MECHTRACE: RELATION-COMPLETE EVIDENCE WEAVES FOR MECHANISTIC BIOMEDICAL QA

Anonymous authors
Paper under double-blind review

ABSTRACT

A mechanistic biomedical answer can name the correct drug, target, and outcome while leaving their relation unsupported, substituting endpoint plausibility for inspection of the mechanism. MechTrace makes requested relations inspectable with a connected evidence weave of typed edges, aligned snippets, provenance, and stable edge–snippet identifiers. The weave is both generator context and legal citation set; pre- and post-decoding checks expose the first relation without a valid witness. A path-matched comparison fixes upstream path discovery, generator, evidence snapshot, and retrieval budget, then contrasts this closed interface package with TypedPath-Open. On 600 fixed questions, MechTrace raises all-question correctness from 58.2% to 63.0% and reduces unsupported-relation escapes from 69 to 41, while coverage changes by 0.5 points. Among answered outputs, citation correctness improves by +2.5 points on BioASQ-mech and coherence by +2.7 points on DrugMechQA. The citation advantage grows from +0.8 to +4.6 points with BioASQ-mech link demand, persists with Qwen2.5-72B-Instruct, and accompanies 78.0% versus 49.0% detection when a required bridge is removed. These decision-level gains and missing-edge signatures demonstrate the verification value of relation-complete evidence weaves.

1 INTRODUCTION

Consider "How does imatinib treat chronic myeloid leukemia?" An answer can name imatinib, BCR–ABL1, and remission yet cite only those endpoints. Without a witness for the BCR–ABL1-to-pathway step, its mechanism cannot be checked. We call the target *intermediate-relation visibility*: every asserted relation must be locatable in the evidence shown to the generator and reader.

MechTrace realizes this target with one retained evidence weave W: typed edges, aligned snippets, provenance, and stable edge–snippet identifiers from a connected binding. Before decoding, each requested relation needs a connected witness; during decoding, only identifiers in W are legal citations; afterward, every asserted relation is checked against its cited witness. A failure returns the first unsatisfied relation. Thus the relation chain, serialized context, citation space, and failure signal remain views of one technical object.

TRACE constructs connected triple chains, Attribute First then Generate reuses selected spans as attributions, GAVEL validates subclaim–evidence identifiers, and S2G-RAG emits structured retrieval gaps (Fang et al., 2024; Slobodkin et al., 2024; Xu et al., 2026; Li et al., 2026). MedGraphRAG links graph retrieval to medical sources, while MedCite and LongCite sharpen claim-level attribution (Wu et al., 2025; Wang et al., 2025; Zhang et al., 2025). MechTrace contributes their biomedical relation-level conjunction: typed slots, connected edge–snippet witnesses, pre-decoding closure, a weave citation allowlist, typed-relation validation, and localized missing-link signaling.

The evaluation separates this interface from upstream path quality. TypedPath-Open shares normalized entities, path-discovery outputs, generator, response protocol, 2024 snapshot, and 3,000-token budget with MechTrace, but opens context, citation, and validation together. The estimand is therefore the complete closed-versus-open package. All-question correctness and unsupported-relation escapes retain all 600 questions; answered-output supportability is reported separately. Link-demand strata, generator transfer, and removed-edge intervention test whether the effect tracks missing intermediate relations.

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.

With just three fewer substantive answers, the closed package returns 29 more correct mechanisms and allows 28 fewer unsupported-relation escapes than TypedPath-Open. Two mechanism-specific signatures sharpen that decision-level result: the BioASQ-mech citation advantage increases from +0.8 to +4.6 points as minimum link demand rises, and removed-bridge detection reaches 78.0% versus 49.0%. Positive citation and coherence effects across both generators provide a complementary check that the ordering is not tied to one decoder.

Contribution. The paper connects an annotation-ready target—intermediate-relation visibility—to an executable evidence interface and decision-level evaluation. An omitted bridge becomes a localized system outcome, and the path-matched study identifies when preserving it changes biomedical QA behavior.

2 Related Work

Mechanistic verification requires a connected relation chain, textual support for each edge, and citations that remain bound to those witnesses. Table 1 locates MechTrace by the object retained across these requirements.

Evidence attribution and contracts. Attribute First then Generate reuses selected source segments as local attributions; LongCite and MedCite align claims with fine-grained spans (Slobodkin et al., 2024; Zhang et al., 2025; Wang et al., 2025). GAVEL binds atomic subclaims to evidence units and validates identifiers and quotations (Xu et al., 2026). MechTrace changes the granularity: each citation atom joins a biomedical typed edge to a supporting snippet, and valid atoms must form the requested connected mechanism.

Reasoning chains and structured gaps. TRACE uses connected knowledge-triple chains, directly or through their source documents, as generation context (Fang et al., 2024). S2G-RAG emits structured bridge, relation, or span gaps from a sentence-level context to guide retrieval (Li et al., 2026). MechTrace combines the chain object with edge–snippet witnesses, an object-derived citation allowlist, and relation-local checks before and after decoding.

Biomedical graph retrieval and insufficiency. MedGraphRAG connects medical documents and controlled vocabularies through source-linked graph retrieval (Wu et al., 2025). Self-RAG regulates retrieval through reflection, and KnowGuard identifies insufficiency before answer commitment (Asai et al., 2024; Dang et al., 2026). MechTrace specializes these ideas to ordered biomedical relation slots: an insufficiency names the missing slot, while a valid answer cites a type-compatible witness for every slot.

Comparators in the shared protocol. Domain retrieval systems improve medical passage access and ordering (Jin et al., 2023; Sohn et al., 2025; Li et al., 2025; Zhang et al., 2026; Punuru et al., 2026). ColMedRAG tests late-interaction ranking over the shared pool (Khattab & Zaharia, 2020), and GraphRAG tests structured neighborhoods. TypedPath-Open is the closest intervention: it shares MechTrace's entities, snippets, paths, scores, operating point, and generator, then opens conditioning, citation, and validation jointly. Its paired contrast identifies the downstream interface package.

Table 1: Closest design precedents organized by their retained evidence object and enforcement boundary.

System	Retained object	Continuity mechanism	Citation or failure boundary
Attribute First	selected spans	selection reused for attribution	local attributed spans
TRACE	connected triples	chain used as generation context	chain or source documents
S2G-RAG	sentence context + gap items	iterative sufficiency control	structured next-hop gap
GAVEL	subclaim–evidence bindings	evidence contract + scrutiny	identifier/quote validation
MedGraphRAG	hierarchical medical graph	source-linked graph retrieval	credible source documents
MechTrace	typed edge–snippet weave	connected witnesses for ordered slots	weave allowlist + relation-local validation

MechTrace contributes the final-row conjunction: connected typed slots, textual edge witnesses, one legal citation space, and a localized validation outcome. Section 6.1 evaluates it after path construction is fixed, with ColMedRAG as the complementary passage-ranking comparison.

3 Relation-Complete Evidence Bundles

MechTrace makes the selected relation chain the persistent interface between retrieval and decoding. Given a biomedical question q, it derives required relation slots, links biomedical mentions to canonical entities, enumerates typed paths, and attaches textual support to every path edge. A budgeted union of the ranked path–text units becomes the evidence weave: the generator context, legal citation set, and object retained for post-decoding validation.

Formally, $W = (V_W, E_W, S_W, \phi)$, where V_W and E_W are retained typed nodes and edges, S_W is the set of supporting snippets, and $\phi : E_W \to 2^{S_W}$ maps each edge to its snippets. Let $I_W = \{\mathrm{PairId}(e, s) : e \in E_W,\ s \in \phi(e)\}$ be the stable edge–snippet citation atoms. The evidence-bundle contract imposes three invariants on W. *Support closure* requires an edge–snippet witness for every requested relation. *Citation closure* requires every emitted citation to belong to I_W. *Omission observability* names the relation slot that has no valid witness.

