provenance-aware canonicalization
Second- and third-order interactions add information about joint effects within the studied pilot model	LOQO log loss decreased from 0.678 (ADD) to 0.625 (PAIR) and 0.581 (TRIPLE)	Non-additive joint contribution of evidence units	Interaction decomposition
Scalar attribution and explicit decomposition describe different properties of the same evidence game	For TRIPLE, mean Quotient Shapley deviation was 0.0724 and mean Quotient Banzhaf deviation was 0.0632; Quotient Shapley changed sign in 2 of 51 cases	Redistribution of joint value to individual units	Comparison with scalar attribution
In this regime, a structurally unstable sign was not issued as certified	In High/NEG, MAP coverage was 0.947 with accuracy 0.443, while the dual certificate abstained in all runs	Combined statistical and structural uncertainty	Dual certification
The maximum interaction order sets the trade-off between analytical detail and computational cost	For 20 evidence units, truncation at third order reduced the coalition count by 776.1×	Cost of exhaustive coalition evaluation	Interaction-order truncation

Further work should test PQ-EID on larger datasets and broader evidence collections. The main priorities are better provenance information, evaluation with more retrieval systems and language models, and lower computational cost for large evidence sets.

One direction is to extend the experiments to full conflicting-question corpora, different retrieval mechanisms, and several answer-generation model families. Such studies can examine whether the observed interaction structures remain stable when the retrieved evidence, ranking procedure, and evaluator architecture change. Of particular interest is the behavior of interactions in knowledge-enhanced web question answering systems where provenance is naturally incomplete and heterogeneous.

A complementary direction is a systematic comparison of PQ-EID with complementary attribution approaches, particularly Faith-Shap and ContextCite. Evaluating these methods under a unified experimental protocol would provide deeper insights into their capabilities for identifying individual evidential contributions, higher-order interactions, and sensitivity to provenance uncertainty.

A further direction is empirical estimation of structural uncertainty. The controlled experiments in this paper provide a basis for studying admissible provenance partitions and interaction stability under alternative canonicalizations. Future work can estimate partition weights from semantic, temporal, and provenance features of web sources. Combining these signals with small manually verified subsets could connect formal structural coverage more closely to observed canonicalization quality.

Scalability also warrants scalable decomposition for larger evidence sets. The current results show that restricting the maximum interaction order substantially reduces the number of coalitions. Adaptive order selection, sparse search for informative interactions, and approximate Möbius transforms may further improve scalability while retaining interpretable joint effects and explicit sign control.

Finally, PQ-EID can be integrated into an end-to-end web question answering pipeline that covers retrieval, semantic normalization, provenance-graph construction, coalition evaluation, and explanation generation. In this setting, provenance-aware auditing can operate as a separate analytical layer that accompanies an answer with information about evidence independence, interaction structure, and the stability of conclusions under provenance uncertainty.

8. Conclusions

We developed PQ-EID to examine dependent evidence and interactions in web question answering. It uses claim and provenance information to form evidence units, evaluates how coalitions affect support for competing answers, and applies Möbius decomposition to identify individual and joint contributions. With a consistent canonical representation, adding another copy of an existing passage leaves the coalition game unchanged.

When source information is uncertain, PQ-EID considers multiple compatible groupings. Structural preservation measures whether an interaction keeps its members and order across those groupings. It is a descriptive measure and is not part of the certification rule. A robust interval then tests whether its sign remains consistent. The sign guarantee requires both valid evaluation-error bounds and inclusion of the true provenance structure in the candidate family.

In controlled experiments, PQ-EID recovered predefined interactions and showed how incorrect grouping can change the structure of an interaction. The semivalue comparisons clarified the difference between assigning scores to individual sources and retaining joint effects. In the small QACC pilot, the grouped representation was unaffected by added copies of existing evidence. Adding second- and third-order terms reduced log loss within the tested interaction models, although raw-context counting performed better as an answer predictor. In the 60-question real-LLM experiment, both models showed higher-order interactions and sensitivity to evidence order. The controlled certificate experiment showed why structural coverage matters: excluding the true grouping can lead to incorrect signs.

Overall, PQ-EID provides a way to audit how web evidence works together when sources may be repeated or uncertain.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/fi18110582/s1.

Author Contributions: Conceptualization, K.Z. and I.D.; methodology, K.Z.; software, K.Z.; validation, K.Z., I.D. and O.M.; formal analysis, K.Z. and I.D.; investigation, K.Z. and O.M.; resources, K.Z. and O.M.; data curation, K.Z. and O.M.; writing—original draft preparation, K.Z.; writing—review and editing, K.Z., I.D. and O.M.; visualization, K.Z.; supervision, I.D.; project administration, K.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Question Answering with Conflicting Contexts (QACC) dataset used in this study is publicly available through the Amazon Science repository: https://www.amazon.science/code-and-datasets/open-domain-question-answering-with-conflicting-contexts (accessed on 22 October 2026). The dataset description and accompanying publication materials are available through ACL Anthology: https://aclanthology.org/2025.findings-naacl.99/ (accessed on 22 October 2026); DOI: https://doi.org/10.18653/v1/2025.findings-naacl.99.

