CROSS-MODAL EVIDENTIAL RUPTURE DETECTION AND LOCALIZATION IN SCIENTIFIC RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the growing success of AI systems in scientific research, hallucinations and unsupported outputs remain a critical concern, particularly when plausible errors propagate across stages of scientific analysis and lead to consequential conclusions. In particular, AI systems increasingly generate heterogeneous scientific entities, including text, mathematical formulations, code, and experimental results. While individual entities may appear plausible, the evidential relations connecting them can lose support, making such cross-entity inconsistencies difficult to detect. We propose a novel notion named *cross-modal evidential rupture* for characterizing the loss of support in an expected evidential relation between scientific entities. Starting from this formulation, we derive a topology-dependent characterization of the structural information to identify such ruptures. Our analysis establishes a boundary between local tracking, identity-preserving tracking, and cross-anchor correspondence. In particular, overlapping structures require explicit cross-anchor correspondence because local information is insufficient to preserve relation identity. Based on this characterization, we introduce RUPTURETRACE, a topology-aware framework for cross-modal evidential rupture detection and localization. RUPTURETRACE combines relation-level directional support measurement with topology-aware tracking and uses zigzag persistence when overlapping structures require cross-anchor correspondence. Comprehensive evaluations across 6 backbone families against 12 competitive baselines show that 1) RUPTURETRACE achieves 88.9% AUROC for controlled rupture detection and 0.694 relation-level Top-1 localization accuracy; 2) RUPTURETRACE consistently outperforms all compared baselines while reducing token usage by up to 90%.

1 INTRODUCTION

Recent advances in AI for scientific research are expanding the role of AI across the research process, from scientific reasoning and formulation to implementation and experimentation (Edwards et al., 2026; Zhao et al., 2026; Yamada et al., 2026). As AI systems participate in more stages of research, they increasingly generate heterogeneous scientific entities, including text, mathematical formulations, code, and experimental results. Yet the local plausibility of each entity does not guarantee that the evidential relations connecting them remain supported (Ho et al., 2026; Baumgärtner & Gurevych, 2026). As illustrated in Figure 1, the reported analysis attributes an empirical gain to a consistency regularizer specified in the formulation, but the implementation omits the corresponding operation. Both entities may appear reasonable in isolation, yet the evidential relation connecting them is no longer supported. As AI becomes involved in more stages of scientific research, reliably

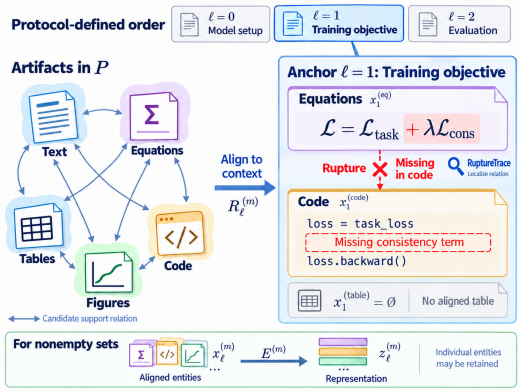


Figure 1: An illustrative cross-modal evidential rupture. RUPTURETRACE detects and localizes the rupture where the code omits the equation's consistency regularizer.

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.

identifying when these evidential relations lose support becomes increasingly important for scientific evaluation.

Existing work addresses several related but narrower aspects. Scientific claim verification asks whether claims are supported by evidence. Recent work extends this setting to multimodal evidence, source-level traceability, and graded claim-evidence alignment (Wang et al., 2025; Ho et al., 2026; Tan et al., 2026; James et al., 2026). LLM-based evaluators have also been used to assess scientific abstracts and AI-generated papers, while recent studies examine whether their scores reliably reflect scientific quality (Zheng et al., 2023; Liu et al., 2023; Yamada et al., 2026; Sajeva & Merialdo, 2026; Miyai et al., 2026). Other work tests the alignment between scientific papers and their implementations, or whether reported experiments can be reproduced (Baumgärtner & Gurevych, 2026; Zhao et al., 2026; Edwards et al., 2026). Scientific-document benchmarks increasingly evaluate integrated reasoning over multimodal evidence and across multiple papers (Ho et al., 2026; Xiong et al., 2026). These methods improve the evaluation of individual scientific entities and specific evidential relations. However, these methods focus on particular entities or predefined relation types, leaving the broader phenomenon of evidential support loss without a general characterization. As a result, it remains unclear what information is required to identify such cases across heterogeneous scientific entities and how this information can be used for systematic detection. In this paper, we study two fundamental research questions: ***Q1:*** *Can we formally characterize the loss of evidential support between scientific entities and determine what structural information is required to identify such ruptures?* ***Q2:*** *Can this characterization be instantiated as a unified framework that efficiently and effectively detects and localizes evidential ruptures across heterogeneous scientific entities?*

To address these questions, we propose a novel notion named *cross-modal evidential rupture* for characterizing the loss of support in an expected evidential relation between scientific entities. Based on this formulation, we characterize the structural information required to identify such ruptures. We represent relation-specific support across aligned entities as a support structure whose structural classes encode the continuity of evidential relations across aligned anchors. Our analysis establishes a topology-dependent boundary between local tracking, identity-preserving tracking, and cross-anchor correspondence. In particular, overlapping structures require explicit cross-anchor correspondence because local relation identity is insufficient to determine how the same evidential relation continues across anchors.

Building on this characterization, we introduce RUPTURETRACE, a topology-aware framework for cross-modal evidential rupture detection and localization. RUPTURETRACE first aligns heterogeneous scientific entities and measures directional support for their evidential relations while preserving relation-specific provenance. It then tracks these relations according to the topology of the induced support structure, where zigzag persistence provides a principled mechanism for retaining the required cross-anchor correspondence in this overlapping regime by tracking structural classes as the support structure changes. By combining relation-level support measurement with topology-aware tracking, RUPTURETRACE detects when an expected evidential relation loses support and localizes the resulting rupture to the affected relation and source evidence. Comprehensive evaluations across 6 backbone families against 12 competitive baselines show that RUPTURETRACE achieves 88.9% AUROC for rupture detection and 69.4% Top-1 localization accuracy. RUPTURETRACE consistently outperforms all compared baselines while reducing token usage by up to 90%.

2 PRELIMINARIES

In this section, we introduce the notation for scientific entities, expected evidential relations, and support structures, and formally define the problem of cross-modal evidential rupture detection and localization.

2.1 CROSS-MODAL EVIDENTIAL SUPPORT

Let $P \in \mathcal{P}$ denote a paper together with the associated entities available for evaluation, such as implementations and reported results. Let $\mathcal{M} = \{1, \ldots, M\}$ index modalities, including text, mathematical formulations, code, figures, tables, and empirical results. For modality m, let $X^{(m)}(P) \subseteq \mathcal{X}_m$ be its observed entity.

As illustrated in Figure 1, an *anchor* is a specified context in the scientific analysis, such as a claim, a methodological step, or an experimental setting, to which relevant entities is aligned. The alignment protocol orders these contexts as $\mathcal{A}(P) = \{0, \dots, T\}$; this order need not be temporal. At anchor $\ell \in \mathcal{A}(P)$, the finite set $x_\ell^{(m)} = R_\ell^{(m)}(P) \subseteq X^{(m)}(P)$ contains the aligned entities of modality m, with $x_\ell^{(m)} = \varnothing$ when that modality is unavailable. For a nonempty set, $z_\ell^{(m)} = E^{(m)}(x_\ell^{(m)}) \in \mathcal{H}$ is its representation, which may retain individual entity representations.

Let $G^\star = (\mathcal{M}, E^\star)$ specify the expected evidential relations. Each directed, labeled *relation channel* $e \in E^\star$ connects source modality $m(e)$ to target modality $n(e)$. The channel is fixed across anchors, although its aligned entities may change. Let $\mathcal{L}_e(P) \subseteq \mathcal{A}(P)$ contain the anchors where both endpoints are available and aligned. We call (e, ℓ), for $\ell \in \mathcal{L}_e(P)$, a *relation instance*. Instances with unavailable endpoints are masked rather than assigned a support loss.

A potentially asymmetric map $S_e : \mathcal{H} \times \mathcal{H} \to [0, 1]$ measures directional support for channel e. At a valid instance, define $\sigma_{e,\ell}(P) = S_e\left(z_\ell^{(m(e))}, z_\ell^{(n(e))}\right), r_{e,\ell}(P) = 1 - \sigma_{e,\ell}(P)$. The assessment also retains relation-specific provenance $q_{e,\ell}(P) \in \mathcal{Q}_e$, such as the source spans or entity locations used to measure support.

Cross-modal evidential rupture. At support scale $\epsilon \in [0, 1]$, an expected relation instance (e, ℓ) is locally unsupported when $r_{e,\ell}(P) > \epsilon$. This judgment is conditional on the specified extraction, alignment, and support map. A local loss does not, by itself, establish how the joint evidential structure changes across anchors.

To describe that structure, let $E_\ell(\epsilon) = \{e \in E^\star : \ell \in \mathcal{L}_e(P),\ r_{e,\ell}(P) \le \epsilon\}$. Let $K^\star$ be the one-dimensional incidence complex of $G^\star$, with modality vertices and labeled relation edges. At anchor ℓ, $K_\ell(\epsilon) \subseteq K^\star$ contains the available vertices and the supported edges in $E_\ell(\epsilon)$. Incidence determines the structural analysis; edge directions, relation identities, and provenance remain available for support measurement and localization. Over a fixed field $\mathbb{F}$, $H_1(K_\ell(\epsilon); \mathbb{F})$ describes cycle classes formed by jointly supported relations.

2.2 Zigzag Persistence

The sequence $K_0(\epsilon), \dots, K_T(\epsilon)$ need not be monotone: a relation can lose support and later regain it. Thus, adjacent complexes need not contain one another. Unavailable or unaligned entities remain masked and do not establish continuity across a gap.

Standard persistent homology tracks features through a monotone filtration $F_0 \subseteq \cdots \subseteq F_J$ (Edelsbrunner et al., 2002; Carlsson, 2009). Zigzag persistence permits maps in both directions (Carlsson & De Silva, 2010). Within an unmasked anchor run, we connect adjacent complexes by

$$K_\ell(\epsilon) \longrightarrow K_\ell(\epsilon) \cup K_{\ell+1}(\epsilon) \longleftarrow K_{\ell+1}(\epsilon)$$

and apply $H_1(-; \mathbb{F})$ to track cycle classes across anchors.

These maps retain correspondence information that separate support measurements or cycle ranks can omit. As formalized in Section 2.3, a unique cycle class admits local tracking, separable classes require identity-preserving tracking, and overlapping classes require explicit cross-anchor correspondence. Zigzag intervals summarize class continuity, but do not uniquely attribute a rupture to an entity. Relation identities and $q_{e,\ell}(P)$ provide the information needed for that localization. We now formally define our problem as below.

Problem 1 (Cross-Modal Evidential Rupture Detection and Localization)**.**

Given: (i) A scientific analysis $P \in \mathcal{P}$, comprising associated entities available for evaluation; **(ii)** an evaluation specification Θ that fixes the ordered anchors, extraction and alignment maps $\{R_\ell^{(m)}\}_{m,\ell}$, encoders $\{E^{(m)}\}_{m=1}^M$, expected-relation schema $G^\star$ and its incidence construction, support maps $\{S_e\}_{e \in E^\star}$, and the rules defining availability masks $\{\mathcal{L}_e(P)\}_{e \in E^\star}$; and **(iii)** a set of support scales $\Pi \subseteq [0, 1]$ at which relation support is evaluated.

Find: An evaluator $f_{\Pi;\Theta}(P) = \big(\widehat{R}_{\Pi;\Theta}(P), \widehat{\Lambda}_{\Pi;\Theta}(P)\big)$ that returns a paper-level rupture score $\widehat{R}_{\Pi;\Theta}(P) \in [0, 1]$ and source-linked localization $\widehat{\Lambda}_{\Pi;\Theta}(P) = \big\{((e, \ell), \widehat{\lambda}_{e,\ell}, q_{e,\ell}(P)) : (e, \ell) \in \mathcal{I}(P)\big\}, \widehat{\lambda}_{e,\ell} \in [0, 1]$.

We develop RUPTURETRACE in three steps. Section 2.3 identifies when local support measurements fail to preserve the identity of an evidential structure across anchors and derives a topology-dependent rule for measuring its continuity. Section 2.4 combines this structural measurement with relation-level support to construct a paper-level rupture certificate, then establishes its exact evaluation and stability. Section 2.5 links the certificate back to relation instances and source evidence.

2.3 TOPOLOGY-DEPENDENT STRUCTURAL ANALYSIS

An expected relation may appear locally supported at several anchors even when the joint evidential structure changes between them. For instance, a cycle of mutually supporting relations can disappear and be replaced by a different cycle without changing the number of cycles observed at either anchor. The following result makes this limitation of pointwise analysis explicit.

Proposition 1 (Pointwise cycle rank does not identify continuity)**.** *If the template $K^\star$ contains two distinct simple cycles C and C', there are anchor graphs $K, K' \subseteq K^\star$ with $\beta_1(K) = \beta_1(K') = 1$ whose cycle classes have different images in $H_1(K^\star; \mathbb{F})$. Hence the sequence of pointwise ranks $\{\beta_1(K_\ell(\epsilon))\}_\ell$ does not, in general, determine whether the same cycle class persists across anchors.*

We therefore retain the correspondence of cycle classes across anchors. At scale ϵ, the graph $K_\ell(\epsilon)$ contains the available vertices and the relation edges with $r_{e,\ell} \le \epsilon$; all such graphs are subgraphs of the same labeled, one-dimensional template $K^\star$. Edge directions determine how support is measured, while the underlying incidence graph determines its cycle classes. Let O_ℓ contain all available relation edges before thresholding. An anchor is *cycle-valid* when $\beta_1(O_\ell) > 0$. Availability gaps divide these anchors into maximal contiguous runs $[a_q, b_q]$; we never connect graphs across a gap.

Within each run, adjacent graphs are linked by their union $B_\ell^\epsilon = K_\ell(\epsilon) \cup K_{\ell+1}(\epsilon)$. Applying $H_1(-; \mathbb{F})$ to $K_\ell(\epsilon) \to B_\ell^\epsilon \leftarrow K_{\ell+1}(\epsilon)$ produces a zigzag barcode $b_{\mathrm{cyc}}^\epsilon$. Its intervals distinguish a class that remains supported across anchors from successive, locally valid classes. Intervals supported only at union bridges cover no anchor positions and contribute zero to the score below.

Cycle continuity. Let N_{cyc} be the number of cycle-valid anchor positions, $\rho_{\max} = \beta_1(K^\star)$, and $m_{\mathrm{cyc}}(I)$ the number of anchor positions covered by interval I, divided by N_{cyc}. When $N_{\mathrm{cyc}} > 0$ and $\rho_{\max} > 0$, define $Q_{\mathrm{cyc}}^\epsilon = \frac{1}{\rho_{\max}} \sum_{I \in b_{\mathrm{cyc}}^\epsilon} m_{\mathrm{cyc}}(I)^2$. Squaring the covered fraction rewards sustained support of the *same* class: two shorter intervals contribute less than one interval covering the same total number of anchor positions.

The topology of the available graphs determines how much computation is needed to obtain this score. Some cycle systems can be followed with ordinary runs; others require cross-anchor class correspondence.