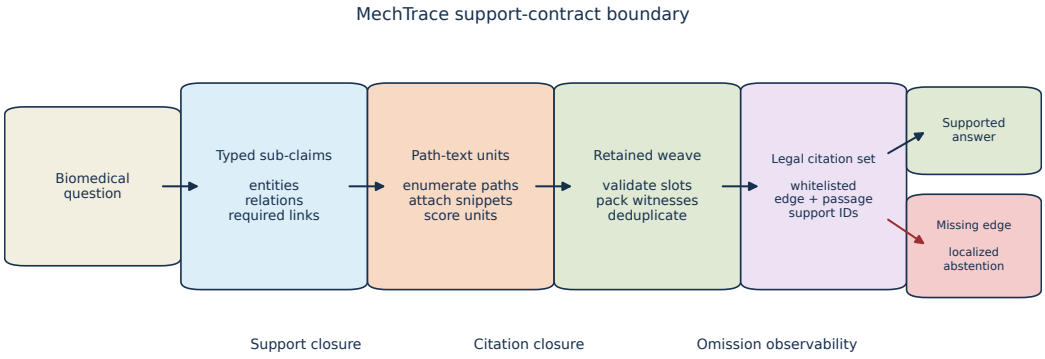

Figure 1: MechTrace maps typed sub-claims to ranked paths, a retained weave, and a legal support-ID set. Generation is conditioned on that weave only. Missing required links therefore surface as explicit abstentions.

Figure 1 exposes the enforcement boundary. Retrieval constructs edge–snippet path units; the closure validator selects a connected witness chain for the required relations; and only the validated weave

and its support-ID whitelist cross into generation. Upstream path discovery and the downstream bundle interface remain separate stages, enabling the controlled comparison in Section 6.1.

3.1 Question analysis and path ranking

Required relations and linking. The decomposer emits a validated record containing normalized `entities`, ordered `subclaims`, `required_relations`, and a `path_template`. Each required relation r_i specifies typed source and destination arguments and an allowed relation family. Coordination creates separate slots; "how" and "why" questions retain intervening target–pathway and pathway–phenotype slots rather than collapsing them into endpoint pairs. Appendix Table 9 gives the finite template inventory. Mentions resolve first by exact normalized alias in UMLS or DrugBank and otherwise to the type-compatible candidate with highest SapBERT cosine similarity (Liu et al., 2021); unresolved required entities trigger a linking-specific abstention. Similarity ties are broken by canonical identifier.

Paths, snippets, and scores. The evidence graph $G = (V, E)$ contains typed drugs, genes, diseases, pathways, phenotypes, and publications (Appendix Figure 3). Each edge stores relation type, source family, database or PMID identifier, and source confidence. For each seed, incident type-compatible edges are ordered by SapBERT similarity between q and the verbalized edge, then by source confidence and canonical edge ID; the first K destination nodes define the expanded neighborhood. MechTrace enumerates simple paths $\pi = (v_1, e_1, \ldots, e_k, v_{k+1})$ of length $k \leq L$ within that neighborhood.

Text is attached before scoring. A sentence is eligible for edge $e = (v, \rho, v')$ only when its normalized mentions include v and v' and its extracted or template-matched relation is compatible with ρ; extracted relations must meet the graph's 0.85 confidence threshold. Among eligible sentences, $\phi(e)$ is the span maximizing SapBERT similarity to q concatenated with a verbalization of e; stable support identifier breaks ties. Otherwise $\phi(e) = \varnothing$, and the edge cannot witness a relation. The raw semantic score u_{sem} averages $(\cos + 1)/2$ for supported edges and 0 for unsupported edges over the entire path. The coherence score u_{coh} is the fraction of requested relation slots realized by type-compatible path edges. The provenance score u_{prov} averages source confidence: curated edges use the database confidence (or 1 when ungraded), and extracted edges use relation-extractor confidence.

Each component is min–max normalized within the question's candidate set; a component constant over that set is assigned 1. With $\widehat{u}$ denoting the normalized value, paths are scored by

$$s(\pi \mid q) = 0.35\,\widehat{u}_{\text{sem}}(\pi, q) + 0.35\,\widehat{u}_{\text{coh}}(\pi) + 0.30\,\widehat{u}_{\text{prov}}(\pi). \tag{1}$$

Ties are resolved by fewer edges and then lexicographically by edge identifiers. A disjoint 60-question calibration selects $L = 3$, $K = 40$, $N = 120$, $\tau = 0.4$, and weights $0.35/0.35/0.30$ before evaluation; Appendix Section B.2 gives the grid, objective, and tie rule.

3.2 Closure, packing, and decoding

Paths with $s(\pi \mid q) \geq \tau$ enter the qualified pool. The validator enumerates complete bindings $b = (w_1, \ldots, w_m)$ in which witness $w_i = (\pi, e, s)$ realizes r_i, adjacent slots share the same normalized intermediate entity, and all slots follow the requested order. It selects the binding with greatest summed path score, breaking ties by fewer distinct paths and canonical IDs. If no complete binding exists, the first unsatisfied or disconnected slot is named in the abstention. The selected connected witnesses are packed first into W; additional paths for which every edge has an eligible snippet are then added in score order. Merge and deduplication canonicalize repeated entities and sentence spans. A path record is never truncated between an edge and its support. If the required witness binding exceeds the $N = 120$ node cap or the shared 3,000-token evidence budget, MechTrace names the first relation that cannot be retained and abstains. Budget pruning can therefore remove optional context but cannot silently break support closure.

Each retained edge and snippet have stable component IDs, such as `E12` and `P07`; the legal citation atom is their pair, `E12:P07`. The serializer emits ordered typed triples followed immediately by edge-aligned snippets, provenance, and pair IDs. The fixed response template requests an answer, an

intermediate relation chain, and an evidence table. Generation sees only W and may cite only its pair-ID whitelist.

Post-decoding validation checks that every cited ID is legal and that every asserted mechanistic relation resolves through ϕ to a cited witness with compatible endpoint and relation types. A failed check rejects the decode and returns the corresponding relation-specific insufficiency signal. Pre-decoding closure thus asks whether the retained evidence can support the requested chain; post-decoding validation asks whether the generated chain stayed within that evidence. Appendix Algorithm 1 specifies the complete sequence, and Appendix Table 17 reports sensitivity to the retained-node cap.

Naive enumeration of paths has complexity $O(d^L)$, where d is the pruned neighborhood degree. In practice, applying entity constraints, type templates, and semantic filters leaves roughly 200–800 candidate paths per query; CPU scoring precedes delivery of the final weave to the shared generator.

4 BENCHMARKS

The evaluation spans three complementary linkage regimes. BioASQ-mech anchors the study in externally authored mechanistic questions, DrugMechQA supplies controlled drug–target–pathway chains with known relation demand, and PubMedQA represents literature-grounded short-chain questions (Tsatsaronis et al., 2015; Wei et al., 2019; Jin et al., 2019). Every method retrieves from one fixed 2024 snapshot of PubMed abstracts, DrugBank 5.1.10, Reactome v84, DisGeNET v7.0, STRING v11.5, and UMLS 2024AA (Wishart et al., 2018; Jassal et al., 2020; Piñero et al., 2020; Szklarczyk et al., 2021; Bodenreider, 2004). Source roles and graph scale appear in Appendix Tables 8 and 11.

Table 2: Dataset source-pool or collection sizes and the evidence snapshot available to every method.

Dataset	#Questions	Evidence sources
BioASQ-mech	1,200	PubMed + KG
PubMedQA	1,000	PubMed
DrugMechQA	800	PubMed + KG

BioASQ-mech retains 1,200 public questions requiring a literature-supported causal, therapeutic, regulatory, or pathway relation; PubMedQA starts from 1,000 expert-labeled source questions. Two annotators mark the shortest sufficient chain and PMID spans. Seed 2027 fixes stratified 200-question samples with one/two/three-or-more-link counts of $68/87/45$ and $94/78/28$, respectively. Appendix Section B specifies eligibility and deterministic screening.

DrugMechQA joins DrugBank drug–target edges to Reactome pathway chains and aligned PubMed support. Grouping exact normalized full chains before the 400/200/200 train/development/test assignment prevents rendered variants of one mechanism from crossing splits; measured subpair overlap is 2.1% for drug–target and 1.8% for target–pathway pairs. Quality control accepts 113/120 sampled forms (94.2%, $\kappa = 0.84$) before replacements and freezing. DrugMechQA is thus a template-aligned relation-continuity diagnostic, complementing externally authored BioASQ-mech. Calibration items are disjoint from the 600 evaluation questions, and systems never receive annotation-side link count. Appendix Section B gives construction details; Tables 10, 18, and 23 give examples and chain-demand analyses.

5 MATCHED-BUDGET PROTOCOL

Resource matching identifies the decision consequence of the complete evidence-bundle interface. Snapshot, normalization, generator, response and citation formats, and token ceilings are fixed; evidence selection and the bundle interface vary. TypedPath-Open fixes path discovery while testing the closed interface package, ColMedRAG tests passage ranking, Self-RAG tests adaptive control, and GraphRAG tests structured context without closure. Three-seed aggregation yields 600 paired question units. All-question correctness and unsupported-relation escapes retain every question; answered-output supportability is conditional on substantive responses.