Acknowledgments: During the preparation and revision of this manuscript, the authors used ChatGPT (OpenAI) to assist with language editing, translation, terminology consistency, manuscript formatting, development and debugging of experimental code, and revision of the manuscript in response to reviewers' comments. The authors reviewed and edited the AI-assisted material, including experimental scripts and the revised text and take full responsibility for the accuracy and integrity of the published work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, S.; Ning, Q.; Halder, K.; Qi, Z.; Xiao, W.; Htut, P.M.; Zhang, Y.; John, N.A.; Min, B.; Benajiba, Y.; et al. Open Domain Question Answering with Conflicting Contexts. In *Findings of the Association for Computational Linguistics: NAACL 2025*; Association for Computational Linguistics: Albuquerque, NM, USA, 2025; pp. 1838–1854. [CrossRef]
2. Lebo, T.; Sahoo, S.; McGuinness, D. (Eds.) PROV-O: The PROV Ontology. *W3C Recommendation*, 30 April 2013. Available online: https://www.w3.org/TR/prov-o/ (accessed on 26 September 2026).
3. Nematov, I.; Kalai, T.; Kuzmenko, E.; Fugagnoli, G.; Sacharidis, D.; Hose, K.; Sagi, T. Source Attribution in Retrieval-Augmented Generation. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: ECML PKDD 2025 Workshops, Part II*; Springer: Berlin/Heidelberg, Germany, 2026; Volume 2840, pp. 317–332. [CrossRef]
4. Shapley, L.S. A Value for n-Person Games. In *Contributions to the Theory of Games II*; Kuhn, H.W., Tucker, A.W., Eds.; Princeton University Press: Princeton, NJ, USA, 1953; pp. 307–317.
5. Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; de Melo, G.; Gutierrez, C.; Kirrane, S.; Labra Gayo, J.E.; Navigli, R.; Neumaier, S.; et al. Knowledge Graphs. *ACM Comput. Surv.* **2021**, *54*, 71. [CrossRef]
6. Diefenbach, D.; Lopez, V.; Singh, K.D.; Maret, P. Core Techniques of Question Answering Systems over Knowledge Bases: A Survey. *Knowl. Inf. Syst.* **2018**, *55*, 529–569. [CrossRef]
7. Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3580–3599. [CrossRef]
8. Dong, X.L.; Berti-Equille, L.; Srivastava, D. Truth Discovery and Copying Detection in a Dynamic World. *Proc. VLDB Endow.* **2009**, *2*, 562–573. [CrossRef]
9. Rahadi, I. Counting Copies as Evidence: Confidence Inflation from Dependent Evidence in Retrieval-Augmented Generation (RAG). *Preprint* **2026**. [CrossRef]
10. Wang, S.; Zhou, Y.; Fan, L.; Zhang, F. GraphEcho: Structural Redundancy and Evidence Provenance in LLM Graph Agents. *arXiv* **2026**, arXiv:2609.17695. [CrossRef]
11. Lin, C.; Yuan, W.; Wang, X.; Ngai, E.C.H. Beyond Memory Majority: Latent-Source Reasoning for Multi-Agent Memory Arbitration. *arXiv* **2026**, arXiv:2608.19701. [CrossRef]
12. Qin, H.; Wei, C.; Chen, Y.; Wang, Y. Counterfactual Reasoning for Retrieval-Augmented Generation. In Proceedings of the International Conference on Learning Representations (ICLR 2026). 2026. Available online: https://proceedings.iclr.cc/paper_files/paper/2026/hash/1c078897dc08d46091d0d361d9955c6b-Abstract-Conference.html (accessed on 22 October 2026).
13. Jiang, P.; Zhu, X.; Zhu, Q. Semantic Cooperative Games for Contribution Attribution in LLM-Based Multi-Agent Systems. *arXiv* **2026**, arXiv:2607.18255. [CrossRef]

14. Burnat, F.A.D.; Davidson, B.I. Quotient Semivalues for False-Name-Resistant Data Attribution. *arXiv* **2026**, arXiv:2605.07663. [CrossRef]
15. Harsanyi, J.C. A Simplified Bargaining Model for the n-Person Cooperative Game. *Int. Econ. Rev.* **1963**, *4*, 194–220. [CrossRef]
16. Qu, X.; Yu, Z.; Liu, D.; Wei, W.; Liu, D.; Dong, J.; Cheng, Y. Cooperative or Competitive? Understanding the Interaction between Attention Heads From A Game Theory Perspective. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics ((Volume 1: Long Papers))*; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 14079–14099. [CrossRef]
17. Kang, J.S.; Butler, L.; Agarwal, A.; Erginbas, Y.E.; Pedarsani, R.; Yu, B.; Ramchandran, K. SPEX: Scaling Feature Interaction Explanations for LLMs. In Proceedings of the 42nd International Conference on Machine Learning, Vancouver, BC, Canada, 13–19 July 2024. Available online: https://proceedings.mlr.press/v267/kang25a.html (accessed on 22 October 2026).
18. Liao, J. Auditing Provenance Sensitivity in LLM Agent Action Selection. *arXiv* **2026**, arXiv:2607.20827. [CrossRef]
19. Sundararajan, M.; Dhamdhere, K.; Agarwal, A. The Shapley Taylor Interaction Index. In *Proceedings of the 37th International Conference on Machine Learning*; PMLR: Birmingham, UK, 2020; Volume 119, pp. 9259–9268. Available online: https://proceedings.mlr.press/v119/sundararajan20a.html (accessed on 22 October 2026).
20. Tsai, C.-P.; Yeh, C.-K.; Ravikumar, P. Faith-Shap: The Faithful Shapley Interaction Index. *J. Mach. Learn. Res.* **2023**, *24*, 1–42. Available online: https://jmlr.org/papers/v24/22-0202.html (accessed on 22 October 2026).
21. Cohen-Wang, B.; Shah, H.; Georgiev, K.; Madry, A. ContextCite: Attributing Model Generation to Context. In Proceedings of the Advances in Neural Information Processing Systems 37, Vancouver, BC, Canada, 10–15 December 2024. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.