Proposition 2 (Topology-adaptive structural computation)**.** *Fix ϵ and consider the cycle-valid runs $[a_q, b_q]$.*

(i) Unique cycle. *Suppose C is the only cycle realizable in every cycle-valid available graph O_ℓ. Set $c_\ell = \max_{e \in C} r_{e,\ell}$ and let $\mathfrak{R}_\epsilon$ be the maximal contiguous cycle-valid runs on which $c_\ell \le \epsilon$. Then $Q_{\mathrm{cyc}}^\epsilon = \rho_{\max}^{-1} \sum_{R \in \mathfrak{R}_\epsilon} (|R|/N_{\mathrm{cyc}})^2$.*

(ii) Fixed basis. *Suppose fixed cycles $C_1, \dots, C_d$ are independent in $H_1(K^\star; \mathbb{F})$, and at every anchor the cycles among them that are present in $K_\ell(\epsilon)$ span $H_1(K_\ell(\epsilon); \mathbb{F})$. If $\mathfrak{R}_{\epsilon,k}$ contains the maximal contiguous cycle-valid runs supporting C_k, then $Q_{\mathrm{cyc}}^\epsilon = \rho_{\max}^{-1} \sum_{k=1}^{d} \sum_{R \in \mathfrak{R}_{\epsilon,k}} (|R|/N_{\mathrm{cyc}})^2$.*

(iii) General case. *For $a_q \le i \le j \le b_q$, let $H_{ij}^\epsilon = \bigcap_{\ell=i}^{j} K_\ell(\epsilon)$. Without either reduction above, the exact score is*

$$Q_{\mathrm{cyc}}^\epsilon = \frac{1}{\rho_{\max} N_{\mathrm{cyc}}^2} \sum_q \sum_{a_q \le i \le j \le b_q} \big(2 - \mathbf{1}\{i = j\}\big)\, \beta_1(H_{ij}^\epsilon). \tag{1}$$

The first two cases replace class tracking with scans over a unique cycle or fixed cycle identities. In the general case, including overlapping cycles whose identities cannot be followed through a fixed basis, the intersection ranks retain the cross-anchor information encoded by the zigzag barcode. Equation 1 computes its *score* without requiring an explicit interval decomposition; the barcode remains useful when individual class trajectories are needed for inspection.

2.4 RUPTURETRACE CERTIFICATE CONSTRUCTION

We now instantiate the support maps in the evaluation specification Θ and aggregate support at three levels: relation instances, relation continuity, and cycle continuity.

Directional support at an anchor. For modality m, let $\mathcal{Z}_\ell^{(m)} = \{z_{\ell,i}^{(m)}\}_{i=1}^{n_{m,\ell}}$ be the representations aligned to anchor ℓ. Local entities are aligned within the anchor; cross-section and external entities enter only through explicit references or the predeclared alignment maps in Θ. For a valid directed relation $e : m \to n$, we instantiate the support and loss from the preliminaries as $\sigma_{e,\ell}(P) = \frac{1}{n_{m,\ell}} \sum_{i=1}^{n_{m,\ell}} \max_{1\le j\le n_{n,\ell}} \frac{1+\cos\left(z_{\ell,i}^{(m)}, z_{\ell,j}^{(n)}\right)}{2}$, $r_{e,\ell} = 1 - \sigma_{e,\ell}(P)$. Averaging the best target match for each source entity makes the measurement directional: the reverse relation need not have the same support. We retain the selected matches and their source locations as $q_{e,\ell}(P)$. In Figure 1, this record can point from a formulation term to the implementation in which the corresponding operation is missing.

Continuity within a relation channel. Let $\Omega(P) = \{e \in E^\star : \mathcal{L}_e(P) \neq \varnothing\}$. At scale ϵ, supported anchors of e satisfy $r_{e,\ell} \le \epsilon$. Write $\mathcal{I}_e^\epsilon$ for their maximal runs, connecting only anchors that were consecutive in the original axis. Thus a missing or unaligned instance splits a run even if the next valid instance is supported. We set $Q_e^\epsilon = \sum_{I\in\mathcal{I}_e^\epsilon} \left(\frac{|I|}{|\mathcal{L}_e(P)|}\right)^2$. The runs $\mathcal{I}_e^\epsilon$ give the relation's support intervals at that scale. They preserve which channel and anchors contribute to continuity, while $q_{e,\ell}(P)$ retains the underlying entities.

From continuity to rupture. Write $\omega_{\mathrm{cyc}} = \mathbf{1}\{N_{\mathrm{cyc}} > 0, \rho_{\max} > 0\}$ and assume $\omega_{\mathrm{cyc}} + |\Omega(P)| > 0$; an analysis with no evaluable relation instances is reported as unevaluable. We combine only the available channels:

$$Q^\epsilon(P) = \frac{\omega_{\mathrm{cyc}} Q_{\mathrm{cyc}}^\epsilon + \sum_{e\in\Omega(P)} Q_e^\epsilon}{\omega_{\mathrm{cyc}} + |\Omega(P)|}, \widehat{F}_\Theta(P) = \int_0^1 Q^\epsilon(P)\, d\epsilon, \widehat{R}_{\Pi;\Theta}(P) = \widehat{F}_\Theta^\star(P) - \widehat{F}_\Theta(P). \quad (2)$$

Here we take the scale domain in Problem 1 to be $\Pi = [0, 1]$. The mask-matched attainable value $\widehat{F}_\Theta^\star(P)$ is computed after setting all *valid* relation losses to zero while preserving the original anchors, alignments, and availability masks. For every scale, $0 \le Q_e^\epsilon, Q_{\mathrm{cyc}}^\epsilon \le 1$ when the corresponding channel is available. Adding supported edges cannot reduce either continuity score, so $0 \le \widehat{F}_\Theta(P) \le \widehat{F}_\Theta^\star(P) \le 1$. The difference in Eq. 2 measures lost attainable support without turning missing entities into rupture.

Although the certificate integrates over all scales, only observed relation losses can change a thresholded graph. The following result gives its finite implementation.

Proposition 3 (Exact evaluation at critical scales)**.** *Let $\Pi(P) = \{0, 1\} \cup \{r_{e,\ell} : (e, \ell) \in \mathcal{I}(P)\}$ have distinct ordered values $0 = \eta_0 < \eta_1 < \cdots < \eta_J = 1$. Then $Q^\epsilon(P)$ is constant on each $[\eta_j, \eta_{j+1})$, and $\widehat{F}_\Theta(P) = \sum_{j=0}^{J-1} (\eta_{j+1} - \eta_j) Q^{\eta_j}(P)$.*

Thus $\Pi(P)$ is an exact evaluation grid for the integrated certificate, not a numerical approximation. The score also remains controlled when the measured losses change slightly.

Theorem 1 (Conditional stability of the integrated certificate)**.** *Suppose P and $\widetilde{P}$ share the same ordered anchors, relation schema, alignment, availability masks, and channel normalizations. If $\delta = \max_{(e,\ell)\in\mathcal{I}(P)} |r_{e,\ell}(P) - r_{e,\ell}(\widetilde{P})|$, then $|\widehat{F}_\Theta(P) - \widehat{F}_\Theta(\widetilde{P})| \le \delta, \quad |\widehat{R}_{\Pi;\Theta}(P) - \widehat{R}_{\Pi;\Theta}(\widetilde{P})| \le \delta.$*

The thresholded graphs of the two analyses are interleaved by a shift of at most δ. Relation and cycle continuity are bounded, monotone functions of the scale; the two analyses have the same mask-matched attainable value. This guarantee concerns the *integrated score* under fixed structural inputs. It does not assert that every interval endpoint or source attribution is invariant to perturbation.

2.5 RUPTURE DETECTION AND LOCALIZATION

The paper-level score indicates how much attainable support has been lost. To identify where the loss occurs, we return each evaluable instance's measured loss and the entities used to calculate it:

$$\widehat{\lambda}_{e,\ell} = r_{e,\ell}, \widehat{\Lambda}_{\Pi;\Theta}(P) = \big\{((e,\ell), \widehat{\lambda}_{e,\ell}, q_{e,\ell}(P)) : (e,\ell) \in \mathcal{I}(P)\big\}, f_{\Pi;\Theta}(P) = \big(\widehat{R}_{\Pi;\Theta}(P), \widehat{\Lambda}_{\Pi;\Theta}(P)\big). \tag{3}$$

The pair $(\widehat{\lambda}_{e,\ell}, q_{e,\ell})$ identifies a relation, an anchor, and source locations for inspection. Its value measures local support loss; interpreting an instance as unsupported at a particular scale still uses the condition $r_{e,\ell} > \epsilon$ from the preliminaries.

Optional structural attribution. When the cycle channel is available, we can additionally quantify how repairing each relation channel would recover cycle continuity. For $S \subseteq \Omega(P)$, define $P^{\oplus S}$ by setting $r_{e,\ell} = 0$ for every valid instance of each $e \in S$, leaving all other losses and all masks fixed. Then we have

$$v_P(S) = \int_0^1 Q^{\epsilon}_{\mathrm{cyc}}(P^{\oplus S})\, d\epsilon, \qquad \phi_e(P) = \mathrm{Shapley}_e(v_P). \tag{4}$$

The Shapley efficiency identity gives $\sum_{e\in\Omega(P)} \phi_e(P) = v_P(\Omega(P)) - v_P(\varnothing)$: these values allocate the *recoverable cycle-support loss*.

3 EXPERIMENTS

We conduct a comprehensive evaluation of RUPTURETRACE across six backbone families and 10 baseline families. Our experiments address the following five questions: ***Q1:*** *How effective does* RUPTURETRACE *detect and localize cross-modal evidential ruptures?* ***Q2:*** *How efficient does* RUPTURETRACE *detect and localize cross-modal evidential ruptures?* ***Q3:*** *How consistently does* RUPTURETRACE *perform across backbone families?* ***Q4:*** *How does* RUPTURETRACE *use the topology of evidential support to detect and localize a rupture, and what evidence supports its localization?* ***Q5:*** *Can* RUPTURETRACE *be instantiated as an advanced reviewing scaffold?*

3.1 EXPERIMENTAL SETTINGS

Baselines. We compare with 12 competitive baselines in four categories: (1) ***general-purpose LLMs***, including GPT-5, GPT-5.4 (Singh et al., 2025), Gemini-2.5-Pro (Comanici et al., 2025), Gemini-3.1-Pro (Google DeepMind, 2026), DeepSeek-V3.2 (DeepSeek-AI, 2025), Qwen3 (235B, VL-235B) (Team, 2025), Qwen3.5 (Team, 2026), GLM-4.6V (Team et al., 2025), Mistral-Large-3 (Mistral AI, 2025), and Phi-4-reasoning (Abdin et al., 2025), evaluated in zero-shot and in-context-learning settings; (2) ***reviewer models***, including CycleReviewer (Weng et al., 2025) and DeepReviewer (Zhu et al., 2025); (3) ***agentic scientific-evaluation models***, including AI Scientist-V2 (Yamada et al., 2025); and (4) ***structured-review scaffolds***, including LimitGen (Xu et al., 2025) and TreeReview (Chang et al., 2025). Full model versions, prompts, and input interfaces are reported in Appendix B.

Backbones. We evaluate on 6 backbone families, including GPT2 (Radford et al., 2019), SciBERT (Beltagy et al., 2019), Specter2 (Singh et al., 2023), CLIP-ViT (Yang et al., 2024), SigLIP (Tschannen et al., 2025), and Qwen-3.5(2B, 4B, 9B, 27B, 122B and 397B) (Team, 2026).

Benchmark. We construct an ICLR (OpenReview, 2017-2026) corpus from OpenReview submissions spanning 2017-2026. ICLR 2026 is held-out evaluation.

Metrics. Following prior work (Kang et al., 2018; Zhu et al., 2025), downstream evaluation reports accuracy (Acc), Accept-F1, Macro-F1, Spearman ρ, and MSE against reviewer-assigned overall assessment (OA), coherence (Coh.), and novelty (Nov.). Direct support evaluation reports AUROC, relation Top-k, anchor F1, and clean-ruptured score separation. Full dataset statistics, baseline configurations, and metric definitions are in Appendix B.

3.2 EFFECTIVENESS IN RUPTURE DETECTION AND LOCALIZATION

Table 1 compares paper-level alignment with reviewer outcomes. RUPTURETRACE leads all listed methods across the nine reported metrics. It achieves 0.772 accuracy, 0.641 Accept-F1, and 0.737

Macro-F1, exceeding the strongest competing result on each metric by 0.033, 0.039, and 0.039, respectively. Its OA, coherence, and novelty correlations reach 0.552, 0.548, and 0.482; the corresponding MSE values of 1.104, 0.142, and 0.198 are also the lowest in the table. The agreement across classification, ranking, and prediction-error metrics supports an association between the RUPTURETRACE score and reviewer assessments. Reviewer outcomes do not provide relation-level rupture labels, so this comparison does not establish detection or localization accuracy.

Table 1: Detailed Performance of all evaluation methods on decision-level and scoring metrics, including overall assessment (OA), coherence (Coh.), and novelty (Nov.). Superscripts denote the input modality: † direct document file input, ‡ OCR-derived text from paper documents, and § page-level images of the document. Best results are **bold** and second-best results are underlined.

Method	Acc ↑	Accept F1 ↑	Macro-F1 ↑	OAρ ↑	OA MSE ↓	Coh.ρ ↑	Coh. MSE ↓	Nov.ρ ↑	Nov. MSE ↓
Zero-shot									
GPT-5.4†	0.725	0.553	0.677	0.484	1.231	0.463	0.155	0.409	0.331
GPT-5†	0.650	0.583	0.641	0.461	2.035	0.468	0.370	0.381	0.640
Gemini 2.5 Pro†	0.377	0.504	0.333	0.392	7.200	0.364	0.987	0.370	1.524
Gemini 3.1 Pro†	0.687	0.601	0.672	0.505	1.809	0.506	0.225	0.451	0.364
DeepSeek-V3.2‡	0.634	0.522	0.613	0.343	2.447	0.334	0.655	0.277	1.180
Qwen3-235B-A22B‡	0.327	0.490	0.251	0.310	5.897	0.396	1.028	0.312	1.799
Qwen3-VL-235B-A22B§	0.340	0.493	0.274	0.330	5.644	0.298	0.892	0.272	1.461
Qwen3.5-397B-A17B‡	0.456	0.529	0.443	0.408	3.096	0.425	0.685	0.324	0.993
GLM-4.6V§	0.312	0.476	0.238	0.037	6.942	0.470	0.771	0.433	1.678
Mistral-Large-3§	0.335	0.486	0.272	0.357	4.345	0.328	0.728	0.270	1.481
Phi-4-reasoning‡	0.407	0.504	0.383	-0.019	3.585	0.142	0.919	0.077	1.737
Qwen3.5-397B-A17B§	0.350	0.492	0.295	0.396	4.329	0.379	0.934	0.370	1.102
In Context Learning									
GPT-5.4†	0.739	0.574	0.693	0.462	1.319	0.465	0.153	0.413	0.216
GPT-5†	0.705	0.587	0.679	0.490	1.692	0.474	0.218	0.419	0.378
Gemini 2.5 Pro†	0.395	0.516	0.354	0.471	7.577	0.320	0.674	0.411	1.168
Gemini 3.1 Pro†	0.728	0.602	0.698	0.493	1.970	0.499	0.189	0.443	0.291
DeepSeek-V3.2‡	0.691	0.368	0.582	0.470	2.076	0.498	0.235	0.341	0.433
Qwen3-235B-A22B‡	0.348	0.498	0.285	0.382	5.581	0.401	0.471	0.312	0.908
Qwen3-VL-235B-A22B§	0.334	0.491	0.264	0.344	7.097	0.291	0.478	0.289	0.896
GLM-4.6V§	0.330	0.489	0.259	0.291	8.800	0.236	0.625	0.178	1.290
Mistral-Large-3§	0.500	0.545	0.495	0.371	3.167	0.332	0.378	0.286	0.771
Phi-4-reasoning‡	0.357	0.500	0.299	0.127	13.269	0.307	0.552	0.119	1.053
Qwen3.5-397B-A17B§	0.390	0.512	0.349	0.363	4.807	0.411	0.736	0.307	0.855
Qwen3.5-397B-A17B‡	0.554	0.567	0.553	0.477	2.772	0.442	0.370	0.364	0.603
Specialized Evaluation methods									
AI-Scientist‡	0.662	0.118	0.455	0.161	4.170	0.156	0.290	0.078	0.546
CycleReviewer‡	0.539	0.428	0.521	0.137	2.561	0.161	0.234	0.100	0.350
DeepReviewer‡	0.647	0.476	0.605	0.305	1.780	0.300	0.170	0.170	0.290
RUPTURETRACE **(Ours)**	**0.772**	**0.641**	**0.737**	**0.552**	**1.104**	**0.548**	**0.142**	**0.482**	**0.198**

3.3 EFFICIENCY IN RUPTURE DETECTION AND LOCALIZATION

Table 2 compares accuracy, latency, and token use under the reported interfaces. RUPTURETRACE achieves the highest accuracy (0.772) with the lowest average latency (6.15s) and token count (3.4K). Compared with GPT-5.4 ICL, the strongest listed accuracy competitor, it improves accuracy by 0.033 while using 4.1× less average time and 6.2× fewer counted tokens. Even relative to Gemini-3.1-Pro, the least token-intensive comparator, RUPTURETRACE uses 49.3% fewer tokens and achieves higher accuracy (0.772 versus 0.687). These comparisons describe the observed accuracy-cost trade-off under each system's reported interface; token counts across different input interfaces are not a controlled measure of identical computation.