5.1 COMPARATORS

DenseRAG uses a BM25+Contriever passage pool (Izacard et al., 2022; Karpukhin et al., 2020). Our GraphRAG control appends normalized one-hop entity neighborhoods to the same retrieved passages; ProvenanceRAG reweights passages by source family; MedCPT-RAG applies MedCPT and a biomedical cross-encoder (Jin et al., 2023); and Self-RAG adds critique-triggered retrieval (Asai et al., 2024). ColMedRAG applies ColBERT-style late interaction over the same PubMed snapshot and reranks the shared top-100 pool (Khattab & Zaharia, 2020). Table 3 fixes the evaluated implementations and operating settings before the 600-question study; all cross-system comparisons refer to these configured rows.

TypedPath-Open copies MechTrace's entity links, candidate paths, edge snippets, path scores, and L, K, N, τ. Per question, it receives the identical typed-triple block and the same text-token allotment, but fills that allotment with question-ranked PubMed passages irrespective of path membership. Every in-context passage ID is legal, while sub-claim closure, relation-localized missing-edge signaling, and post-decoding edge–citation validation are disabled. The comparison thus tests the downstream bundle interface as a package after holding path discovery, ranking, and token allocation fixed; Appendix Section E.1 specifies ordering and overflow.

Table 3: Operational baseline configuration. Every row uses the shared generator, answer template, evidence snapshot, and 3,000-token retrieval cap.

Method	Selection unit	Distinguishing control	Citation unit
DenseRAG	BM25 + Contriever passages	top-10 passages	passage IDs + spans
GraphRAG	passages + one-hop neighbors	up to 50 neighbors/entity	passage IDs + graph facts
Prov. RAG	source-reweighted passages	top-10 passages	passage IDs + spans
MedCPT-RAG	MedCPT + cross-encoder	top-100 to top-10	passage IDs + spans
Self-RAG	BM25 + Contriever passages	up to 2 critique rounds	final passage IDs
ColMedRAG	late-interaction passages	top-100 to top-10	passage IDs + spans
TypedPath-Open	paths + open text	closure checks disabled	any context passage ID
MechTrace	closed evidence weave	at most 120 nodes	edge-aligned weave IDs

5.2 SHARED-BUDGET PROTOCOL

All methods share normalized entity constraints and the corpus snapshot. Passage methods consume text, GraphRAG adds serialized neighbors, TypedPath-Open combines selected paths with unrestricted passages, and MechTrace consumes the closed weave. Evidence units enter in method rank order; typed paths, graph records, and edge–snippet pairs remain atomic, while a passage may use only the complete-sentence prefix fitting the 3,000-token cap. The answer template standardizes sub-claim and citation slots without changing each method's selection or citation unit. Closure and support-ID validation apply only to MechTrace. Table 4 summarizes fixed resources and varied representations.

Table 4: Fair-budget configuration. Fixed resources remove generator, source, and context-window advantages while permitting each evidence representation to allocate the shared cap.

Component	Fixed across methods	Intentionally varied
Evidence	PubMed 2024 snapshot, fixed KG, shared normalization	passage, neighborhood, path, or weave representation
Generator	Llama-3.1-70B-Instruct; temperature 0.2; top-p 0.95; 512 new tokens	none
Budget	3,000 retrieval tokens; 8,192-token context; common response template	allocation among selected evidence units
Retrieval	same source snapshot and top-level candidate access	reranking, expansion, reflection, or path selection

Table 5: Decision-level behavior for the path-matched interface packages over the same 600 questions. Abstentions remain in the all-question denominator. Brackets are 95% question-bootstrap intervals; counts are in parentheses.

Method	Coverage ↑	All-Q correct. ↑	Cond. correct. ↑	Escapes ↓
TypedPath-Open	91.2 [88.9,93.4] (547)	58.2 [54.2,62.1] (349)	63.8 (349/547)	69
MechTrace	90.7 [88.3,92.9] (544)	63.0 [59.1,66.8] (378)	69.5 (378/544)	41

PubMed indexing uses BM25+Contriever fusion with $\alpha_{\text{BM25}} = 0.4$ and $\alpha_{\text{dense}} = 0.6$, implemented with FAISS IVF-PQ (Johnson et al., 2021; Robertson & Zaragoza, 2009). ColMedRAG reranks token-level late-interaction scores; MedCPT-RAG uses its biomedical cross-encoder; and the two path systems enumerate approximately 200–800 type-valid candidates per query before retaining the operating-point path set. The global graph counts in Appendix Table 11 describe the retrieval substrate, whereas this candidate range describes a single query.

5.3 Models, hardware, and measurement protocol

Retrieval runs on dual AMD EPYC 7713 CPUs with 512 GB RAM. Generation uses 8×NVIDIA A100 80GB GPUs, CUDA 12.2, driver 535.104.05, and vLLM 0.4.2 with tensor parallelism. Generator-transfer experiments substitute Qwen2.5-72B-Instruct while preserving all other controls. Retrieval latency excludes offline indexing and final LLM decoding.

After a disjoint 20-question calibration, two biomedical literature annotators evaluate the fixed 200 questions per dataset, independently and then by adjudication. Method names, assignments, and internal retrieval state are hidden; randomized presentation masks condition order. Annotators see the generator's evidence and citation representation because that interface is the object evaluated. Faithfulness measures factual-sentence support, citation correctness measures relation-level claim–span alignment, and coherence requires the ordered mechanism; pre-adjudication κ is $0.78/0.74/0.72$.

Binary outcomes use a two-of-three majority. All-question correctness is primary: abstention and substantive error both score zero, with coverage and escapes decomposing the total. Answered-seed fractional scores are averaged within questions only for conditional supportability. Intervals use 1,000 question-level bootstrap resamples, paired for displayed differences; stress and preference rates use Wilson intervals. Prespecified conditional endpoints are MechTrace–TypedPath-Open citation correctness on BioASQ-mech and coherence on DrugMechQA; others are unadjusted secondary analyses. Appendix C gives complete denominators and rules.

For the usefulness study, biomedical evaluators blinded to condition labels and assignment order compare MechTrace with TypedPath-Open, the path-matched interface intervention, on 150 stratified questions (50 per dataset). They view only the question, answer, and cited evidence—including the interface format whose verification utility is being compared—and choose the more useful answer or a tie. The missing-edge stress test removes one adjudicated required edge from 200 questions while retaining plausible endpoint passages. Detection requires the output to name that relation as unsupported, whether validator-triggered or model-generated. The test also records intervened escapes, intact-control false abstention and correctness, and retrieval-side latency.

6 Results

6.1 Decision-level correctness on the complete question set

Table 5 retains every question and scores abstentions as incorrect. Over the same 600 questions, MechTrace produces 378 correct responses versus 349 for TypedPath-Open and reduces unsupported-relation escapes from 69 to 41. This is not broad refusal: MechTrace answers 544 questions versus 547, while conditional correctness is 69.5% versus 63.8%.

Appendix Table 20 resolves the aggregate: DrugMechQA adds 17 correct answers and removes 16 escapes, while externally authored BioASQ-mech adds 8 and removes 8.

Table 6: Conditional supportability effects of MechTrace minus TypedPath-Open (TP) and ColMedRAG (CM) among answered outputs. Cells report point deltas with 95% paired bootstrap intervals.

Dataset	TP Δ Citation	TP Δ Coherence	CM Δ Citation	CM Δ Coherence
BioASQ-mech	+2.5 [+0.9,+4.1]	+2.5 [+0.8,+4.2]	+4.8 [+3.0,+6.6]	+4.2 [+2.2,+6.2]
PubMedQA	+1.7 [+0.3,+3.1]	+1.2 [-0.1,+2.5]	+2.1 [+0.5,+3.7]	+1.0 [-0.4,+2.4]
DrugMechQA	+2.6 [+0.8,+4.3]	+2.7 [+0.9,+4.5]	+5.3 [+3.3,+7.3]	+5.9 [+3.7,+8.1]

Table 7: Supportability among answered outputs from fixed 200-question samples. Brackets are 95% question-bootstrap intervals; significance marks compare MechTrace with Self-RAG.

Setting	Method	Faithfulness ↑	Citation corr. ↑	Coherence ↑
Aggregated	Self-RAG	82.2	76.7	69.5
	ColMedRAG	84.8	78.4	71.5
	TypedPath-Open	85.8	80.2	73.0
	MechTrace	87.2	82.4	75.2
BioASQ-mech	Self-RAG	81.0 [77.5,84.4]	75.6 [72.0,79.1]	68.9 [65.2,72.5]
	ColMedRAG	83.8 [80.3,87.2]	77.6 [73.9,81.2]	70.9 [67.1,74.6]
	TypedPath-Open	84.6 [81.2,87.9]	79.9 [76.4,83.3]	72.6 [68.9,76.2]
	MechTrace	86.2** [82.9,89.4]	82.4** [79.0,85.7]	75.1** [71.5,78.6]
PubMedQA	Self-RAG	83.2 [79.7,86.6]	78.1 [74.5,81.6]	70.2 [66.4,73.9]
	ColMedRAG	85.8 [82.3,89.2]	79.1 [75.5,82.6]	71.8 [68.0,75.5]
	TypedPath-Open	85.4 [81.9,88.8]	79.5 [75.9,83.0]	71.6 [67.8,75.3]
	MechTrace	86.1* [82.7,89.4]	81.2** [77.7,84.6]	72.8* [69.1,76.4]
DrugMechQA	Self-RAG	82.4 [78.7,86.0]	76.5 [72.7,80.2]	69.3 [65.4,73.1]
	ColMedRAG	84.9 [81.2,88.5]	78.4 [74.5,82.2]	71.7 [67.6,75.7]
	TypedPath-Open	87.3 [83.7,90.8]	81.1 [77.3,84.8]	74.9 [70.9,78.8]
	MechTrace	89.4** [86.1,92.6]	83.7** [80.2,87.1]	77.6** [74.0,81.1]

6.2 Conditional supportability under path matching

TypedPath-Open preserves path discovery and scoring while opening text context, citation legality, relation-local signaling, and post-decoding validation. Its paired contrast estimates the complete interface package. Among answered outputs, Table 6 shows positive citation effects throughout, including the prespecified +2.5 points on BioASQ-mech; the other prespecified endpoint is +2.7 points in DrugMechQA coherence.

On short-chain PubMedQA, citation correctness improves while the smaller coherence interval crosses zero. Larger effects against ColMedRAG on BioASQ-mech and DrugMechQA show that passage ranking alone does not reproduce relation continuity. Appendix Table 19 adds paired faithfulness effects.

6.3 Competitive context across linkage regimes

Table 7 compares the four strongest configured implementations. MechTrace leads every answered-output cell; separation is smallest for PubMedQA faithfulness and largest for citation correctness and coherence on the higher-link datasets.

Faithfulness margins over Self-RAG are 5.2, 2.9, and 7.0 points across the three datasets; those over MedCPT-RAG are 6.4, 4.7, and 8.5.

6.4 Mechanism signatures and verification utility

Within BioASQ-mech, the citation effect over TypedPath-Open grows from +0.8 on one-link to +4.6 on three-or-more-link questions, a +1.85-point paired trend per bin [95% CI +0.42,+3.28], $p = 0.012$. Qwen2.5-72B-Instruct preserves positive citation and coherence effects on every dataset, tying the package effect to relation demand rather than one decoder.

Usefulness and removed-edge stress. Evaluators select MechTrace on 87 of 150 preference questions, TypedPath-Open on 43, and tie on 20 (66.9% among non-ties). Removing one required bridge raises detection from 98/200 to 156/200 and lowers escapes from 47 to 18. On intact controls, MechTrace records 122 correct answers and 12 false abstentions versus 114 and 8, exposing both sensitivity gain and specificity cost. Appendix Tables 21–26 report complete supporting analyses.

7 Discussion

With only three fewer answered questions, MechTrace returns 29 more correct mechanisms and 28 fewer unsupported-relation escapes than TypedPath-Open. Because that comparator shares path discovery and scoring but opens context, citation, signaling, and validation together, the estimate belongs to the complete interface package: one biomedical relation chain remains the conditioning, citation, and validation object.

The package advantage grows with BioASQ-mech link demand, persists under Qwen2.5-72B-Instruct, and becomes pronounced when a required edge is removed but endpoint passages remain. PubMedQA supplies the short-chain regime, where faithfulness is close but citation correctness improves. The 87–43 preference result with 20 ties further links relation continuity to both answer behavior and the reader's evidence surface.

The evaluated regime covers normalized biomedical entities, mechanisms expressible in the typed inventory, the fixed 2024 snapshot, and the reported question strata and generators. DrugMechQA probes template-aligned chains; externally authored BioASQ-mech tests the behavior beyond that construction. Evaluators were blinded to labels and order, while seeing the evidence format whose utility was measured. Stronger missing-edge detection carries 6.0% false abstention on intact controls versus 4.0% for TypedPath-Open; this tradeoff, and relations outside the inventory, surface as localized insufficiency rather than unsupported mechanistic prose.

The weave also localizes engineering action: illegal identifiers point to decoding, absent witnesses to selection or graph coverage, and type mismatches to relation resolution. The same separation applies wherever scientific conclusions depend on intermediate relations.

8 Conclusion

MechTrace keeps connected biomedical relations, textual witnesses, legal citations, and relation-local validation aligned through decoding. Against the path-matched open package, all-question correctness rises from 58.2% to 63.0% and unsupported escapes fall from 69 to 41 at nearly unchanged coverage. Link-demand growth, generator transfer, and 78.0% removed-bridge detection locate the effect where intermediate relations must survive. Relation-complete evidence weaves turn hidden mechanistic omissions into localized outcomes while improving biomedical QA correctness.

Ethics Statement

The study evaluates model outputs over public or licensed evidence and uses no patient records or participant intervention. Its intended setting is auditable research assistance, in which readers inspect cited chains before acting on an answer.

AI use statement

Generative AI tools assisted with language and presentation editing during manuscript preparation, including formatting and consistency checks. The authors are responsible for the final manuscript.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004. doi: 10.1093/nar/gkh061. URL https://doi.org/10.1093/nar/gkh061.

Xilin Dang, Kexin Chen, Xiaorui Su, Ayush Noori, Iñaki Arango, Lucas Vittor, Xinyi Long, Yuyang Du, Marinka Zitnik, and Pheng-Ann Heng. KnowGuard: Knowledge-driven abstention for multi-round clinical reasoning. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://iclr.cc/virtual/2026/poster/10008150.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. TRACE the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8472–8494. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.496. URL https://aclanthology.org/2024.findings-emnlp.496/.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jKN1pXi7b0.

Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, Fred Loney, Bruce May, Marija Milacic, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1):D498–D503, 2020. doi: 10.1093/nar/gkz1031. URL https://doi.org/10.1093/nar/gkz1031.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259/.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023. doi: 10.1093/bioinformatics/btad651. URL https://doi.org/10.1093/bioinformatics/btad651.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572. URL https://doi.org/10.1109/TBDATA.2019.2921572.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075. URL https://dl.acm.org/doi/10.1145/3397271.3401075.

Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. BiomedRAG: A retrieval augmented large language model for biomedicine. *Journal of Biomedical Informatics*, 162:104769, 2025. doi: 10.1016/j.jbi.2024.104769. URL https://www.sciencedirect.com/science/article/pii/S1532046424001874.

Minghan Li, Junjie Zou, Xinxuan Lv, Chao Zhang, and Guodong Zhou. S2G-RAG: Structured sufficiency and gap judging for iterative retrieval-augmented QA. In *Proceedings of the 64th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 25846–25862. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1185. URL https://aclanthology.org/2026.acl-long.1185/.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4228–4238, 2021. doi: 10.18653/v1/2021.naacl-main.334. URL https://aclanthology.org/2021.naacl-main.334/.

Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I. Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 2020. doi: 10.1093/nar/gkz1021. URL https://doi.org/10.1093/nar/gkz1021.

Pranav Punuru, Nabil Ibtehaz, Swagarika Giri, Harsha Srirangam, Emilia Tugolukova, and Daisuke Kihara. Queryome: orchestrating retrieval, reasoning, and synthesis across biomedical literature. *Journal of Biomedical Semantics*, 2026. doi: 10.1186/s13326-026-00366-5. URL https://doi.org/10.1186/s13326-026-00366-5.

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

Eric W. Sayers, Evan E. Bolton, J. Rodney Brister, Kathi Canese, Jessica Chan, Donald C. Comeau, Ryan Connor, Kathryn Funk, Chris Kelly, Sunghwan Kim, Tom Madej, Aron Marchler-Bauer, Christopher Lanczycki, Stacy Lathrop, Zhiyong Lu, Françoise Thibaud-Nissen, Terence Murphy, Lon Phan, Yuri Skripchenko, Tony Tse, Jiyao Wang, Rebecca Williams, Barton W. Trawick, Kim D. Pruitt, and Stephen T. Sherry. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 50(D1):D20–D26, 2022. doi: 10.1093/nar/gkab1112. URL https://doi.org/10.1093/nar/gkab1112.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3344. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.182. URL https://aclanthology.org/2024.acl-long.182/.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. Rationale-guided retrieval augmented generation for medical question answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12739–12753, 2025. doi: 10.18653/v1/2025.naacl-long.635. URL https://aclanthology.org/2025.naacl-long.635/.