3.4 ROBUSTNESS ACROSS DIFFERENT BACKBONE FAMILIES AND DIFFERENT MODEL SCALES

Table 3 examines backbone choice with RUPTURETRACE dependency modeling held fixed. Across the six family representatives, accuracy ranges from 0.671 to 0.805, while OA correlation ranges from 0.356 to 0.591. SPECTER2 performs best among the non-Qwen backbones; Qwen3.5-122B achieves the highest values overall. Because these checkpoints differ substantially in size, the cross-family comparison demonstrates backbone sensitivity but does not isolate an architectural effect.

Table 2: Per-paper operational cost under the reported interfaces. Token counts are interface-specific; controlled local results are in Table 10.

Method	Acc. ↑	Avg. Latency (s)↓	Med. Latency (s) ↓	Avg. Tokens ↓
GPT-5.4	0.725	19.43	18.34	19.5K
GPT-5.4 ICL	0.739	25.29	23.71	21.2K
Gemini-3.1-Pro	0.687	18.03	16.87	6.7K
Gemini-3.1-Pro ICL	0.728	17.55	15.29	7.3K
AI-Scientist	0.662	32.84	32.31	34.0K
CycleReviewer	0.539	706.71	828.29	15.2K
DeepReviewer	0.647	65.83	57.76	9.3K
RUPTURETRACE	**0.772**	**6.15**	**5.72**	**3.4K**

Table 3: Performance of RUPTURETRACE across backbone families and model scales. **(a)** One representative checkpoint per family, with dependency modeling held fixed. **(b)** Model-scale comparison within Qwen3.5.

(a) Across backbone families

Backbone	Acc. ↑	Accept-F1 ↑	OA ρ ↑	OA MSE ↓
GPT-2	0.671	0.522	0.356	1.39
SciBERT	0.718	0.574	0.447	1.21
SPECTER2	0.739	0.601	0.489	1.12
CLIP-ViT	0.704	0.559	0.421	1.27
SigLIP	0.723	0.583	0.462	1.18
Qwen3.5-122B	0.805	0.744	0.591	0.997

(b) Model scale within Qwen3.5

Model	Acc. ↑	Accept-F1 ↑	OA ρ ↑	OA MSE ↓
Qwen3.5-2B	0.772	0.641	0.552	1.104
Qwen3.5-4B	0.779	0.658	0.559	1.086
Qwen3.5-9B	0.786	0.683	0.568	1.061
Qwen3.5-27B	0.797	0.724	0.583	1.018
Qwen3.5-122B	0.805	0.744	0.591	0.997

Within the Qwen3.5 family, all four metrics improve from 2B through 122B. Accuracy rises from 0.772 to 0.805, Accept-F1 from 0.641 to 0.744, and OA correlation from 0.552 to 0.591, while OA MSE falls from 1.104 to 0.997. The final increase from 27B to 122B adds 0.008 accuracy and 0.008 OA correlation, indicating smaller gains at the upper end of the tested range. Together, the two panels show how paper-level reviewer-outcome performance varies across backbone families and improves with scale within Qwen3.5.

3.5 TOPOLOGY HELPS DETECT AND TRACE RUPTURES

Table 4 isolates cross-anchor structural modeling. In the no-switch stratum, the classwise run scan and zigzag persistence obtain the same AUROC (0.878) and repair Top-1 (0.670), consistent with the fixed-basis condition. When cycle correspondence changes across anchors, zigzag persistence increases AUROC from 0.866 to 0.914 and repair Top-1 from 0.636 to 0.718 relative to the classwise scan. Direct anchor F1 remains 0.618 across variants because the anchor score $r_{e,\ell}$ is held fixed.

Table 4: Topology-sensitive rupture detection and tracing. Structural variants with the backbone, relation scores, and evaluation split held fixed. No-switch cases admit a fixed cycle basis; switch cases require cross-anchor cycle correspondence. Repair Top-1 evaluates repair-attribution, whereas anchor F1 evaluates $\widehat{\lambda}_{e,\ell} = r_{e,\ell}$.

	Rupture AUROC ↑			Repair Top-1 ↑		Anchor F1 ↑
Method	Overall	No-switch	Switch	No-switch	Switch	Overall
Relation only	0.848	0.862	0.817	0.620	0.600	0.618
Pointwise β_1	0.861	0.867	0.843	0.642	0.628	0.618
Classwise run scan	0.875	**0.878**	0.866	**0.670**	0.636	0.618
Zigzag persistence	**0.889**	**0.878**	**0.914**	**0.670**	**0.718**	0.618

Table 5 separates structural continuity from attribution. Structural continuity raises rupture AUROC from 0.827 to 0.889 and the clean-corrupted score gap from 0.196 to 0.247. On the same cycle-eligible interventions and candidate set, PRA replaces the no-PRA ranking $1 - \int Q_e^{\epsilon} d\epsilon$ by ϕ_e, raising Top-1 from 0.422 to 0.642 and anchor F1 from 0.447 to 0.618 while preserving the document-level

Table 5: Controlled cross-modal rupture detection and localization. Anchor F1 measures recovery of the modified structural positions.

Method	Rupture AUROC ↑	Relation Top-1 ↑	Relation Top-2 ↑	Anchor F1 ↑	Δ_{score} ↑
Feature concatenation	0.684	-	-	-	0.083
Global persistent homology	0.761	-	-	-	0.141
Cycle-only zigzag	0.793	-	-	-	0.158
Edge scores w/o persistence	0.827	0.563	0.748	0.514	0.196
RUPTURETRACE + persistent-weakness readout	0.889	0.422	0.860	0.447	0.247
RUPTURETRACE + repair-attribution readout	**0.889**	**0.642**	**0.861**	**0.618**	**0.247**

score. The direct-edge comparator instead ranks mean $r_{e,\ell}$ (Top-1 0.563). The paper-level score contributes rupture sensitivity; the optional repair-attribution readout improves relation and anchor localization under the stated eligibility conditions.

3.6 ADVANCED REVIEWING SCAFFOLD

Table 6 shows that RUPTURETRACE-guided GPT-5.4 performs best on all four metrics. Compared with the strongest alternative, TreeReview-style prompting, it improves accuracy by 0.120, OA correlation by 0.050, rupture AUROC by 0.144, and edge localization accuracy by 0.251. These results support using explicit, paper-specific evidential relations as a reviewing scaffold for detecting and localizing support failures.

Table 6: Controlled Review Scaffold comparison. Edge Loc. Acc. denotes top-1 relation-edge localization accuracy.

Reviewing scaffold	**Acc.** ↑	**OA** ρ ↑	**Rupture AUROC** ↑	**Edge Loc. Acc.** ↑
Generic prompt	0.725	0.484	0.702	0.371
LimitGen-style	0.739	0.503	0.762	0.457
TreeReview-style	0.753	0.523	0.806	0.516
RUPTURETRACE**-guided**	**0.873**	**0.573**	**0.950**	**0.767**

4 RELATED WORK

Automated scientific evaluation. LLM-as-judge and science-specific reviewer systems score scientific content or generate review-oriented feedback (Zheng et al., 2023; Liu et al., 2023; Dycke et al., 2023; Liang et al., 2023; Yuan et al., 2022; Gottweis et al., 2026; Chan et al., 2024). Multimodal evaluators extend the evidence base to figures and tables (Singh et al., 2025).

Topological analysis. Persistent homology summarizes structural features across scales (Carlsson, 2009; Carlsson & Gabrielsson, 2018) and has been applied to language-model representations and dynamics (Yin et al., 2024; Perez & Reinauer, 2022; Gardinazzi et al., 2025). Zigzag persistence extends this framework to non-monotone sequences with both additions and removals (Carlsson & De Silva, 2010) and has been used for dynamic networks, time series, evolving point clouds, and recent LLM analyses (Su et al., 2026; Pham et al., 2025; Shafei et al., 2025).

5 CONCLUSION

Recent advances in AI for scientific research are expanding the role of AI across the research process. We propose a novel notion named *cross-modal evidential rupture* for characterizing the loss of evidential support between scientific entities. Starting from this formulation, e derive a topology-dependent characterization to capture rupture. Building on this characterization, we introduce RUPTURETRACE, a topology-aware framework for cross-modal evidential rupture detection and localization. Comprehensive evaluations across 6 backbone families against 12 competitive baselines show that RUPTURETRACE achieves 88.9% AUROC for rupture detection and 69.4% Top-1 localization accuracy. It consistently outperforms all compared baselines while reducing token usage by up to 90%.

REPRODUCIBILITY STATEMENT

We provide the definitions, assumptions, and theoretical characterization proofs underlying RUPTURETRACE in Appendix A. Appendix B documents dataset collection and split construction, PDF processing, modality extraction and alignment, representation and support computation, backbone and baseline configurations, certificate evaluation, evaluation metrics, controlled rupture construction, randomization procedures, and uncertainty estimation. Appendix C provides additional ablations, sensitivity analyses, and validation experiments; Appendix D reports the LLM evaluation prompts; and Appendix E specifies the computational resources used.

REFERENCES

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, et al. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*, 2025.

Tim Baumgärtner and Iryna Gurevych. SciCoQA: Quality assurance for scientific paper–code alignment. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38740–38770, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1795. URL https://aclanthology.org/2026.acl-long.1795/.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 3613–3618. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1371. URL https://doi.org/10.18653/v1/D19-1371.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46:255–308, 2009.

Gunnar Carlsson and Vin De Silva. Zigzag persistence. *Found Comput Math 10*, pp. 367–405, 2010.

Gunnar E. Carlsson and Rickard Brüel Gabrielsson. Topological approaches to deep learning. *CoRR*, abs/1811.01122, 2018. URL http://arxiv.org/abs/1811.01122.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. Treereview: A dynamic tree of questions framework for deep and efficient llm-based scientific peer review. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 15651–15682. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.790. URL https://doi.org/10.18653/v1/2025.emnlp-main.790.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models. *CoRR*, abs/2512.02556, 2025. doi: 10.48550/ARXIV.2512.02556. URL https://doi.org/10.48550/arXiv.2512.02556.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. Nlpeer: A unified resource for the computational study of peer review. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 5049–5073. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.277. URL https://doi.org/10.18653/v1/2023.acl-long.277.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discret. Comput. Geom.*, 28(4):511–533, 2002. doi: 10.1007/S00454-002-2885-2. URL https://doi.org/10.1007/s00454-002-2885-2.

Nicholas Edwards, Yukyung Lee, Yujun Audrey Mao, Yulu Qin, Sebastian Schuster, and Najoung Kim. Rexbench: Can coding agents autonomously implement AI research extensions? In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 16380–16417. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.745. URL https://doi.org/10.18653/v1/2026.acl-long.745.

Yuri Gardinazzi, Karthik Viswanathan, Giada Panerai, Alessio Ansuini, Alberto Cazzaniga, and Matteo Biagetti. Persistent topological features in large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/gardinazzi25a.html.

Rita González-Márquez and Dmitry Kobak. Learning representations of learning representations. *arXiv preprint arXiv:2404.08403*, 2024.

Google DeepMind. Gemini 3.1 Pro. https://deepmind.google/models/gemini/pro/, 2026. Accessed: 2026-05-05.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. Accelerating scientific discovery with co-scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y.

Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Tian Cheng Xia, Florian Boudin, Andre Greiner-Petter, and Akiko Aizawa. SciClaimEval: Cross-modal claim verification in scientific papers. In Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral (eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 11060–11071, Palma de Mallorca, Spain, May 2026. ELRA Language Resource Association. doi: 10.63317/4ap9rg2gnwmf. URL https://aclanthology.org/2026.lrec-1.864/.

Maximilian Idahl and Zahra Ahmadi. Openreviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 550–562, 2025.

Joseph James, Chenghao Xiao, Yucheng Li, Nafise Sadat Moosavi, and Chenghua Lin. RIGOURATE: Quantifying scientific exaggeration with evidence-aligned claim evaluation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34022–34043, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1699. URL https://aclanthology.org/2026.findings-acl.1699/.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard H. Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and NLP applications. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1647–1661. Association for Computational Linguistics, 2018. doi: 10.18653/V1/N18-1149. URL https://doi.org/10.18653/v1/n18-1149.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *CoRR*, abs/2310.01783, 2023. doi: 10.48550/ARXIV.2310.01783. URL https://doi.org/10.48550/arXiv.2310.01783.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.153. URL https://doi.org/10.18653/v1/2023.emnlp-main.153.

Mistral AI. Mistral Large 3 675B Instruct 2512, 2025. URL https://huggingface.co/mistralai/Mistral-Large-3-675B-Instruct-2512.

Atsuyuki Miyai, Mashiro Toyooka, Zaiying Zhao, Kenta Watanabe, Toshihiko Yamasaki, and Kiyoharu Aizawa. Paper reconstruction evaluation: Evaluating presentation and hallucination in ai-written papers. *CoRR*, abs/2604.01128, 2026. doi: 10.48550/ARXIV.2604.01128. URL https://doi.org/10.48550/arXiv.2604.01128.

OpenReview. ICLR Conference. https://openreview.net/group?id=ICLR.cc, 2017-2026. International Conference on Learning Representations.

Jiefu Ou, William Gantt Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, et al. Claimcheck: How grounded are llm critiques of scientific papers? In *EMNLP (Findings)*, pp. 21712–21735, 2025.

Ilan Perez and Raphael Reinauer. The topological BERT: transforming attention into topology for natural language processing. *CoRR*, abs/2206.15195, 2022. doi: 10.48550/ARXIV.2206.15195. URL https://doi.org/10.48550/arXiv.2206.15195.