Damian Szklarczyk, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T. Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars Juhl Jensen, and Christian von Mering. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 2021. doi: 10.1093/nar/gkaa1074. URL https://doi.org/10.1093/nar/gkaa1074.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138, 2015. doi: 10.1186/s12859-015-0564-6. URL https://doi.org/10.1186/s12859-015-0564-6.

Xiao Wang, Mengjue Tan, Qiao Jin, Guangzhi Xiong, Yu Hu, Aidong Zhang, Zhiyong Lu, and Minjia Zhang. MedCite: Can language models generate verifiable text for medicine? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18891–18913, 2025. doi: 10.18653/v1/2025.findings-acl.967. URL https://aclanthology.org/2025.findings-acl.967/.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. PubTator Central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593, 2019. doi: 10.1093/nar/gkz389. URL https://doi.org/10.1093/nar/gkz389.

David S. Wishart, Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018. doi: 10.1093/nar/gkx1037. URL https://doi.org/10.1093/nar/gkx1037.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28443–28467. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1381. URL https://aclanthology.org/2025.acl-long.1381/.

Ruoyu Xu, Gaoxiang Li, and Victor S. Sheng. GAVEL: Evidence-contract debate with mechanized scrutiny for provenance-grounded fact-checking. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35907–35920. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1789. URL https://aclanthology.org/2026.findings-acl.1789/.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5098–5122, 2025. doi: 10.18653/v1/2025.findings-acl.264. URL https://aclanthology.org/2025.findings-acl.264/.

Yiqing Zhang, Xiaozhong Liu, and Fabricio Murai. PubMed Reasoner: Dynamic reasoning-based retrieval for evidence-grounded biomedical question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16753–16763, 2026. doi: 10.18653/v1/2026.acl-long.763. URL https://aclanthology.org/2026.acl-long.763/.

A DATA, GRAPH, AND DISTRIBUTION DETAILS

Table 8: Source versions, access conditions, and roles in the fixed retrieval snapshot.

Source	Version	Access or license note	Role in this paper
PubMed abstracts	2024 snapshot	Publicly accessible via NCBI services (Sayers et al., 2022)	Literature retrieval corpus
DrugBank	5.1.10	Academic research access; commercial reuse requires separate licensing	Drug–target relations and indications
Reactome	v84	Reactome data are distributed under public data terms (current data license: CC0)	Pathway relations and pathway identifiers
DisGeNET	v7.0	Free academic access for non-commercial research; broader reuse requires DISGENET terms compliance	Gene–disease associations
STRING	v11.5	STRING downloads are available under CC BY 4.0	Protein association edges
UMLS	2024AA	Active UMLS Metathesaurus license required for download and reuse	Identifier mapping and normalization

The distribution view complements mean scores by showing whether gains are concentrated in a small tail or shift the typical answered output. Figure 2 reports the medians and dispersion used for that comparison.

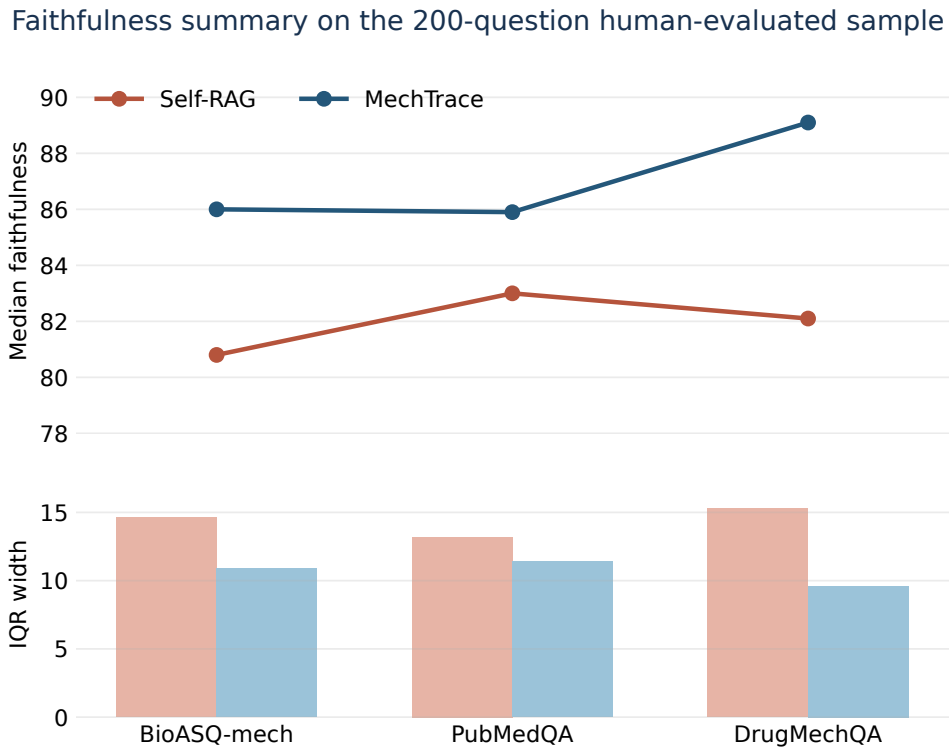

Figure 2: Faithfulness median and IQR on the human-evaluated sample. Exact values appear in Table 16.

The typed schema in Figure 3 defines the endpoint types and relation families that the path enumerator may connect. Publication nodes contribute textual witnesses and provenance rather than substitute endpoints in the requested mechanism.

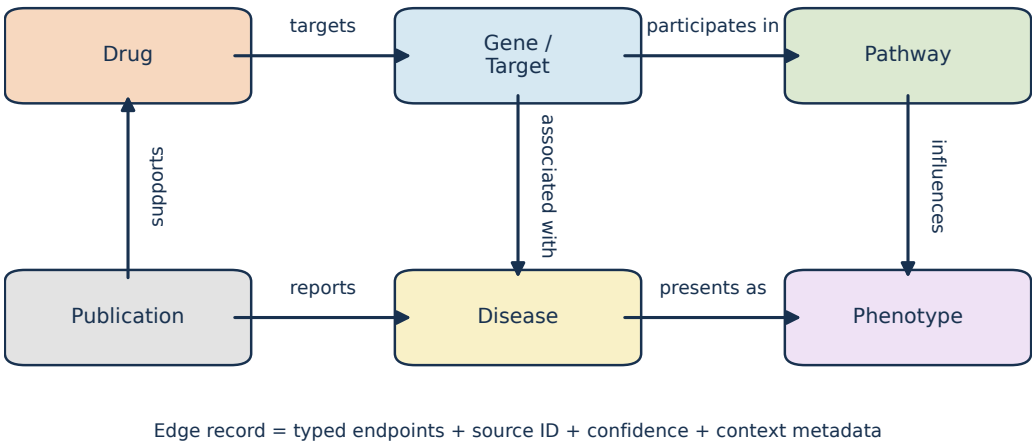

Figure 3: Typed evidence graph schema. MechTrace retrieves selected typed paths and exposes only their retained weave to the generator.

Table 9: Relation templates used by question decomposition. Each slot admits schema aliases with the same typed endpoints.

Question intent	Ordered required relation slots
Therapeutic mechanism	drug $\xrightarrow{\text{targets}}$ gene/protein $\xrightarrow{\text{modulates}}$ pathway $\xrightarrow{\text{affects}}$ disease/phenotype
Gene mechanism	gene $\xrightarrow{\text{participates in}}$ pathway $\xrightarrow{\text{affects}}$ disease/phenotype
Variant effect	variant $\xrightarrow{\text{alters}}$ gene/protein $\xrightarrow{\text{affects}}$ phenotype
Direct association	biomedical entity $\xrightarrow{\text{associated with}}$ disease/phenotype
Intervention outcome	intervention $\xrightarrow{\text{affects}}$ outcome, with an intermediate slot when requested by "how" or "why"

Table 10: Representative question forms and their annotation-side chains. Bracketed arguments denote normalized entities rather than additional evaluation records.

Dataset	Question form	Shortest sufficient chain
BioASQ-mech	How does imatinib treat chronic myeloid leukemia?	imatinib → BCR–ABL1 → signaling pathway → CML phenotype
PubMedQA	Does [exposure] affect [clinical outcome]?	exposure → outcome; add mediators explicitly required by the abstract
DrugMechQA	Through which target and pathway does [drug] affect [phenotype]?	drug → target → pathway → phenotype

Table 11: Evidence graph summary after normalization and pruning. These global substrate counts are distinct from the 200–800 type-valid candidate paths enumerated for one query.

Item	Count	Source or notes
Nodes	2,147,832	genes, diseases, drugs, pathways, phenotypes, publications
Edges (curated)	18,412,507	DrugBank, Reactome, DisGeNET, STRING, UMLS
Edges (extracted)	11,738,291	PubMed abstracts, confidence ≥ 0.85
Avg. degree	28.6	measured on the pruned typed projection
Provenance types	6	curated DB, literature, extraction, dataset, ontology, mapping

The schema and template inventory delimit what counts as a recoverable mechanistic link. The graph statistics describe the global search substrate; each question exposes only the type-compatible neighborhood and relation slots defined above.

B Benchmark Construction and Sampling

BioASQ-mech. We screen BioASQ factoid and explanatory questions whose gold answer and cited PubMed material identify a biomedical intervention, entity, process, or phenotype. Eligibility requires at least one directed causal, therapeutic, regulatory, or pathway relation. We exclude definition-only, enumeration-only, and yes/no items; questions lacking two normalized endpoint entities; duplicate paraphrases; and items whose supplied literature cannot support a minimal relation chain. Two biomedical annotators independently mark the shortest sufficient ordered chain and supporting PMID spans, then adjudicate disagreements. From the resulting 1,200 questions, seed 2027 draws the fixed 200-question sample stratified as 68 one-link, 87 two-link, and 45 three-or-more-link questions.

PubMedQA. The 1,000-question expert-labeled subset is the source pool reported in Table 2. Seed 2027 permutes that pool; in order, we retain questions with a normalized biomedical entity in both the question and supporting abstract until the 94 one-link, 78 two-link, and 28 three-or-more-link strata are filled. The same annotators assign the minimal directed relation sequence supported by the abstract before an item fills a stratum. Required-link count is the number of edges in the adjudicated shortest sufficient chain; it is never supplied to a system.

B.1 DrugMechQA quality control

DrugMechQA joins each DrugBank drug–target edge to a Reactome target–pathway edge and, when present, a pathway–phenotype or disease edge supported by the fixed snapshot. The normalized path is the gold chain; its source records and aligned PubMed spans form the reference support. Questions ask for the mechanism connecting the drug to the downstream process or phenotype, and answers verbalize the complete chain. Before rendering, records with the same exact normalized full-chain key are grouped and assigned with seed 2027 to 400/200/200 train/development/test questions. This full-chain grouping prevents question variants of one complete mechanism from crossing splits; it does not impose disjointness on every constituent relation pair. Cross-split overlap is therefore measured explicitly: 2.1% for drug–target pairs and 1.8% for target–pathway pairs, and no test question shares more than 60% of its normalized entity set with a training question under Jaccard similarity. Two annotators independently review a stratified 120-question construction sample; 113 are valid and unambiguous (94.2%), with $\kappa = 0.84$. The seven rejected forms are removed and replaced by the next group-matched eligible chains before freezing the collection. Operating-point and rubric-calibration items are excluded from all fixed evaluation samples.

B.2 Operating-point selection

The calibration set contains 20 non-evaluation questions from each dataset. We search $L \in \{2, 3\}$, $K \in \{20, 40, 80\}$, $N \in \{60, 90, 120, 180\}$, $\tau \in \{0.3, 0.4, 0.5\}$, and semantic/coherence weights in 0.05 increments summing to 0.70 with $\lambda_{\text{prov}} = 0.3$. The equal-weight mean of citation correctness and coherence is the selection objective; coverage and then lower latency break ties. This procedure

selects $L = 3$, $K = 40$, $N = 120$, $\tau = 0.4$, and weights $0.35/0.35/0.30$, which are shared across datasets.

C Human Evaluation Protocol

Two annotators with biomedical literature-curation training evaluated the same fixed 200-question sample per dataset for every method. They viewed the question, one seeded answer, inline support identifiers, and cited snippets, but not the method name, condition assignment, or internal retrieval state. Evidence formats were retained because the presented verification interface is part of the evaluated output; "blinded" therefore refers to hidden labels and randomized assignment rather than format masking. A separate 20-question calibration round established the rubric and was excluded from evaluation. The three seeded outputs for every question–method pair were shown independently in randomized order. Annotators labeled each output independently and adjudicated only after both label sets were complete.

Faithfulness is annotated per factual sentence and requires every claim in the sentence to follow from displayed evidence. Citation correctness is annotated per sub-claim–citation pair and requires the cited span to state the claimed relation with compatible endpoints. Mechanistic coherence is an answer-level binary label requiring the cited relations to form the complete ordered backbone. Vacuous text and citation-only strings are not substantive answers. A localized insufficiency response is an abstention; it is omitted from fractional answered-seed averages and scored incorrect for all-question correctness.

The primary decision-level observation is all-question correctness: every question contributes one binary value, and abstention or substantive incorrectness scores zero. Coverage and escape status use the same two-of-three majority to decompose that outcome. Correctness requires the reference endpoint and required mechanism without a contradictory factual claim; escape requires a substantive answer with at least one unsupported required relation. Conditional supportability is reported separately. Any question with at least one substantive seeded answer enters that answered-output denominator, and its fractional metrics average the answered seeds only. Binary coherence is one only if at least two seeds are coherent and is zero otherwise, including when just one seed is answered; a three-seed abstention has no conditional supportability score. Unweighted Cohen's κ is computed on pre-adjudication labels at the sentence, pair, and answer levels, giving 0.78, 0.74, and 0.72, respectively. $^{*}/^{**}$ mark paired intervals excluding zero at 95%/99%.

For decision-level outcomes, one bootstrap draw resamples the same 200 question identifiers jointly for both methods and retains zero-valued abstentions. For conditional supportability, each method's answered-output mean is recomputed over its available scores in the same draw before subtraction. An asymmetric three-seed abstention remains missing only from that conditional mean; it is neither imputed nor used to delete the other method's observation. The shared resampling frame is therefore paired by question, while the answered-output estimand remains explicitly conditional on each method's substantive responses.

D Full Results and Supporting Analyses

Table 12: Macro-average answered-output supportability. Rows are equal-weight means of the corresponding per-dataset estimates in Table 13.

Method	Faithfulness ↑	Citation corr. ↑	Coherence ↑
DenseRAG	71.4	62.8	55.1
GraphRAG	75.9	67.2	60.3
ProvenanceRAG	78.3	70.6	63.7
MedCPT-RAG	80.7	74.2	66.9
Self-RAG	82.2	76.7	69.5
ColMedRAG	84.8	78.4	71.5
TypedPath-Open	85.8	80.2	73.0
MechTrace	87.2	82.4	75.2

Table 13: Per-dataset supportability on answered outputs from the same fixed 200-question samples. Brackets are 95% question-bootstrap intervals; significance marks compare MechTrace with Self-RAG.