Phu Pham, Quang-Thinh Bui, Ngoc Thanh Nguyen, Robert Kozma, Philip S. Yu, and Bay Vo. Topological data analysis in graph neural networks: Surveys and perspectives. *IEEE Trans. Neural Networks Learn. Syst.*, 36(6):9758–9776, 2025. doi: 10.1109/TNNLS.2024.3520147. URL https://doi.org/10.1109/TNNLS.2024.3520147.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Alessandro Sajeva and Paolo Merialdo. Robustness and authorship bias of large language models in scientific abstracts scoring. *Discov. Artif. Intell.*, 6(1):530, 2026. doi: 10.1007/S44163-026-01295-Z. URL https://doi.org/10.1007/s44163-026-01295-z.

Lukas Selch, Yufang Hou, Muhammad Jehanzeb Mirza, Sivan Doveh, James R Glass, Rogerio Feris, and Wei Lin. Prismm-bench: A benchmark of peer-review grounded multimodal inconsistencies. In *International Conference on Learning Representations*, volume 2026, pp. 132494–132533, 2026.

Hassan A. Shafei, Yang Zhao, Christian Martin, Javad M. Alizadeh, Karin M. Eyrich-Garg, Omar Martinez, Ying Ding, Chiu C. Tan, Huanmei Wu, and Yuzhou Chen. Safety and public protection: Predicting and analyzing incidents with large language model-based zigzag graph neural networks. In Xintao Wu, Myra Spiliopoulou, Can Wang, Vipin Kumar, Longbing Cao, Yanqiu Wu, Yu Yao, and Zhangkai Wu (eds.), *Advances in Knowledge Discovery and Data Mining - 29th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2025, Sydney, NSW, Australia, June 10-13, 2025, Proceedings, Part II*, volume 15871 of *Lecture Notes in Computer Science*, pp. 302–313. Springer, 2025. doi: 10.1007/978-981-96-8173-0_24. URL https://doi.org/10.1007/978-981-96-8173-0_24.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 5548–5566. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.338. URL https://doi.org/10.18653/v1/2023.emnlp-main.338.

Guijin Son, Jiwoo Hong, Honglu Fan, Heejeong Nam, Hyunwoo Ko, Seungwon Lim, Jinyeop Song, Jinha Choi, Gonçalo Paulo, Youngjae Yu, et al. When ai co-scientists fail: Spot-a benchmark for automated verification of scientific research. *arXiv preprint arXiv:2505.11855*, 2025.

Zhe Su, Xiang Liu, Layal Bou Hamdan, Vasileios Maroulas, Jie Wu, Gunnar Carlsson, and Guo-Wei Wei. Topological data analysis and topological deep learning beyond persistent homology: a review. *Artif. Intell. Rev.*, 59(2):58, 2026. doi: 10.1007/S10462-025-11462-W. URL https://doi.org/10.1007/s10462-025-11462-w.

Neset Tan, Minghao Li, and Mark Gahegan. SciTrue: Evidence-grounded claim verification in science. In Danilo Croce, Jochen Leidner, and Nafise Sadat Moosavi (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 397–406, Rabat, Marocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-382-1. doi: 10.18653/v1/2026.eacl-demo.27. URL https://aclanthology.org/2026.eacl-demo.27/.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Qwen Team. Qwen3.5-omni technical report. *CoRR*, abs/2604.15804, 2026. doi: 10.48550/ARXIV.2604.15804. URL https://doi.org/10.48550/arXiv.2604.15804.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025. doi: 10.48550/ARXIV.2502.14786. URL https://doi.org/10.48550/arXiv.2502.14786.

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. SciVer: Evaluating foundation models for multimodal scientific claim verification. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8562–8579, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.420. URL https://aclanthology.org/2025.acl-long.420/.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth*

International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net, 2025. URL https://openreview.net/forum?id=bjcsVLoHYs.

Lei Xiong, Huaying Yuan, Zheng Liu, Zhao Cao, and Zhicheng Dou. Paperscope: A multi-modal multi-document benchmark for agentic deep research across massive scientific papers. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 8015–8040. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.394. URL https://doi.org/10.18653/v1/2026.findings-acl.394.

Zhijian Xu, Yilun Zhao, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. Can llms identify critical limitations within scientific research? A systematic evaluation on AI research papers. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 20652–20706. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1009. URL https://doi.org/10.18653/v1/2025.acl-long.1009.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob N. Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *CoRR*, abs/2504.08066, 2025. doi: 10.48550/ARXIV.2504.08066. URL https://doi.org/10.48550/arXiv.2504.08066.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Chris Lu, Shengran Hu, Jakob N. Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *CoRR*, abs/2606.15497, 2026. doi: 10.48550/ARXIV.2606.15497. URL https://doi.org/10.48550/arXiv.2606.15497.

Han Yang, Minxia Xu, Zhiyong Sun, Bo Song, and Erkang Cheng. Clip-vit detector: Side adapter with prompt for vision transformer object detection. In *7th International Conference on Algorithms, Computing and Artificial Intelligence, ACAI 2024, Guangzhou, China, December 20-22, 2024*, pp. 1–8. IEEE, 2024. doi: 10.1109/ACAI63924.2024.10899669. URL https://doi.org/10.1109/ACAI63924.2024.10899669.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. Characterizing truthfulness in large language model generations with local intrinsic dimension. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57069–57084. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yin24c.html.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *J. Artif. Intell. Res.*, 75:171–212, 2022. doi: 10.1613/JAIR.1.12862. URL https://doi.org/10.1613/jair.1.12862.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, Biqing Qi, and Bowen Zhou. ReviewRL: Towards automated scientific review with RL. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16931–16943, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.857. URL https://aclanthology.org/2025.emnlp-main.857/.

Tianmai M Zhang and Neil F Abernethy. Reviewing scientific papers for critical problems with reasoning llms: Baseline approaches and automatic evaluation. *arXiv preprint arXiv:2505.23824*, 2025.

Xuanle Zhao, Zilin Sang, Yuxuan Li, Qi Shi, Weilun Zhao, Shuo Wang, Duzhen Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. Autoreproduce: Automatic AI experiment reproduction with paper lineage. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026, pp. 21920–21942. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1001. URL https://doi.org/10.18653/v1/2026.acl-long.1001.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 29330–29355. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1420. URL https://doi.org/10.18653/v1/2025.acl-long.1420.

A Proofs and Additional Theoretical Results

A.1 Notation and Support-Map Assumptions

Throughout, the structural axis, alignment, global availability mask, and local relation masks are fixed when comparing P and $\widetilde{P}$. As in Sections 2.1 and 2.4, $\mathcal{L}_e(P)$ contains exactly the anchors with aligned, nonempty endpoint object sets that satisfy the channel's predeclared link gate; the primary e_{13} channel additionally requires a resolved textual reference to the code or stated algorithm, as specified in Appendix B.3. Further, $\Omega(P) = \{e \in E^\star : \mathcal{L}_e(P) \neq \varnothing\}$. The shared representations $\mathcal{Z}_\ell^{(m)} = E^{(m)}(x_\ell^{(m)})$ retain individual objects; $r_{e,\ell} = 1 - S_e(\mathcal{Z}_\ell^{(m(e))}, \mathcal{Z}_\ell^{(n(e))})$ is defined only on $\mathcal{L}_e(P)$. Comparisons below are on the common evaluable instance set $\mathcal{I}(P) = \mathcal{I}(\widetilde{P})$; the theorem in the main text assumes the measured field satisfies $\delta = \max_{(e,\ell)\in\mathcal{I}(P)} |r_{e,\ell}(P) - r_{e,\ell}(\widetilde{P})|$. When no paper argument is displayed, Ω abbreviates the common set $\Omega(P) = \Omega(\widetilde{P})$. The score and attribution applicability conditions are stated in Appendix B.6.

Cycle normalizations. Let O_ℓ be the locally available graph and $N_{\mathrm{cyc}} = |\{\ell : \beta_1(O_\ell) > 0\}|$. For each cycle interval, only its original anchor nodes enter its mass. We set $\rho_{\max} = \beta_1(K^\star)$ for the incidence template fixed before evaluation. This gives $\rho_{\max} = 2$ for the primary six-edge G_θ and its alternate single-chord sensitivity templates; the nested canonical five-edge diagnostic has $\rho_{\max} = 1$. Local masking can reduce $\beta_1(O_\ell)$ or make the channel unavailable, but it does not change the template-level normalization. The same value is retained under virtual repair and in every paired comparison, as required by the stability argument.

Availability affects the score ceiling. If the valid relation runs have sizes $n_{e,1}, \ldots, n_{e,k}$ and $N_e = |\mathcal{L}_e(P)|$, even perfect support gives $Q_e^1 = \sum_q (n_{e,q}/N_e)^2$, which can be below one. Thus masking does not assign a negative support value to missing evidence, but it can change continuity, channel composition, and the attainable score. Comparisons across masks must be interpreted with their coverage and normalization. No invariance to deleting or relocating masked anchors follows from the stability theorem.

Assumption 1 (Fixed-object support comparison, A1)**.** When comparing object-level coverage directly, the two evaluations have matched nonempty source and target object sets with unchanged cardinalities and object identities; the corresponding encoded vectors are unit normalized. Let $s_{ij}^e = [1 + \langle z_i^{(m(e))}, z_j^{(n(e))}\rangle]/2$ and define $d_{\mathrm{obj},e} = \max_{i,j} |s_{ij}^e - \widetilde{s}_{ij}^e|$. Then the support map in Section 2.4 obeys

$$|S_e - \widetilde{S}_e| \le d_{\mathrm{obj},e}.$$

Indeed, the maximum over target objects is non-expansive in the supremum norm on matched pairwise similarities, and averaging over source objects preserves the bound. For matched unit vectors, $d_{\mathrm{obj},e} \le \frac{1}{2}(\Delta_{m(e)} + \Delta_{n(e)})$, where Δ_m is the maximum Euclidean displacement among matched objects of modality m on valid anchors. This optional object-level consequence does not strengthen Theorem 1, which is stated directly in terms of the loss-field distance δ. Changes to cardinality, object matching, or alignment are outside A1 and are probed empirically in Appendix C.3.6.

B Detailed Experimental Setup

This appendix provides additional details on benchmark construction, baseline evaluation protocols, target construction, and evaluation metrics.

B.1 Data Collection and Statistics

We constructed our dataset from the ICLR paper metadata released by González-Márquez & Kobak (2024). We further queried OpenReview with the API version corresponding to each venue year and downloaded the submission records together with the original submission PDFs.[1] This setup provides the paper-level metadata and raw submission files required for our paper evaluation setting.

[1] https://docs.openreview.net/getting-started/using-the-api

Table 7: Final dataset statistics after decision filtering, PDF sanitization, manual test-set quality control.

Split	Years	Total	Accepted	Accept rate	Reject	Reject rate
Development	2017-2024	19,185	6,897	35.95%	12,288	64.05%
Test	2025	7,281	2,337	32.10%	4,944	67.90%
Test	2026	13,763	4,954	36.00%	8,809	64.00%

For each submission, we normalized the final decision string and applied a consistent labeling rule. Submissions whose decision contained `withdraw` or `desk reject` were removed, submissions whose decision contained `accept` were labeled as accepted, and all remaining valid submissions were labeled as `Rejected`. We exclude withdrawn and desk-rejected submissions because they often lack complete metadata or usable submission files and are not part of the final accept-reject adjudication pool.

ICLR 2017-2024 forms the development split used for structural design, calibration, and fixed in-context examples. ICLR 2025 and 2026 are held out for evaluation; no certificate evidence is fitted on their decisions or reviewer scores.

Before model input, we sanitized the retained submission PDFs to reduce label leakage. Our Python redaction pipeline combined text-pattern matching with region-based PDF masking to remove or obscure page headers, footers, acknowledgment sections, camera-ready and publication markers, and other visible cues that could reveal acceptance status, venue version, or author identity. This preprocessing follows the same idea as the PDF sanitization procedure used in ReviewRL (Zeng et al., 2025), while applying the procedure to the PDFs used in our evaluation pipeline.

Because the test set is used only for evaluation, we manually reviewed all automatically sanitized PDFs. This audit identified cases in which erroneous submission templates, non-standard author annotations, custom comment or markup expressions, or unusual PDF layouts caused the automatic pattern-matching and region masks to fail. The manual review was limited to PDF validity and leakage material.

The development split contains 19,185 papers, including 6,897 accepted papers and 12,288 rejected papers. The 2025 test split was processed using the same decision filtering and PDF sanitization pipeline, yielding 7,281 test papers, including 2,337 accepted papers and 4,944 rejected papers. The 2026 test split contains 13,763 papers, including 4,954 accepted papers and 8,809 rejected papers, corresponding to an acceptance rate of 36%. The final dataset statistics are summarized in Table 7.

B.2 Benchmark Construction

We construct an ICLR corpus from OpenReview submissions spanning ICLR 2017-2026. The earlier years form the development split, while ICLR 2025 and ICLR 2026 are reserved as held-out test sets. Unless otherwise specified, the main comparison in Section 3 reports results on the ICLR 2026 test set. For the numerical comparison, all methods use the same $\mathcal{P}_{\text{score}}$ intersection with that test set. The majority-class comparator on this same cohort is $8{,}221/12{,}846 = 0.640$. The full unfiltered corpus has $8{,}809/13{,}763 = 0.640$ rejected submissions, a separate descriptive denominator. All held-out RuptureTrace rows use the single predeclared six-edge multi-cycle G_θ specified in Appendix C.2; the five-edge graph appears only in the rank-one diagnostic. Thus the outcome, direct-task, and structural results are evaluations of the same score definition with different eligible paper subsets.

Each paper is associated with its PDF, metadata, final decision, and reviewer-provided scores. We parse each paper into five modalities: text, formulation, code, figures, and tables. The text modality is extracted from the paper body. The formulation modality contains displayed equations and mathematical expressions. The code modality contains code blocks, pseudocode, or algorithmic snippets when available. The figure and table modalities contain extracted figures and tables, respectively. Because real submissions often expose only a subset of these modalities, we record a modality-availability mask $a(P) = (a_1, \ldots, a_5) \in \{0, 1\}^5$, where $a_m = 1$ indicates that modality m is observed. Unavailable modalities are masked rather than treated as negative evidence.

B.3 Modality Extraction, Alignment, and Shared Representations

This subsection operationalizes the extraction maps $R_\ell^{(m)}$, local alignment sets $\mathcal{L}_e(P)$, and object collections $\mathcal{Z}_\ell^{(m)}$. The extraction, reference-resolution, and alignment rules are fixed before held-out evaluation and use no conference decisions, reviewer scores or comments, corruption labels, or predictions from the evaluated methods.

Document axis and modality objects. We parse each sanitized PDF into an ordered hierarchy of sections and typed document blocks, retaining the page, character span, object identifier, and bounding box of every block. Bibliography entries, headers, footers, and acknowledgments are excluded. The structural anchors $\ell = 0, \ldots, T$ are nonempty leaf sections in document order; unheaded front matter is attached to the first leaf section. The five modality object types are:

Modality	Included objects	Canonical representation input
Text	Body paragraphs and claim-bearing list items	Original text with section title
Formulation	Displayed equations and mathematical blocks	Parsed LaTeX or OCR math string
Code	Algorithms, pseudocode, and code listings	Ordered raw or OCR-derived code string
Figure	Figure crop, caption, and first reference sentence	Crop plus caption/reference text
Table	Table crop or cells, caption, and first reference sentence	Crop/serialized cells plus caption/reference text

Each extracted object retains one original source location. Let $\mathcal{O}_\ell^{(m)}$ contain the objects whose source location is in leaf section ℓ. The aligned set $x_\ell^{(m)}$ includes these objects and an object sourced from another section only when an explicit reference at anchor ℓ resolves to it. A referenced object is included once in that anchor, even if it has several local references. Its original source pointer is preserved; it is never copied into intervening anchors solely because its source and use lie in different sections. If an associated entity outside the PDF is evaluated, its link, object boundaries, and anchor assignments must be specified by Θ before evaluation. The ICLR results here use the sanitized submission PDF as the entity source. Neither review labels nor nearest-neighbor similarity across sections creates an alignment.