Dataset	Method	Faithfulness	Citation corr.	Coherence
BioASQ-mech	DenseRAG	70.1 [66.7,73.4]	61.5 [57.8,65.1]	54.0 [50.3,57.6]
	GraphRAG	75.0 [71.5,78.4]	66.2 [62.6,69.7]	60.1 [56.4,63.7]
	ProvenanceRAG	77.2 [73.8,80.5]	69.4 [65.8,72.9]	63.0 [59.3,66.6]
	MedCPT-RAG	79.8 [76.3,83.2]	73.1 [69.5,76.6]	66.4 [62.7,70.0]
	Self-RAG	81.0 [77.5,84.4]	75.6 [72.0,79.1]	68.9 [65.2,72.5]
	ColMedRAG	83.8 [80.3,87.2]	77.6 [73.9,81.2]	70.9 [67.1,74.6]
	TypedPath-Open	84.6 [81.2,87.9]	79.9 [76.4,83.3]	72.6 [68.9,76.2]
	MechTrace	86.2** [82.9,89.4]	82.4** [79.0,85.7]	75.1** [71.5,78.6]
PubMedQA	DenseRAG	73.8 [70.1,77.4]	65.7 [61.9,69.4]	57.8 [53.9,61.6]
	GraphRAG	76.4 [72.8,79.9]	68.1 [64.4,71.7]	60.9 [57.1,64.6]
	ProvenanceRAG	79.1 [75.5,82.6]	72.6 [68.9,76.2]	64.3 [60.5,68.0]
	MedCPT-RAG	81.4 [77.9,84.8]	75.8 [72.2,79.3]	67.6 [63.8,71.3]
	Self-RAG	83.2 [79.7,86.6]	78.1 [74.5,81.6]	70.2 [66.4,73.9]
	ColMedRAG	85.8 [82.3,89.2]	79.1 [75.5,82.6]	71.8 [68.0,75.5]
	TypedPath-Open	85.4 [81.9,88.8]	79.5 [75.9,83.0]	71.6 [67.8,75.3]
	MechTrace	86.1* [82.7,89.4]	81.2** [77.7,84.6]	72.8* [69.1,76.4]
DrugMechQA	DenseRAG	70.2 [66.2,74.1]	61.3 [57.1,65.4]	53.6 [49.4,57.7]
	GraphRAG	76.2 [72.4,79.9]	67.4 [63.3,71.4]	60.0 [55.9,64.0]
	ProvenanceRAG	78.5 [74.7,82.2]	69.7 [65.7,73.6]	63.8 [59.7,67.8]
	MedCPT-RAG	80.9 [77.2,84.5]	73.6 [69.7,77.4]	66.7 [62.7,70.6]
	Self-RAG	82.4 [78.7,86.0]	76.5 [72.7,80.2]	69.3 [65.4,73.1]
	ColMedRAG	84.9 [81.2,88.5]	78.4 [74.5,82.2]	71.7 [67.6,75.7]
	TypedPath-Open	87.3 [83.7,90.8]	81.1 [77.3,84.8]	74.9 [70.9,78.8]
	MechTrace	89.4** [86.1,92.6]	83.7** [80.2,87.1]	77.6** [74.0,81.1]

The complete roster preserves the ordering summarized in the main text. The smallest separation is faithfulness on PubMedQA; citation correctness and coherence separate more strongly on BioASQ-mech and DrugMechQA, where more relations must remain visible.

Table 14: Answered-output supportability in the ablation study on the same 600-question sample. Each row removes one component while holding the remainder of the pipeline fixed.

Variant	Faithfulness ↑	Citation corr. ↑	Coherence ↑
MechTrace	87.2 [83.8,89.1]	82.4 [78.9,84.8]	75.2 [70.9,77.4]
w/o provenance	82.1 [79.3,84.8]	74.3 [71.1,77.4]	72.8 [69.5,76.0]
w/o path retrieval	79.4 [76.5,82.2]	71.6 [68.3,74.8]	64.7 [61.2,68.1]
w/o decomposition	80.2 [77.3,83.0]	76.1 [72.9,79.2]	66.9 [63.4,70.3]

On DrugMechQA, MechTrace achieves 78.6% gold-chain recall [75.4,81.7] and 64.3% precision [60.8,67.7], compared with Self-RAG's 66.9% recall and 51.2% precision. Missing-link errors comprise 52–57% of errors on the higher-hop datasets, while wrong-link errors comprise 22–30%.

Removing the DrugBank-attributed evidence family lowers every score, yet the configured MechTrace implementation retains a 9.1-point faithfulness, 10.3-point citation-correctness, and 10.2-point coherence advantage over the configured Self-RAG implementation. This source-sensitivity analysis complements the path-matched package comparison in Table 19.

Table 15: DrugMechQA source-family holdout with DrugBank drug–target edges and PubMed spans aligned exclusively to those edges removed from every method's candidate pool. The full-graph row exactly matches Table 13.

Method	Faithfulness	Citation corr.	Coherence
Self-RAG (holdout)	72.1 [68.5,75.6]	64.3 [60.5,68.0]	58.2 [54.3,62.0]
MechTrace (holdout)	81.2 [77.8,84.5]	74.6 [71.0,78.1]	68.4 [64.7,72.0]
MechTrace (full graph)	89.4 [86.1,92.6]	83.7 [80.2,87.1]	77.6 [74.0,81.1]

Table 16: Medians and IQRs behind Figure 2, computed over answered outputs. The final column records the fixed sampled-question frame.

Dataset	Method	Median	IQR	Sampled Q
BioASQ-mech	Self-RAG	80.8	14.6	200
BioASQ-mech	MechTrace	86.0	10.9	200
PubMedQA	Self-RAG	83.0	13.2	200
PubMedQA	MechTrace	85.9	11.4	200
DrugMechQA	Self-RAG	82.1	15.3	200
DrugMechQA	MechTrace	89.1	9.6	200

Table 17: Sensitivity to retained weave size. The 120-node row intentionally duplicates the main MechTrace operating point. Latency measures wall-clock retrieval time on the dual-EPYC server, with indexing and LLM decoding excluded.

Max nodes	Faith. ↑	Cit. corr. ↑	Coh. ↑	Latency (ms) ↓
60	83.2 [80.4,85.9]	78.1 [75.0,81.1]	70.5 [67.1,73.8]	108
90	85.1 [82.4,87.7]	80.2 [77.1,83.2]	72.9 [69.5,76.2]	142
120	87.2 [83.8,89.1]	82.4 [78.9,84.8]	75.2 [70.9,77.4]	185
180	86.7 [84.0,89.3]	81.8 [78.8,84.7]	74.0 [70.7,77.2]	268

Increasing the node cap helps through 120 nodes and then slightly degrades all three supportability measures while adding latency. The selected cap is therefore the observed accuracy–cost turning point rather than the largest tested weave.

Table 18: Descriptive minimum typed-chain length after normalization. The derivation differs for DrugMechQA and the two annotation-side datasets, so these values provide context rather than a shared causal axis.

Dataset	Indicative min. links	Derivation
BioASQ-mech	2.3	annotation-side minimal support chains
PubMedQA	1.7	annotation-side minimal support chains
DrugMechQA	2.6	template-derived support chains

These averages motivate the within-dataset bins below but are not pooled as a cross-dataset causal variable because DrugMechQA chains originate from its construction graph.

E Complete Contract Deltas and Pipeline

E.1 TypedPath-Open token matching

When MechTrace constructs a weave, the common serializer records B_{path}, the token count of its ordered typed triples and provenance fields, and B_{text}, the token count of its packed snippets,

with $B_{\text{path}} + B_{\text{text}} \leq 3{,}000$. TypedPath-Open receives the identical triple block and exactly B_{text} text-token slots. If MechTrace returns a relation-localized abstention before a weave is available, a shared fallback serializer packs qualified path triples alone in the same score order under N and the 3,000-token cap; that block defines B_{path}, and $B_{\text{text}} = 3{,}000 - B_{\text{path}}$. MechTrace still abstains, while TypedPath-Open fills the fallback text allotment. It ranks the shared PubMed pool by question–passage SapBERT similarity, removes duplicates, and retains passages in rank order irrespective of path membership. The last included passage is the longest complete-sentence prefix fitting B_{text}, and unused remainder stays empty. Entity and response-format validation remain shared; bundle-specific closure and post-decoding relation checks remain disabled.

This section expands the primary table to all three supportability metrics. Paired intervals quantify the complete downstream interface package against the path-matched TypedPath-Open intervention and separately measure the gap from the strongest passage-side comparator.

Table 19: Paired MechTrace improvement over the path-matched interface and late-interaction comparators. Each cell reports a point delta and 95% paired bootstrap interval.

Comparator	Dataset	Δ Faithfulness	Δ Citation corr.	Δ Coherence
TypedPath-Open	BioASQ-mech	+1.6 [+0.4,+2.8]	+2.5 [+0.9,+4.1]	+2.5 [+0.8,+4.2]
	PubMedQA	+0.7 [-0.3,+1.7]	+1.7 [+0.3,+3.1]	+1.2 [-0.1,+2.5]
	DrugMechQA	+2.1 [+0.6,+3.6]	+2.6 [+0.8,+4.3]	+2.7 [+0.9,+4.5]
ColMedRAG	BioASQ-mech	+2.4 [+1.1,+3.8]	+4.8 [+3.0,+6.6]	+4.2 [+2.2,+6.2]
	PubMedQA	+0.3 [-0.7,+1.3]	+2.1 [+0.5,+3.7]	+1.0 [-0.4,+2.4]
	DrugMechQA	+4.5 [+2.8,+6.2]	+5.3 [+3.3,+7.3]	+5.9 [+3.7,+8.1]

All six citation-correctness deltas are positive with intervals above zero. Coherence is likewise positive, with the two smallest lower bounds occurring in the lower-link PubMedQA setting.