Local alignment and masks. The primary schema is $E^\star = \{e_{12}, e_{23}, e_{34}, e_{45}, e_{51}, e_{13}\}$, with e_{13} measuring text-to-code support. A directed instance $e = (m, n)$ is evaluable at anchor ℓ when both aligned endpoint sets are nonempty and the predeclared relation-specific link $A_{e,\ell}$ is present. For the five original relations this link requires local co-occurrence or a resolved cross-section reference; for e_{13} it additionally requires a resolvable textual pointer to the stated algorithm or code object. The extra edge is not created from unrelated text and code merely occupying a section. Writing this fixed link rule into Θ gives

$$\mathcal{L}_e(P) = \{\ell : x_\ell^{(m)} \neq \varnothing, x_\ell^{(n)} \neq \varnothing, A_{e,\ell} = 1\}, \qquad \Omega(P) = \{e \in E^\star : \mathcal{L}_e(P) \neq \varnothing\}.$$

The link is an evaluability check, not a support or entailment judgment. Globally observed modalities can have no jointly aligned instance, in which case their relation is absent from $\Omega(P)$. An empty endpoint set masks that local instance, and no relation path or union bridge crosses a masked structural gap. We retain original anchor indices after masking rather than compacting the axis. Global availability $a_m(P) = \mathbf{1}[\bigcup_\ell \mathcal{O}_\ell^{(m)} \neq \varnothing]$ is reported for corpus stratification; it does not define $\Omega(P)$.

Backbone-specific rendering and pooling. The default checkpoint is `Qwen/Qwen3.5-2B`, loaded in evaluation mode through its official multimodal processor. Text, formulation, and code use the canonical serializations above. A figure or table is supplied as one early-fusion input containing its crop and textual metadata; these two views are not encoded by separate towers. We extract final-layer states after multimodal fusion and mean-pool all non-padding content tokens, excluding processor control tokens. Text-only decoder ablations use the same final-layer mean pooling on non-padding tokens. Encoder-only ablations use their official pooled or `[CLS]` output, and dual-tower ablations use the official projected embedding. Inputs longer than a checkpoint's context limit are divided into consecutive non-overlapping chunks; no content-dependent truncation is used.

Table 8: Audit of modality extraction and structural alignment.

Audited evidence	Precision	Recall	F1 / Accuracy
Modality block type	0.958	0.949	0.953
Leaf-section assignment	-	-	0.943
Figure/table reference resolution	0.918	0.887	0.902
Local relation eligibility	0.901	0.873	0.887
Provenance-pointer exact match	-	-	0.969

For chunks c of object o, define

$$h_o^{(m)} = \mathrm{norm}\left(\frac{1}{|\mathcal{C}_o|}\sum_{c\in\mathcal{C}_o}\mathrm{norm}(E^{(m)}(c))\right), \qquad \mathcal{Z}_\ell^{(m)} = \{h_o^{(m)} : o \in x_\ell^{(m)}\}.$$

Here $\mathrm{norm}(x) = x/\|x\|_2$. We retain the object set $\mathcal{Z}_\ell^{(m)}$ rather than averaging all objects in a section into one vector. An empty object set is unavailable by construction. If an encoder returns an exact zero vector, that object is marked unavailable; if no object remains, the corresponding anchor-modality pair is masked. A single backbone configuration places all available modality objects in one space $\mathcal{H}$; no decision-supervised projection or relation-specific prediction head is fit. The source pointers attached to each $h_o^{(m)}$ are retained in $q_{e,\ell}(P)$ for every evaluated relation instance.

Extraction and alignment audit. We sample 200 held-out papers stratified by subject area and modality availability. Two annotators independently verify block type, section assignment, figure/table reference resolution, relation eligibility, and provenance pointers without observing scores or rupture predictions. Disagreements are adjudicated before metric computation.

B.4 Baselines and Input Protocols

We compare RuptureTrace with three categories of automated scientific evaluators: (1) general-purpose LLM-as-judge baselines, (2) task-specialized reviewer models/systems, and (3) agentic scientific-evaluation systems. Since these systems support different input interfaces, we adopt modality-specific preprocessing protocols following prior document-understanding and automated-review practice.

Direct document input. For LLM system with native document-level PDF support, we provide the sanitized PDF directly as input (Liang et al., 2023; Zhang & Abernethy, 2025). This protocol follows recent full-manuscript review settings and preserves both textual and visual document evidence when supported by the model interface, which aims to preserve all information aligning with the latest practice [2].

OCR-derived text input. For text-only systems, each PDF is converted into a Markdown-style textual representation using the Datalab API.[3] This protocol follows recent scientific-document parsing and document-AI practice, where PDFs or document images are converted into machine-readable text or markup while retaining equations, tables, and layout-derived structure when available (Zhang & Abernethy, 2025; Idahl & Ahmadi, 2025; Ou et al., 2025; Son et al., 2025). The resulting Markdown includes section content, extracted equations when available, and OCR-derived descriptions of document structure.

Page-image input. For multimodal systems that accept images but not native PDFs, we render each page of the paper as an image and provide the sequence of page images as input. This protocol follows recent visual document understanding and document-RAG practice, where document pages are directly processed as images to preserve visual evidence such as layout, figures, charts, and tables (Ou et al., 2025; Selch et al., 2026; Son et al., 2025). It also avoids relying exclusively on OCR-derived text, which can discard or distort visual and layout information in visually rich documents.

[2] https://developers.openai.com/api/docs/guides/file-inputs

[3] https://www.datalab.to/

Specialized reviewer-oriented and agentic systems. For task-specialized reviewer systems and agentic scientific-evaluation systems, we follow their native input and inference protocols whenever possible (Zhu et al., 2025; Yamada et al., 2025; Weng et al., 2025). When a system requires text-only manuscripts, we provide the same OCR-derived Markdown representation used for text-only LLM-as-judge baselines. This keeps preprocessing comparable while respecting each system's intended deployment interface.

B.5 Baseline Experiment Settings

For LLM-as-judge baselines, we evaluate two prompting settings: zero-shot inference and in-context learning. All baselines receive the full paper content under the input protocol described in Appendix B.4; we do not use paper summaries as substitutes for the target manuscript.

Zero-shot inference. In the zero-shot setting, the model receives the paper content together with an evaluation instruction. The model is asked to predict: (i) a binary accept/reject decision, (ii) an overall assessment score, (iii) a coherence score, and (iv) a novelty score.

In-context learning. In the in-context-learning setting, the model is additionally provided with representative paper-review examples from the development split. Each example contains the paper content, the final decision, and the corresponding reviewer-derived scores. The test paper is then evaluated using the same output format as in the zero-shot setting. The in-context examples are fixed across all test instances for each model to avoid instance-specific prompt tuning.

Model serving. The GPT-5.4 and GLM-4.6V configurations are identified by their provider documentation. Commercial models are evaluated through the OpenAI API[4] and the OpenRouter API[5]. For locally served open-weight models, we load the official checkpoints on a server with 4 NVIDIA A100 and 2 H200 GPUs and use the model's official tokenizer and generation interface. Task-specialized reviewer models/systems and agentic scientific-evaluation systems are evaluated using their native inference protocols whenever possible; when a system requires text-only manuscripts, we provide the same OCR-derived Markdown representation used for text-only LLM-as-judge baselines.

Decoding configuration. For reasoning-capable models, we use the default reasoning configuration when the provider does not expose an explicit reasoning-control parameter. When an explicit reasoning-effort parameter is available, we set it to `medium`. For generation, we set the sampling temperature to 0 whenever supported by the model interface. For models or providers that do not expose temperature control or do not allow temperature 0, we use the provider's default decoding configuration. We set the maximum completion budget to 12K tokens to avoid truncating the structured prediction.

B.6 RuptureTrace Certificate and Score Evaluation

Availability and applicability. The evaluable relation set is the main-text set $\Omega(P) = \{e \in E^\star : \mathcal{L}_e(P) \neq \varnothing\}$. Global modality availability alone does not create an evaluable relation. The score domain is therefore exactly

$$\mathcal{P}_{\text{score}} = \{P \in \mathcal{P} : \Omega(P) \neq \varnothing\}. \tag{5}$$

Every included relation automatically has $|\mathcal{L}_e(P)| > 0$. When $\Omega(P) = \varnothing$, ZigScore and the mask-matched rupture score are unavailable; the input remains in the coverage denominator and receives no imputed numerical score.

Support and structural continuity. For each $e \in \Omega(P)$ and $\ell \in \mathcal{L}_e(P)$, the implementation in Section 2.4 uses the directional object-level coverage map S_e and loss $r_{e,\ell} = 1 - S_e$. The relation path D_e connects only originally consecutive valid anchors. Cycle-valid anchors are $\mathcal{A}_{\text{cyc}} = \{\ell : \beta_1(O_\ell) > 0\}$; union bridges connect adjacent anchors only inside its maximal contiguous valid runs. Homology is computed on the undirected incidence complex of the labeled relation edges, retaining channel identities and directions for relation-level measurement.

[4] `https://platform.openai.com/docs`

[5] `https://openrouter.ai/docs`

Score and audit record. Using the main-text anchor masses, the channel values are

$$Q_e^\epsilon = \sum_{I \in \mathcal{I}_e^\epsilon} m_e(I)^2, \qquad Q_{\mathrm{cyc}}^\epsilon = \frac{1}{\rho_{\max}} \sum_{I \in b_{\mathrm{cyc}}^\epsilon} m_{\mathrm{cyc}}(I)^2.$$

The second expression is evaluated only when $\omega_{\mathrm{cyc}} = 1$; bridge-only intervals receive zero anchor mass. For $P \in \mathcal{P}_{\mathrm{score}}$, the paper-level outputs in Equation (2) are

$$\begin{aligned} Q^\epsilon(P) &= \frac{\omega_{\mathrm{cyc}} Q_{\mathrm{cyc}}^\epsilon(P) + \sum_{e \in \Omega(P)} Q_e^\epsilon(P)}{\omega_{\mathrm{cyc}} + |\Omega(P)|}, \\ \mathrm{ZigScore}(P) &= \int_0^1 Q^\epsilon(P)\, d\epsilon = \sum_{j=0}^{J} (\eta_{j+1} - \eta_j) Q^{\eta_j}(P), \\ \widehat{F}_\Theta^\star(P) &= \int_0^1 Q^\epsilon(P^{\oplus \Omega(P)})\, d\epsilon, \\ \widehat{R}_{\Pi(P);\Theta}(P) &= \widehat{F}_\Theta^\star(P) - \mathrm{ZigScore}(P). \end{aligned} \tag{6}$$

The attainable reference sets $r_{e,\ell} = 0$ only on valid instances and holds all masks, the cycle indicator, and channel normalizations fixed. The critical values $\Pi(P) = \{0, 1\} \cup \{r_{e,\ell} : (e, \ell) \in \mathcal{I}(P)\}$ give the exact integral; the virtual full-repair integrand is likewise constant between these values. The score deficit is $1 - \mathrm{ZigScore}$ only in the special case $\widehat{F}_\Theta^\star = 1$.

For source-linked localization, the specified local score remains $\widehat{\lambda}_{e,\ell} = r_{e,\ell}$ with provenance $q_{e,\ell}(P)$. Each relation retains $c_{e,\Pi} = (b_e, R_e, \Lambda_e, \{\mathcal{I}_e^\epsilon\}_{\epsilon \in \Pi})$. Together with cycle barcodes and applicable repair attributions, these are the components of $\mathfrak{C}_{\Pi,\Theta}^\omega(P)$ in Section 2.4. The relation-level $\widehat{\lambda}$ and the repair attribution ϕ_e answer different localization questions: the former scores an instance; the latter ranks channels by their contribution to a hypothetical restoration of the cycle signal.

Persistent repair attribution. The cycle-based repair game applies on

$$\mathcal{P}_{\mathrm{PRA}} = \{P \in \mathcal{P}_{\mathrm{score}} : \omega_{\mathrm{cyc}}(P) = 1\}. \tag{7}$$

For a coalition $S \subseteq \Omega(P)$, set $r_{e,\ell} = 0$ only for $e \in S$ at its valid aligned anchors. All masks, provenance, and normalization denominators stay fixed. Evaluate $v_P(S) = \int_0^1 Q_{\mathrm{cyc}}^\epsilon(P^{\oplus S})\, d\epsilon$ and compute its Shapley values ϕ_e as defined in the main text. On the primary G_θ graph there are at most $2^6 = 64$ repair coalitions, so the attribution admits exact enumeration. Explicitly, for $n = |\Omega|$,

$$\phi_e(P) = \sum_{S \subseteq \Omega \setminus \{e\}} \frac{|S|!(n - |S| - 1)!}{n!} \big[v_P(S \cup \{e\}) - v_P(S)\big].$$

The Shapley efficiency identity gives $\sum_e \phi_e = v_P(\Omega) - v_P(\varnothing)$. Monotonicity of the graph-only cycle score makes these repair benefits nonnegative. They concern an idealized change in estimated compatibility, not a causal claim about repairing the source experiment. For the relation Top-1/Top-2 evaluation, rank channels in descending ϕ_e on $\mathcal{P}_{\mathrm{PRA}}$, using the tie rule in Appendix B.12. This evaluation ranking is an additional structural attribution; it does not redefine the instance-level localization score $\widehat{\lambda}_{e,\ell} = r_{e,\ell}$ or replace the provenance in $\widehat{\Lambda}_{\Pi;\Theta}$.

B.7 Target Construction

For paper i, let $y_i \in \{\mathrm{accept}, \mathrm{reject}\}$ be the final decision, $\mathcal{R}_i$ its reviews, and o_i its mean overall assessment. For reviewer $r \in \mathcal{R}_i$, let u_{ir}, p_{ir}, and c_{ir} denote soundness, presentation, and contribution. We evaluate three reviewer-derived targets:

$$s_i^{\mathrm{overall}} = o_i, \qquad s_i^{\mathrm{coherence}} = \frac{1}{2|\mathcal{R}_i|} \sum_{r \in \mathcal{R}_i} (u_{ir} + p_{ir}), \qquad s_i^{\mathrm{novelty}} = \frac{1}{|\mathcal{R}_i|} \sum_{r \in \mathcal{R}_i} c_{ir}.$$

The coherence target is a soundness/presentation proxy, and contribution is an imperfect novelty proxy. These targets measure association with reviewer outcomes; none is used to construct the certificate or its rupture traces.

B.8 Decision-Level Metrics

For paper P_i, let $z_i = \mathrm{ZigScore}(P_i) \in [0, 1]$. Conference acceptance is external to the certificate, so binary predictions use one development-set threshold τ_{dec}:

$$\hat{y}_i = \begin{cases} \text{accept}, & z_i \geq \tau_{\mathrm{dec}}, \\ \text{reject}, & z_i < \tau_{\mathrm{dec}}. \end{cases}$$

We select τ_{dec} once and freeze it before held-out evaluation. We report accuracy, Accept F1, and macro-F1 using their standard definitions, with accept as the positive class. This threshold affects only the external decision mapping; it does not change z_i, the conditional certificate, or the rupture attribution.

B.9 Score-Level Metrics

For each target $d \in \{\text{overall}, \text{coherence}, \text{novelty}\}$, we report Spearman correlation between the same native score vector $\mathbf{z} = (z_1, \ldots, z_N)$ and the reviewer-score vector $\mathbf{s}^{(d)} = (s_1^{(d)}, \ldots, s_N^{(d)})$. Thus, the three correlations do not use dimension-specific prediction heads.

For MSE only, the native score must be placed on the corresponding reviewer scale. We fit $g_d(z) = a_d z + b_d$ on the development split, freeze (a_d, b_d), and report

$$\mathrm{MSE}^{(d)} = \frac{1}{N} \sum_{i=1}^{N} \big(g_d(z_i) - s_i^{(d)}\big)^2$$

on the held-out split. This affine map is an evaluation-only scale calibration; it is not used by the certificate, ranking, localization, or persistent repair attribution.

B.10 Efficiency Metrics

We report mean and median wall-clock latency and mean token usage per paper on the ICLR 2026 test split. Baseline token counts include input and output tokens. For RuptureTrace, they count encoder-consumed text after parsing and modality extraction. Its latency includes parsing, extraction, encoding, certificate construction, and aggregation. Because RuptureTrace does not generate from the raw PDF, these measurements describe the reported deployment interfaces rather than architecture-normalized compute.

B.11 Cross-Modal Rupture Localization Benchmark

The primary controlled design samples a 1,200 distinct held-out papers for which all six G_θ channels co-occur at an original anchor in a cycle-valid run. Before model output is examined, allocate 80 papers to each of the 5×3 cells defined by the five adjudicated target relations and the three corruption families. Each paper contributes one clean–corrupted pair for the assigned family; edit realizations are repeated with five fixed seeds, and all uncertainty resamples the paper rather than individual edits or anchors. For each selected paper, we target an observed relation $e_{mn} \in \Omega(P)$ at a valid aligned anchor $\ell \in \mathcal{L}_{mn}(P)$. The cross-paper family replaces the aligned segment of M_n with a length-, format-, and topic-matched segment from another held-out paper. The within-paper family instead reassigns an appropriate segment from a different section of the same paper; the hard-corruption family edits the targeted source object in place. Each intervention preserves all other modality content, section structure, and availability masks. An edit that breaks the channel's reference/link gate is excluded rather than converted into a masked relation; masks and the six candidates must remain identical within the clean–corrupted pair. The targets are text-formulation (e_{12}), formulation-code (e_{23}), code-figure (e_{34}), figure-table (e_{45}), and table-text (e_{51}). The replaced units are respectively the formulation, implementation or pseudocode, figure, table, and supported textual claim.

The targeted relation and modified anchors provide the primary localization labels. Replacing one modality can also affect its other incident relations, including auxiliary e_{13}, so the primary target is not assumed to be the only affected edge. All six channels are localization candidates; Top-2 recall captures whether the target is retained among the two highest-ranked relations.

Hard-corruption strength is the number $s \in \{1, 2, 3\}$ of nested relation-specific atomic modifications. For a fixed paper and target relation, the level-s corruption modifies a contiguous block of s eligible anchors, with $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \mathcal{B}_3$.

B.12 Rupture Localization Metrics

Evaluation cohort. Numerical score comparisons require defined scores for both members of each clean-corrupted pair. A comparison including a cycle-only method additionally requires an available cycle channel for both. For relation and anchor localization, use the same corrupted material for every relation-resolved method, with $P_j^{\mathrm{corr}} \in \mathcal{P}_{\mathrm{PRA}}$ and $e_j^\star \in \Omega(P_j^{\mathrm{corr}})$. The candidate pool and the score-, cycle-, and localization-eligible cohorts have different denominators.

Eligibility is fixed from availability, alignment, and the benchmark labels before comparing predictions. Empty predictions remain in the eligible cohort.

Rupture AUROC. Use the main-text mask-matched rupture score $\widehat{R}_{\Pi(P);\Theta}(P) = \widehat{F}^\star_\Theta(P) - \mathrm{ZigScore}(P)$, with clean papers labeled negative and corrupted papers labeled positive. For a clean-corrupted pair with identical masks, its attainable reference is unchanged; the paired score decrease below equals the increase in rupture score. Across papers with different masks, $1 - \mathrm{ZigScore}$ need not give the same AUROC ranking.

Relation ranking and ties. The full model ranks $\phi_e(P)$ in descending order for Top-1/Top-2. This channel ordering is distinct from the Problem 1 instance score $\widehat{\lambda}_{e,\ell} = r_{e,\ell}$. The ranking scores for w/o-PRA and the edge-score baseline without persistence are, respectively,

$$1 - \int_0^1 Q_e^\epsilon(P)\, d\epsilon, \qquad \frac{1}{N_e} \sum_{\ell \in \mathcal{L}_e} r_{e,\ell}.$$

Round each ranking score to 12 decimal places using round-to-even before sorting. Equal scores follow the fixed six-channel order $e_{12} \prec e_{13} \prec e_{23} \prec e_{34} \prec e_{45} \prec e_{51}$. The same ordering gives Top-1 and the first two distinct relations for Top-2. This precision convention handles numerical ties and is not fitted to localization labels. Writing the ordered relations as $(\widehat{e}_j^{(1)}, \widehat{e}_j^{(2)}, \ldots)$,

$$\mathrm{Top1} = \frac{1}{N_{\mathrm{loc}}} \sum_j \mathbf{1}[\widehat{e}_j^{(1)} = e_j^\star],$$

$$\mathrm{Top2} = \frac{1}{N_{\mathrm{loc}}} \sum_j \mathbf{1}[e_j^\star \in \{\widehat{e}_j^{(1)}, \widehat{e}_j^{(2)}\}].$$

PRA is not replaced by a no-cycle fallback.

Score decrease. For the common eligible clean-corrupted pairs,

$$\Delta_{\mathrm{score}} = \frac{1}{N_{\mathrm{pair}}} \sum_j \big[\mathrm{ZigScore}(P_j) - \mathrm{ZigScore}(P_j^{\mathrm{corr}})\big].$$

Paired uncertainty. For the headline controlled table, preserve the $1{,}200$ base-paper IDs across all methods and compare each score and six-candidate ranking on the same intervention realization. Repeat mutation realizations with five fixed seeds, first compute one statistic per seed, then resample independent papers with the paired method outputs kept together. The paired paper-bootstrap 95% intervals are $[0.875, 0.901]$ for the primary-score 0.889 AUROC, $[0.668, 0.718]$ for its 0.694 PRA Top-1, and $[0.102, 0.161]$ for the 0.131 paired PRA-minus-direct-edge Top-1 difference.

C Additional Results

C.1 Exact-Scan Verification

Under fixed representations and masks, restricting the primary six-edge graph to its original five-edge cycle gives a diagnostic whose rank-one scan matches the general zigzag computation to numerical

Table 9: Supplementary results on the held-out ICLR 2025 corpus. These results assess a second conference year; checkpoint dates and documented data exposure are reported separately for contamination auditing. Superscripts denote the input modality: † direct document file input and ‡ OCR-derived text from paper documents.

Method	Acc ↑	Acc. F1 ↑	Macro-F1 ↑	OAρ ↑	OA MSE ↓	Coh.ρ ↑	Coh. MSE ↓	Nov.ρ ↑	Nov. MSE ↓
Zero-shot									
GPT-4o†	0.315	0.479	0.239	0.116	8.976	0.027	1.038	0.086	1.882
GPT-5†	0.584	0.533	0.579	0.506	1.943	0.348	0.358	0.325	0.625
GPT-5 w/ reasoning†	0.573	0.515	0.567	0.485	1.957	0.415	0.350	0.383	0.602
GPT o3†	0.575	0.508	0.567	0.426	1.627	0.367	0.304	0.294	0.441
GPT o3, w/ reasoning†	0.577	0.513	0.570	0.360	1.747	0.325	0.311	0.238	0.470
Llama-3-70B-Instruct‡	0.309	0.472	0.236	-0.047	6.739	0.108	0.918	0.047	1.622
Llama-3.3-70B-Instruct‡	0.309	0.472	0.236	0.152	10.607	0.163	1.164	-0.051	2.366
Qwen-2.5-72B-Instruct‡	0.314	0.470	0.248	0.055	4.278	0.088	0.907	0.020	1.566
Specialized paper evaluation methods									
AI-Scientist v1‡	0.691	0.081	0.448	-0.042	4.784	0.049	0.364	0.018	0.554
RUPTURETRACE	0.781$\pm$0.015	0.643$\pm$0.018	0.742$\pm$0.015	0.558$\pm$0.020	1.110$\pm$0.041	0.551$\pm$0.021	0.151$\pm$0.007	0.490$\pm$0.023	0.199$\pm$0.012

Table 10: Supplementary efficiency comparison on locally runnable models. Frontier API baselines expose latency and token usage but not directly comparable local GPU time or FLOPs; this table therefore reports the controlled local comparison on runnable Qwen3.5 models under the zero-shot OCR setting. Units are seconds per paper for GPU Time, TFLOPs per paper for FLOPs, and generated tokens per paper for Gen. Tokens. Superscript ‡ denotes OCR-derived text from paper documents.

Method	GPU Time ↓	TFLOPs ↓	Gen. Tokens ↓
Qwen3.5-2B‡	17.39	34.24	48.36
Qwen3.5-4B‡	80.63	91.35	284.32
Qwen3.5-9B‡	81.27	176.03	284.24
RUPTURETRACE	**10.24**	**19.31**	**0.00**

precision (mean absolute difference 2.4×10^{-13}; Spearman 1.000) while running $11.7\times$ faster. This is an implementation verification of Proposition 2, not evidence that a full decomposition is necessary on that five-edge subgraph. The overlapping-cycle experiments below use the *primary six-edge graph* used for Tables 1 and 5.

C.2 Detailed Structural and Specificity Results

This section expands the structural comparison in Table 22. The graph used by all primary paper-level and direct-task results is

$$G_\theta = (\{M_1, \ldots, M_5\}, \{e_{12}, e_{23}, e_{34}, e_{45}, e_{51}, e_{13}\}),$$

where the chord e_{13} is text-to-code support and the five-cycle retains all five reviewer-labeled categories. The triangle through formulation and the cycle through figures and tables span two independent classes; the original five-cycle is their sum over $\mathbb{F}_2$. Edges retain the semantic direction of their endpoint relation, while homology is computed on the underlying undirected graph. Its cycle rank is two, so all variants use the fixed normalization $\rho_{\max} = 2$.

For Panel A of Table 22, retain a 1,842-paper diagnostic subset of the primary-template eligible cohort in which the two fixed triangle/four-edge basis cycles span every nonzero anchor space at every observed critical scale; this filter depends only on the support filtration and is applied before outcome scoring. Panel B instead uses all 3,516 template-eligible papers, with its sensitive column defined by the observed switch subset ($n = 684$). Different panels have different OA denominators, so their full-zigzag correlations need not agree. Neither panel substitutes a new relation schema for the primary paper score.

Rank-run baseline. For each threshold ϵ, let

$$\beta_\ell^\epsilon = \dim H_1(K_\ell(\epsilon)), \qquad A_h^\epsilon = \{\ell \in \mathcal{A}_{\text{cyc}} : \beta_\ell^\epsilon \geq h\}.$$

Let $\mathcal{R}_h^\epsilon$ denote the maximal contiguous runs of A_h^ϵ within the valid structural runs. We define

$$Q^\epsilon_{\mathrm{rank}}(P) = \frac{1}{\rho_{\max}} \sum_{h=1}^{\rho_{\max}} \sum_{R \in \mathcal{R}_h^\epsilon} \left(\frac{|R|}{N_{\mathrm{cyc}}} \right)^2, \qquad Q_{\mathrm{rank}}(P) = \int_0^1 Q^\epsilon_{\mathrm{rank}}(P)\, d\epsilon.$$

This baseline depends only on the ordered anchor-wise Betti ranks and does not retain cycle identity or inclusion-induced maps. It agrees with the cycle scalar in the fixed rank-one regime but can differ on G_θ.

C.2.1 Classwise scan baselines

We define a rank-matched non-module comparator explicitly. For the fixed ambient graph G_θ, identify every anchor and bridge cycle space with its image in $H_1(G_\theta; \mathbb{F}_2)$:

$$W_\ell^\epsilon = \mathrm{im}[H_1(K_\ell(\epsilon)) \to H_1(G_\theta)], \qquad U_\ell^\epsilon = \mathrm{im}[H_1(B_\ell(\epsilon)) \to H_1(G_\theta)].$$

Let $\beta_\ell^\epsilon = \dim W_\ell^\epsilon$ and $\mathcal{C}_\theta = H_1(G_\theta; \mathbb{F}_2) \setminus \{0\}$. For each $\gamma \in \mathcal{C}_\theta$, construct the binary class sequence recording whether γ belongs to each W_ℓ^ϵ and U_ℓ^ϵ, and let $\mathfrak{R}_\gamma^\epsilon$ denote its maximal anchor-supported runs. For $R \in \mathfrak{R}_\gamma^\epsilon$, define

$$m_\gamma(R) = \frac{1}{N_{\mathrm{cyc}}} \sum_{\ell \in R} \sqrt{\frac{\beta_\ell^\epsilon}{\rho_{\max}(2^{\beta_\ell^\epsilon} - 1)}}, \qquad Q^\epsilon_{\mathrm{cls}} = \sum_{\gamma \in \mathcal{C}_\theta} \sum_{R \in \mathfrak{R}_\gamma^\epsilon} m_\gamma(R)^2,$$

with zero weight when $\beta_\ell^\epsilon = 0$. If the same rank-r subspace persists across a complete run, then

$$Q^\epsilon_{\mathrm{cls}} = (2^r - 1) \frac{r}{\rho_{\max}(2^r - 1)} = \frac{r}{\rho_{\max}},$$

which equals the full-zigzag contribution. Thus differences cannot be attributed to static rank scaling. The exhaustive classwise scan can differ only when the subspaces change, because it treats the nonzero classes independently rather than respecting their joint linear dependence in a vector-space module.

The fixed-basis scan applies the same binary-run construction only to a predetermined basis $\{\gamma_1, \gamma_2\}$, assigning each present basis class weight $1/\sqrt{\rho_{\max}}$. It is therefore basis-dependent and can miss the non-basis class $\gamma_1 + \gamma_2$. All classwise scores are integrated over the same critical filtration intervals and combined with the same relation-level terms as full zigzag.

C.2.2 Overlapping-class matched pairs

The exact G_θ construction uses three consecutive anchors with cycle-space dimensions $(1, 2, 1)$ and rank-two union bridges. Let $V = H_1(G_\theta; \mathbb{F}_2)$. The coherent member is

$$\langle \gamma \rangle \to V \leftarrow V \to V \leftarrow \langle \gamma \rangle,$$

whereas the switching member replaces the last term by $\langle \gamma' \rangle$ for $\gamma' \neq \gamma$. The two members have the same ordered dimensions, supported-edge counts, and rupture-value multiset; the switching member is obtained by permuting those rupture values among edges so that only the endpoint embedding changes. For an explicit realization at one threshold, take the middle anchor to support all six edges and the first anchor to support $G_\theta \setminus \{e_{45}\}$, whose sole cycle is the triangle. Give the coherent member the same five-edge final anchor; give the switching member $G_\theta \setminus \{e_{23}\}$ as its final anchor, whose sole cycle is the four-edge cycle. Both final anchors have five supported edges and one cycle, and the middle bridge has rank two. Their anchor-projected full-zigzag contributions are, respectively,

$$\frac{1}{2}\left[\left(\frac{3}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = \frac{5}{9}, \qquad \frac{1}{2}\left[\left(\frac{2}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = \frac{4}{9}.$$

Thus the comparison holds local mass and rank fixed while changing whether one latent cycle summand survives the expansion-contraction window.

For perturbed observed pairs, candidate windows must have the same $(1, 2, 1)$ rank pattern and endpoint relation coverage. We pair coherent and switching windows exactly on supported-edge

counts and within a caliper of 0.02 for normalized integrated support mass and 0.02 for each relation-level coherence term. Matching uses no reviewer score, decision, corruption label, or method output. Pair accuracy is the fraction for which a global structural summary assigns the coherent member the larger cycle score; relation-level terms are excluded from this metric and numerical ties count as one half. We repeat matching with five prespecified seeds and report the mean and standard deviation. The same pair IDs and tie rule are used for every structural summary.

C.2.3 Sensitivity across overlapping templates

We prespecify one primary six-edge template G_θ on the design split and apply it to *all* held-out RuptureTrace rows. Two single-chord alternatives test graph-choice sensitivity: the formulation–figure template replaces e_{13} with e_{24}, and the code–table template replaces it with e_{35}; the original five edges remain present in both. Each connected six-edge template has cycle rank two and the same $\rho_{\max} = 2$. For e_{24}, a formulation is evaluated against a figure only when the figure or its local discussion explicitly refers to the specified equation or predicted quantity; e_{35} analogously requires a table's methods or local discussion to reference the specified code or algorithm. These alternate link gates are frozen along with the primary e_{13} gate. A paper is template-eligible when all six channels required by that particular template co-occur at an original anchor in a contiguous valid run; these alternate schemas are only sensitivity experiments, not pooled with the primary score. A natural cycle-class switch is counted when inclusion-induced cycle spaces contain a three-anchor $(1, 2, 1)$ pattern with different endpoint one-dimensional classes. Templates, eligibility, and normalizations are fixed before held-out outcomes are evaluated.

Table 11: Held-out prevalence and sensitivity across semantically specified overlapping-cycle templates. Percentages use the full held-out corpus ($N = 13{,}763$); "switch/eligible" reports the fraction of template-eligible papers containing a natural cycle-class switch. $\Delta\rho$ is full zigzag minus the rank-matched exhaustive classwise scan.

Template	Eligible (corpus)	Switch (corpus)	Switch/eligible	All $\Delta\rho$	Switch $\Delta\rho$
Primary text–code chord	3,516 (25.5%)	684 (5.0%)	19.5%	0.017	0.049
Formulation–figure chord	3,278 (23.8%)	603 (4.4%)	18.4%	0.014	0.043
Code–table chord	3,384 (24.6%)	627 (4.6%)	18.5%	0.015	0.046

Across all three templates, roughly one quarter of held-out papers expose the required relation topology, and 4.4-5.0% of the full corpus exhibits a natural cycle-class switch. Full zigzag has a positive $\Delta\rho$ for every template, with the larger gains concentrated on the observed switch subsets. These frequencies show that the overlapping regime occurs in a measurable subset. The all-eligible $\Delta\rho$ is conditional on template eligibility, and the switch-subset gain is conditional on an observed switch; neither is a whole-corpus treatment effect. The tables do not identify the corresponding incremental gain among papers without the specified switch.

Switch and non-switch strata. To separate the prevalence of G_θ from the incremental benefit of cross-anchor correspondence, we split its 3,516 eligible papers into 684 with an observed cycle-class switch and 2,832 without one. Table 12 reports $\Delta\rho = \rho_{\text{zigzag}} - \rho_{\text{classwise}}$ under the same matched relation measurements and eligibility rule. The 10,247 papers without an eligible G_θ run are outside this template-specific comparison, so we do not assign them an artificial zero gain. Spearman correlations computed within different strata are not additive; the all-eligible difference cannot be reconstructed by weighting the two subgroup differences.

Relative to the exhaustive classwise baseline, the full-zigzag gains are 0.017 in all-eligible OA correlation, 0.049 in subspace-switch OA correlation, 14.5 matched-pair accuracy points. The paired 95% intervals are respectively $[0.009, 0.026]$, $[0.031, 0.067]$, and $[11.2, 17.8]$.

C.2.4 Matched exact window-intersection computation

To compare two *exact* evaluations of the same structural statistic, run both backends on the same 3,516 G_θ-eligible papers. Freeze each paper's extracted objects, loss field, masks, support-scale mesh, graph template, and all at most $2^6 = 64$ virtual-repair coalitions; force the general computation for both backends even on a paper where a simpler scan is available. The measurement begins

Table 12: Conditional OA associations by naturally observed G_θ topology. Differences compare full zigzag with the exact rank-matched exhaustive classwise scan on the *same paper IDs* within each stratum. Plus/minus entries are marginal paper-bootstrap s.d.

G_θ stratum	Papers	Classwise OA ρ	Zigzag OA ρ	Paired $\Delta\rho$ [95% CI]
Eligible with switch	684	0.452 ± 0.017	0.501 ± 0.016	$0.049\ [0.031, 0.067]$
Eligible without switch	2,832	0.546 ± 0.009	0.555 ± 0.009	$0.009\ [-0.002, 0.020]$
All template-eligible	3,516	0.531 ± 0.008	0.548 ± 0.008	$0.017\ [0.009, 0.026]$
Template-ineligible	10,247	–	–	Not defined

Table 13: Full structural ablation under the primary overlapping-cycle $G_\theta = \{e_{12}, e_{23}, e_{34}, e_{45}, e_{51}, e_{13}\}$. The fixed-basis scan tracks two predetermined classes. The exhaustive classwise scan tracks every nonzero class with rank-matched normalization but aggregates the classes independently.

Topological summary	All eligible OA ρ	Subspace-switch OA ρ	Matched Pair Acc. (%)
Mean support	0.418 ± 0.011	0.292 ± 0.020	50.4 ± 0.9
Pointwise topology	0.449 ± 0.010	0.327 ± 0.019	50.8 ± 1.1
Rank-run coherence	0.503 ± 0.009	0.386 ± 0.018	51.7 ± 1.2
Fixed-basis class scan	0.526 ± 0.009	0.438 ± 0.018	68.9 ± 1.5
Exhaustive classwise scan	0.531 ± 0.008	0.452 ± 0.017	74.6 ± 1.7
Full zigzag	$\mathbf{0.548 \pm 0.008}$	$\mathbf{0.501 \pm 0.016}$	$\mathbf{89.1 \pm 1.5}$

after representation extraction and ends after the integrated paper score and repair attributions are computed. Alternate backend order over five warm passes per paper. Report the median per-paper elapsed time and peak incremental traced allocations above the frozen-input state; brackets describe the across-paper interquartile range of the per-paper medians. This isolates computation from parsing, encoding, and filesystem cache effects.

The second backend intersects the thresholded labeled edge sets incrementally for each left window endpoint and applies Proposition 2. Under matched repairs it also produces identical scalar-based issue predictions and Shapley values up to numerical tolerance. These projections suggest a runtime–memory tradeoff; they provide no detection advantage for barcode decomposition. A barcode can still be retained in the audit record when individual interval locations are inspected.

C.2.5 Coverage, mask placement, and alignment controls

These controls specify the comparisons summarized in Section C.4. For the document-level coverage control, preselect a score-eligible held-out subset by stratifying on global modality count, document-length quintile, and observed relation-coverage quintile, without using outcomes or model scores. On the same selected paper IDs, compare native ZigScore with the coverage-only statistic

$$C_{\mathrm{cov}}(P) = \frac{1}{|E^\star|(T+1)} \sum_{e \in E^\star} |\mathcal{L}_e(P)|,$$

which measures how often an expected relation is observable but uses no directional support loss. A separate availability normalization uses $\mathrm{ZigScore}(P)/\widehat{F}^\star_\Theta(P)$ on this same subset. The denominator is positive whenever $P \in \mathcal{P}_{\mathrm{score}}$; the comparison is an empirical sensitivity check and does not assert invariance to masks.

For mask-position sensitivity, take a paper with a masked gap and an adjacent valid position in at least one relation path. Move that gap by one position within the path while retaining its ordered measured losses, leaving other paths fixed, and recompute the score with the new mask. Only eligible papers enter the paired rank comparison. For the alignment-shuffle control, select papers with at least two valid anchors in the same target-object cardinality bin. Derange the target-side object sets across these anchors within the paper, keeping source sets, relation labels, cardinality bins, and local masks fixed. This breaks the specified local pairing while preserving the eligible denominator. Compute the

Table 14: Paired, backend-only computation on the same G_θ papers and repair coalitions. Entries are medians [interquartile ranges], not uncertainties on the difference. The algebraic identity guarantees the same exact scalar.

Exact backend	Seconds / paper	Peak incremental MiB	Max. $\lvert\Delta\,\mathrm{ZigScore}\rvert$
Full zigzag barcode	$0.124\ [0.087, 0.196]$	$5.6\ [4.2, 7.8]$	reference
Window intersections	$0.177\ [0.109, 0.294]$	$3.8\ [3.1, 5.4]$	$< 10^{-10}$

Table 15: Coverage and perturbation controls. Different paired interventions have their own prespecified eligibility and are not compared across rows. The rank agreement is Spearman correlation with the unperturbed native ZigScore on the corresponding cohort.

Control	Cohort	Outcome
Coverage-only C_{cov}	4,000 papers	OA $\rho = 0.219$
Native ZigScore on same papers	4,000 papers	OA $\rho = 0.539$
ZigScore / attainable reference	4,000 papers	OA $\rho = 0.537$; rank $\rho = 0.992$
One-step mask relocation	2,400 pairs	rank $\rho = 0.988$
Within-paper alignment shuffle	1,800 pairs	score drop 0.173; AUROC 0.907

paired ZigScore decrease and clean-versus-shuffled AUROC using the mask-matched $\widehat{R}_{\Pi;\Theta}$. The benign-versus-hard AUROC is 0.874 ± 0.009, with a paired difference in score change of 0.128.

C.3 Measurement Validation and End-to-End Robustness

We evaluate the measurement layer independently of the structural aggregation through four targeted tests: preservation of single-object omissions, within-paper localization under fixed document identity, support-scale concentration and recalibration, and end-to-end sensitivity to parser/alignment perturbations.

C.3.1 Support-map instantiation and single-object omissions

Let the original section-level baseline replace each anchor-modality object set by its mean normalized vector before cosine scoring. The reported RuptureTrace instantiation instead computes source-object best matches and averages their coverage. We compare both maps with identical anchors, encoders, masks, structural aggregation, and PRA. We additionally construct a single-object omission benchmark: for each eligible anchor with $n \in \{2, 4, 8, 16\}$ source objects, one reviewer- or corruption-identified evidential object is removed or replaced while every other object is held fixed. This directly measures the count-weighting effect identified by the motivating formulation-to-code example.

The primary comparison is paired at the anchor level. The score-drop series exposes the cardinality effect directly: from $n = 2$ to $n = 16$, the section-mean signal falls from 0.112 to 0.024 (21.4% retained), while object max-coverage falls from 0.181 to 0.068 (37.6% retained). The accompanying AUROC and anchor-F1 results test whether that stronger local signal improves rupture detection and localization.

C.3.2 Repository-backed operation and matching stress tests

For each stress family, select a 120 distinct papers with a valid, aligned formulation–code relation and an accessible fixed repository revision before scoring. Independently confirm the equation, the aligned executable operation, and whether a mutation breaks that operation. For each paper, generate one relation-breaking variant and one semantics-preserving code edit while holding its other material, anchors, and availability masks fixed. Repeat edit realizations with five frozen random seeds; sample papers, not edits or anchors, for paired bootstrap uncertainty. Different families may use different eligible papers, so the table does not support cross-family significance claims.

The omitted-operation family removes an executable regularizer calculation or optimizer contribution while retaining its stated formulation. The wrong-direction family reverses a supported sign or update direction with symbol names and layout preserved. The many-to-one family replaces distinct

Table 16: Specificity to relation-breaking rather than benign edits. Benign edits preserve cross-modal meaning consistently across affected modalities.

Edit type	$\lvert\Delta_{\text{score}}\rvert \downarrow$	Flagged as rupture (%) $\downarrow$
Text paraphrase	0.016 ± 0.004	5.8 ± 0.9
Consistent symbol renaming	0.019 ± 0.005	6.4 ± 1.0
Figure restyling	0.022 ± 0.006	7.1 ± 1.1
Table reformatting	0.015 ± 0.004	5.2 ± 0.8
Pooled benign edits	0.018 ± 0.004	6.1 ± 0.8
Matched hard corruptions	0.146 ± 0.011	78.4 ± 1.7

Table 17: Direct support-map ablation. "Single omission" contains exactly one modified evidential object per anchor. All rows share the same encoder and structural aggregation.

Support map	Natural AUROC	Controlled AUROC	Single-omission AUROC	Anchor F1
Section mean-pool cosine	0.751 ± 0.012	0.824	0.742	0.491
Object max-coverage (RUPTURETRACE)	0.803 ± 0.014	0.889	0.862	0.618

evidential targets with a single broadly worded description that can attract several source-object best matches; independent inspection confirms that the required operations are absent. All labels concern the constructed difference at the fixed code revision. A prompted paper–code QA comparator, *motivated by* the question format of SciCoQA (Baumgärtner & Gurevych, 2026), receives the same source and executable-code spans with the same frozen backbone; it is not a reproduction of the SciCoQA benchmark or its published model scores. Its issue score is the probability of a stated paper–code discrepancy.

A controlled alteration of real code does not establish that the illustrative Figure 1 corresponds to a naturally occurring, independently verified code omission.

C.3.3 WITHIN-PAPER CROSS-SECTION SWAP CONTROL

For each eligible relation, we construct a within-paper cross-section swap that keeps the manuscript identity fixed while changing the local evidential pairing. Two same-modality segments from different leaf sections of the same manuscript are exchanged only when their local target relations differ under the manual relation annotation. The operation preserves paper identity, document-level topic, modality counts, segment length bin, formatting class, and availability masks while breaking the original local evidential alignment. Each swap is paired with the untouched manuscript and evaluated with the same relation and anchor labels.

C.3.4 REVIEWER-OUTCOME COVARIATE CONTROL

We quantify the contribution of simple document structure to the correlations in Table 1. On the design split, we fit rank-linear nuisance models for both the RUPTURETRACE score and each reviewer outcome using log document length, leaf-section count, figure count, table count, number of observed modalities, and relation-coverage fraction. We freeze these nuisance models and compute Spearman correlation between held-out residuals. We also report a richer sensitivity model that adds reference count and appendix length when available.

C.3.5 SUPPORT-SCALE CONCENTRATION AND MONOTONE RECALIBRATION

We inspect the empirical rupture scale induced by each support map using the median rupture, the mass below 0.10, and the central critical-value span $W_{90} = q_{0.95}(r) - q_{0.05}(r)$. The object-level max operation shifts more rupture values toward zero: its median falls from 0.181 to 0.112, $\Pr(r < 0.10)$ rises from 0.182 to 0.432, and W_{90} contracts from 0.421 to 0.287. These statistics directly quantify the support-scale compression introduced by best-match coverage.

We then fit a relation-specific empirical-CDF transform on the design split and apply it unchanged to held-out instance losses $r_{e,\ell}$ before thresholding and structural aggregation. This is not a monotone

Table 18: Sensitivity to one missing evidential object as the number of source objects in an anchor increases. Report the paired clean-corrupted score drop on the same anchors.

Support map	$n = 2$	$n = 4$	$n = 8$	$n = 16$
Section mean-pool cosine	0.112	0.073	0.041	0.024
Object max-coverage (RUPTURETRACE)	0.181	0.142	0.101	0.068

Table 19: Held-out reviewer association before and after residualizing document-structure covariates. Nuisance models are fit only on the design split.

Association	OA ρ	Coh. ρ	Nov. ρ
Raw RUPTURETRACE score	0.552	0.548	0.482
Residualized: core structural covariates	0.497	0.489	0.352
Residualized: core + references/appendix length	0.486	0.477	0.331

transform of the final ZigScore: it preserves within-relation order but can change cross-relation scales, cycle thresholds, document rankings, and AUROC. On held-out papers, the recalibrated certificate has score-rank correlation 0.982 with the raw certificate, reported-issue AUROC 0.798 versus 0.803, and relation Top-1 0.561 versus 0.568. The discrimination and localization results therefore remain stable after the rupture scale is expanded from $W_{90} = 0.287$ to 0.900.

C.3.6 PARSER AND ALIGNMENT PERTURBATION

Starting from the manually audited subset in Table 8, we inject controlled upstream errors before encoding: block-type flips, section-assignment shifts to an adjacent leaf section, and broken figure/table reference links. At each perturbation rate $p \in \{2.5, 5, 10, 15\}\%$, the same sampled objects are perturbed across all compared methods and five fixed seeds. We report the absolute score change, rank correlation with the unperturbed certificate, reported-issue AUROC, and relation Top-1.

At a 10% perturbation rate, close to the audited extraction/alignment error scale, score-rank agreement remains 0.956, reported-issue AUROC is 0.773, and relation Top-1 is 0.526. The full curve quantifies the end-to-end degradation as upstream error increases.

C.3.7 CYCLE-CHANNEL WEIGHT SENSITIVITY

The default score gives each observed relation channel and the cycle channel equal channel-level weight, yielding cycle weight $1/7$ for the complete primary six-relation graph. We evaluate a generalized aggregation on the held-out score-eligible cohort, changing only papers with an observable cycle channel,

$$Q^{\epsilon}_{\lambda}(P) = \lambda_{\mathrm{cyc}} Q^{\epsilon}_{\mathrm{cyc}} + (1 - \lambda_{\mathrm{cyc}}) \frac{1}{|\Omega(P)|} \sum_{e \in \Omega(P)} Q^{\epsilon}_{e},$$

while papers without an observable cycle retain their relation-only aggregation. The exact default has paper-specific weight $\lambda_{\mathrm{cyc}} = 1/(|\Omega| + 1)$, equal to $1/7$ on a complete six-channel paper. A fixed-weight sweep row approximates this default when channel coverage varies.

Held-out sweep.

λ_{cyc}	Score-rank ρ	OA ρ	Issue AUROC
0	0.974	0.539	0.790
$1/12$	0.994	0.548	0.799
$1/7$ (fixed)	0.999	0.552	0.803
Paper-specific (default)	1.000	0.552	0.803
$1/4$	0.996	0.551	0.802
$1/3$	0.988	0.548	0.798
$1/2$	0.969	0.540	0.789

The score exhibits a broad plateau around the paper-specific equal-channel default: fixed weights from $1/12$ through $1/3$ keep OA correlation within 0.548–0.552 and reported-issue AUROC within

Table 20: Support-scale concentration and monotone-recalibration sensitivity on the held-out set. $W_{90} = q_{0.95}(r) - q_{0.05}(r)$ measures the span containing the central 90% of rupture values. Score-rank ρ is measured against the raw object-level certificate.

Variant	Median r	$\Pr(r < 0.10)$	W_{90}	Score-rank ρ	Issue AUROC	Rel. Top-1
Section mean-pool, raw	0.181	0.182	0.421	0.912	0.751 ± 0.012	0.503 ± 0.038
Object max-coverage, raw	0.112	0.432	0.287	1.000	0.803 ± 0.014	0.568 ± 0.038
Object max-coverage, rank-calibrated	0.500	0.100	0.900	0.982	0.798 ± 0.011	0.561 ± 0.037

Table 21: End-to-end sensitivity to controlled extraction/alignment noise. Values are averaged over block-type, section-assignment, and reference-link perturbations; per-error-type results follow in the supplement.

Noise rate	$\lvert\Delta_{\text{score}}\rvert$	Score rank ρ	Issue AUROC	Relation Top-1
0%	0	1.000	0.803	0.568
2.5%	0.006 ± 0.002	0.993	0.798	0.561
5%	0.011 ± 0.003	0.984	0.790	0.548
10%	0.023 ± 0.005	0.956	0.773	0.526
15%	0.037 ± 0.007	0.921	0.751	0.497

0.798–0.803. The $1/7$ row is a fixed-weight approximation; the exact primary score uses $1/(|\Omega| + 1)$ and has rank agreement 1.000 with itself. On the same score-eligible cohort, removing the cycle channel changes OA ρ from 0.552 to 0.539; its paired paper-bootstrap interval is still required. PRA attribution is computed before scalar channel weighting and is unchanged by this sweep.

Controlled rupture validity. We use three matched intervention families with progressively tighter semantic control. Cross-paper replacement swaps one aligned modality segment with a topic-matched donor; within-paper cross-section swaps preserve paper identity, modality, length, and formatting while breaking the local relation; relation-specific hard corruptions alter the implicated symbol/assumption, implementation operation, numerical trend, or conclusion. Semantics-preserving paraphrases, symbol renaming, and visual reformatting provide matched non-rupture edits.

C.4 Topological Regimes and Structural Contributions

We test the computation boundary of Proposition 2 while holding representations, rupture values, masks, relation terms, and repair procedures fixed. In the unique-cycle regime, the rank-one scan and full zigzag agree to numerical precision (mean absolute difference 2.4×10^{-13}; Spearman 1.000). In the separable-cycle regime, an identity-aware interval scan is likewise equivalent to full zigzag. In the overlapping-cycle regime, we compare the reported full-zigzag statistic with fixed-basis and exhaustive classwise scans over the same local cycle classes. The exact window-intersection formula yields the same scalar and therefore the same scalar-based predictions. On the same G_θ-eligible papers, median structural-runtime costs are 0.124 versus 0.177 seconds (full barcode versus exact windows), with median peak incremental allocations of 5.6 versus 3.8 MiB; the common-mesh, common-repair benchmark and spread are specified in Appendix C.2.4. A held-out corpus audit finds the primary G_θ template observable in 3,516/13,763 papers (25.5%); 684 papers contain an observed cycle-class switch (5.0% of the corpus; 19.5% of eligible papers), while 2,832 are eligible without a switch. Thus the largest correspondence-specific gain applies to a small, identifiable part of the full corpus. Two alternate θ templates yield switch prevalence of 4.4–4.6% of the corpus. The separately no-switch gain in OA correlation over the exhaustive classwise scan is 0.009 (paired bootstrap 95% interval $[-0.002, 0.020]$) among 2,832 eligible papers, compared with 0.049 $[0.031, 0.067]$ among 684 switch papers. The 10,247 template-ineligible papers have no defined comparison under this template; conditional subgroup correlations cannot be averaged into a corpus-wide effect. Support-topology definitions, eligibility filters, and sensitivity templates are reported in Appendices C.2 and C.2.3.

Table 22 follows the predicted topology-dependent hierarchy. In the separable-cycle regime, retaining cycle identity raises sensitive-subset OA correlation from 0.394 to 0.481 and matched-pair accuracy from 51.4% to 92.4%, while full zigzag matches the identity-aware scan. In the overlapping-cycle

Table 22: Topological evidence across support regimes. Matched-pair accuracy uses only the global topological summary; other metrics retain common relation terms.

Topological summary	All eligible OA ρ	Sensitive OA ρ	Matched Pair Acc. (%)	Detection AUROC
Panel A: cycle-separable G_{multi}				
Rank-run coherence	0.509 ± 0.009	0.394 ± 0.018	51.4 ± 1.2	0.748 ± 0.010
Cycle-ID interval scan	$\mathbf{0.536 \pm 0.008}$	$\mathbf{0.481 \pm 0.017}$	$\mathbf{92.4 \pm 1.6}$	$\mathbf{0.779 \pm 0.009}$
Full zigzag	$\mathbf{0.536 \pm 0.008}$	$\mathbf{0.481 \pm 0.017}$	$\mathbf{92.4 \pm 1.6}$	$\mathbf{0.779 \pm 0.009}$
Panel B: overlapping-cycle G_θ				
Fixed-basis class scan	0.526 ± 0.009	0.438 ± 0.018	68.9 ± 1.5	0.768 ± 0.010
Exhaustive classwise scan	0.531 ± 0.008	0.452 ± 0.017	74.6 ± 1.7	0.781 ± 0.009
Full zigzag	$\mathbf{0.548 \pm 0.008}$	$\mathbf{0.501 \pm 0.016}$	$\mathbf{89.1 \pm 1.5}$	$\mathbf{0.812 \pm 0.009}$

regime, the correspondence-aware score exceeds the exhaustive classwise scan by 0.049 switch-subset OA correlation, 14.5 matched-pair accuracy points, and 0.031 detection AUROC. The exact window-intersection calculation produces the same score as full zigzag and thus has no distinct detection advantage. The no-switch gain (0.009, interval including zero) is smaller than the switch-subset gain (0.049), and the latter covers 5.0% of this corpus. A coverage-only baseline reaches 0.219 OA correlation versus 0.539 for RUPTURETRACE on a coverage-matched subset; availability normalization and relocating masked anchors leave rankings nearly unchanged. Shuffling cross-modal alignments lowers the score by 0.173 and is detected at 0.907 AUROC. These results align with the theoretical distinction: run scans suffice for unique cycles, identity tracking suffices for separable classes, and correspondence-aware computations contribute when classes overlap. The result concerns the structural score, while the source-level entailment question remains governed by the support map.

D PROMPT

Zero-shot LLM Evaluation Prompt

System: You are a careful machine learning conference reviewer.

User: You are acting as an ICLR reviewer evaluating an ICLR submission.

Your task is to predict the mean Soundness score, the mean Presentation score, the mean Contribution score, the mean Overall Rating, and the final conference decision.

Soundness measures how well the paper's claims are supported by correct, rigorous, and appropriate theory, analysis, and/or experiments. Presentation measures how clear, well-structured, and easy to follow and evaluate the paper is. Contribution measures how much new knowledge, insight, or value the paper brings to the ICLR community.

Soundness, Presentation, and Contribution are scored on a 1-4 scale, where 1 = poor, 2 = fair, 3 = good, and 4 = excellent. Overall Rating is scored on a 1-10 scale, where 1 = strong reject, 3 = reject, not good enough, 5 = marginally below the acceptance threshold, 6 = marginally above the acceptance threshold, 8 = accept, good paper, and 10 = strong accept. Scores 2, 4, 7, and 9 may be used for intermediate cases.

Predict mean scores, not a single reviewer's scores. The scoring scales are discrete, but decimal values are allowed. Use the scale meanings as guidance, not as a requirement to output only integers. Treat Soundness, Presentation, Contribution, and Overall Rating as related but distinct quantities; do not compute one mechanically from the others. Predict the final decision separately. It is related to the scores, but is not determined mechanically by them alone.

Use only the provided manuscript content. Do not invent missing details. Focus on overall scientific contribution, novelty, technical soundness, support for the main claims, clarity sufficient for evaluation, and likely value to the ICLR community. Use the scales realistically and avoid unnecessary precision. Use at most two decimal places.

Now evaluate the following paper.

Paper: Title: `{title}`

Full Paper: `{full manuscript content}`

Output only:

Predicted Mean Soundness: `<a real number from 1.0 to 4.0>`
Predicted Mean Presentation: `<a real number from 1.0 to 4.0>`
Predicted Mean Contribution: `<a real number from 1.0 to 4.0>`
Predicted Mean Overall Rating: `<a real number from 1.0 to 10.0>`
Predicted Final Decision: `<Accept/Reject>`

Figure 2: Zero-shot prompt for predicting ICLR review scores and final decision from the target manuscript.

In-context LLM Evaluation Prompt

System: You are a careful machine learning conference reviewer.

User: You are acting as an ICLR reviewer evaluating an ICLR submission.

Your task is to predict the mean Soundness score, the mean Presentation score, the mean Contribution score, the mean Overall Rating, and the final conference decision.

Soundness measures how well the paper's claims are supported by correct, rigorous, and appropriate theory, analysis, and/or experiments. Presentation measures how clear, well-structured, and easy to follow and evaluate the paper is. Contribution measures how much new knowledge, insight, or value the paper brings to the ICLR community.

Soundness, Presentation, and Contribution are scored on a 1-4 scale, where 1 = poor, 2 = fair, 3 = good, and 4 = excellent. Overall Rating is scored on a 1-10 scale, where 1 = strong reject, 3 = reject, not good enough, 5 = marginally below the acceptance threshold, 6 = marginally above the acceptance threshold, 8 = accept, good paper, and 10 = strong accept. Scores 2, 4, 7, and 9 may be used for intermediate cases.

Predict mean scores, not a single reviewer's scores. The scoring scales are discrete, but decimal values are allowed. Use the scale meanings as guidance, not as a requirement to output only integers. Treat Soundness, Presentation, Contribution, and Overall Rating as related but distinct quantities; do not compute one mechanically from the others. Predict the final decision separately. It is related to the scores, but is not determined mechanically by them alone.

Use only the provided manuscript content. Do not invent missing details. Focus on overall scientific contribution, novelty, technical soundness, support for the main claims, clarity sufficient for evaluation, and likely value to the ICLR community. Use the scales realistically and avoid unnecessary precision. Use at most two decimal places.

The examples below come from prior ICLR reviews. Do not copy their wording or overfit to their topics.

Example `{k}`: Title: `{reference title}`

Abstract: `{reference abstract}`

Content of Paper: `{reference paper content}`

Mean Soundness: `{reference soundness}`
Mean Presentation: `{reference presentation}`
Mean Contribution: `{reference contribution}`
Mean Overall Rating: `{reference overall rating}`
Final Decision: `{reference decision}`

Now evaluate the following paper.

Paper: Title: `{title}`

Full Paper: `{full manuscript content}`

Output only:

Predicted Mean Soundness: `<a real number from 1.0 to 4.0>`
Predicted Mean Presentation: `<a real number from 1.0 to 4.0>`
Predicted Mean Contribution: `<a real number from 1.0 to 4.0>`
Predicted Mean Overall Rating: `<a real number from 1.0 to 10.0>`
Predicted Final Decision: `<Accept/Reject>`

Figure 3: In-context prompt for predicting ICLR review scores and final decision using prior reference papers and the target manuscript.

E Computational Resources

Experiments used a local compute pool with 4 NVIDIA A100 GPUs and 2 NVIDIA H200 GPUs. The GPUs were used mainly for encoder inference and batched embedding extraction; persistence computation, calibration, metrics, and post-processing were lightweight CPU-side operations.

F AI Use Statement

During revision, LLMs assisted with drafting and language editing. The authors reviewed, verified, and finalized the mathematical claims, proofs, methodology, experimental design, data processing,

reported numerical results, citations, and conclusions, and take responsibility for the final content of the work.

G ETHIC STATEMENT

This study is based exclusively on publicly available datasets and large language models, and does not involve the use of private data. All scientific concepts, methodological designs, experimental implementations, and resulting conclusions remain entirely the responsibility of the authors.