Algorithm 1 makes the package boundary executable: path scoring proposes candidates, connected binding enforces relation coverage, packing preserves required witnesses, and the final branch validates generated relations against legal pair IDs.

Algorithm 1 MechTrace contract execution.

Require: Question q, graph G, passage index $\mathcal{I}$, limits L, K, N, B, threshold τ
Ensure: Answer a with validated inline citations, or a relation-localized abstention

$R, A \leftarrow \textsc{Decompose}(q)$
$V_0 \leftarrow \textsc{LinkEntities}(A, G)$
if a required entity in A is unresolved **then**
  return $\textsc{Abstain}(\textsc{UnresolvedEntity})$
end if
$\Pi \leftarrow \textsc{EnumerateTypedPaths}(G, V_0, R, L, K)$
for each $\pi \in \Pi$ **do**
  $(S_\pi, \phi_\pi) \leftarrow \textsc{AttachEligibleEdgeSnippets}(\pi, q, \mathcal{I})$
  compute $u_{\text{sem}}(\pi), u_{\text{coh}}(\pi), u_{\text{prov}}(\pi)$, assigning 0 semantic support to each unbound edge
end for
min–max normalize components; score and deterministically sort Π
$\Pi^+ \leftarrow \{\pi \in \Pi : s(\pi \mid q) \geq \tau\}$
$\mathcal{B} \leftarrow \textsc{ConnectedBindings}(R, \Pi^+, \phi)$
if $\mathcal{B} = \varnothing$ **then**
  return $\textsc{Abstain}(\textsc{FirstUnboundRelation}(R, \Pi^+))$
end if
$b^* \leftarrow \arg\max_{b \in \mathcal{B}} \sum_{(\pi,e,s) \in b} s(\pi \mid q)$
$W \leftarrow \textsc{PackWeave}(\Pi^+, b^*, N, B)$
if W omits any required witness in b^* **then**
  return $\textsc{Abstain}(\textsc{FirstOverflowRelation}(b^*))$
end if
$I_W \leftarrow \{\textsc{PairId}(e, s) : e \in E_W,\ s \in \phi(e)\}$
$a \leftarrow \textsc{Generate}(q, \textsc{Serialize}(W), I_W)$
$F \leftarrow \textsc{InvalidCitations}(a, I_W) \cup \textsc{UnsupportedRelations}(a, W, \phi)$
if $F \neq \varnothing$ **then**
  return $\textsc{Abstain}(\textsc{FirstFailedRelation}(F))$
end if
return (a, W)

The witness-selection and packing steps implement support closure and protect required evidence from pruning. Legal-ID construction implements citation closure, while the final validation branch turns a generation-time violation into an observable relation-specific outcome. The global path threshold filters candidates but never substitutes for these per-relation checks.

F Answer Behavior by Dataset

Raw counts separate evidence quality among answered questions from the decision to abstain. This is essential for the main behavioral claim: the correctness increase cannot be explained by answering fewer questions.

Across datasets, MechTrace gives three fewer substantive answers than TypedPath-Open (544 versus 547) but 29 more correct answers and 28 fewer unsupported escapes. DrugMechQA contributes the largest correctness and escape differences, matching its explicit multi-relation construction.

G Generator-Family Transfer

Generator transfer tests whether the retained-object effect depends on one decoder's citation behavior. All retrieval states, question samples, prompts, and budgets remain fixed when Llama is replaced by Qwen.

Relative to the Llama-3.1-70B-Instruct contract margin, the Qwen interaction in citation correctness is +0.2, -0.1, and +0.8 points. The corresponding coherence interactions are +0.3, +0.5, and +0.2, showing stable direction without identical margins.

Table 20: Raw answer behavior by dataset. Percentages use 200 questions per dataset; conditional correctness uses answered questions as the denominator.

Dataset	Method	Answers	Abstains	All-Q correct	Cond. correct	Escapes
BioASQ	Self-RAG	184/200 (92.0%)	16/200 (8.0%)	108/200 (54.0%)	108/184 (58.7%)	31
	ColMedRAG	186/200 (93.0%)	14/200 (7.0%)	114/200 (57.0%)	114/186 (61.3%)	27
	TypedPath-Open	182/200 (91.0%)	18/200 (9.0%)	116/200 (58.0%)	116/182 (63.7%)	22
	MechTrace	181/200 (90.5%)	19/200 (9.5%)	124/200 (62.0%)	124/181 (68.5%)	14
PubMedQA	Self-RAG	188/200 (94.0%)	12/200 (6.0%)	118/200 (59.0%)	118/188 (62.8%)	24
	ColMedRAG	190/200 (95.0%)	10/200 (5.0%)	121/200 (60.5%)	121/190 (63.7%)	22
	TypedPath-Open	187/200 (93.5%)	13/200 (6.5%)	122/200 (61.0%)	122/187 (65.2%)	20
	MechTrace	186/200 (93.0%)	14/200 (7.0%)	126/200 (63.0%)	126/186 (67.7%)	16
DrugMechQA	Self-RAG	176/200 (88.0%)	24/200 (12.0%)	102/200 (51.0%)	102/176 (58.0%)	38
	ColMedRAG	180/200 (90.0%)	20/200 (10.0%)	110/200 (55.0%)	110/180 (61.1%)	31
	TypedPath-Open	178/200 (89.0%)	22/200 (11.0%)	111/200 (55.5%)	111/178 (62.4%)	27
	MechTrace	177/200 (88.5%)	23/200 (11.5%)	128/200 (64.0%)	128/177 (72.3%)	11

Table 21: Qwen2.5-72B-Instruct transfer under the matched snapshot, evidence budget, answer template, question samples, and three-seed aggregation.

Dataset	Method	Faith.	Cit. corr.	Coh.	All-Q corr.	Coverage
BioASQ	Self-RAG	79.4	73.8	66.8	51.0	91.0
	ColMedRAG	82.1	75.9	68.7	54.5	92.0
	TypedPath-Open	83.1	77.9	70.5	55.5	90.5
	MechTrace	84.8	80.6	73.3	59.5	89.5
PubMedQA	Self-RAG	81.6	76.4	68.1	56.5	93.0
	ColMedRAG	84.2	77.5	69.7	59.0	94.0
	TypedPath-Open	83.8	78.1	69.3	59.5	92.5
	MechTrace	84.9	79.7	71.0	61.0	92.0
DrugMechQA	Self-RAG	80.7	74.5	67.0	48.5	86.5
	ColMedRAG	83.2	76.8	69.1	52.0	89.0
	TypedPath-Open	85.1	79.0	72.1	53.5	87.0
	MechTrace	87.7	82.4	75.0	59.5	86.5

H WITHIN-DATASET LINK DEMAND

The annotation-side bins test a mechanism-specific prediction: opening the evidence bundle should become more damaging as more relation witnesses must survive generation and citation.

Table 22: MechTrace minus TypedPath-Open under Qwen2.5-72B-Instruct. Cells report paired deltas with 95% bootstrap intervals.

Dataset	Δ Citation correctness	Δ Coherence
BioASQ-mech	+2.7 [+0.9,+4.5]	+2.8 [+0.8,+4.8]
PubMedQA	+1.6 [+0.2,+3.0]	+1.7 [+0.1,+3.3]
DrugMechQA	+3.4 [+1.4,+5.4]	+2.9 [+0.8,+5.0]

Table 23: Question counts and MechTrace deltas by annotation-side minimum link demand. Delta columns compare against TypedPath-Open (TP) and ColMedRAG (LI).

Dataset	Links	n	TP Δ Cit.	TP Δ Coh.	LI Δ Cit.	LI Δ Coh.
BioASQ	1	68	+0.8	+1.1	+2.2	+2.3
	2	87	+2.4	+2.2	+5.0	+4.8
	3+	45	+4.6	+4.2	+7.1	+7.3
PubMedQA	1	94	+0.4	+0.6	+1.1	+0.9
	2	78	+1.6	+1.0	+2.2	+1.9
	3+	28	+2.9	+2.2	+3.8	+3.1

For both datasets, the MechTrace advantage over each comparator grows monotonically from one to three-or-more links. The positive BioASQ-mech trends for citation correctness and coherence have intervals above zero; PubMedQA shows the same ordering with a weaker coherence trend.

Trend estimator. For each dataset, comparator, and metric, we fit available question–method scores with method, ordinal link-demand code $0, 1, 2$, and their interaction; the interaction is the MechTrace-minus-comparator trend in points per bin. All 200 question identifiers are jointly resampled, and a method-specific missing score remains missing without imputation or complete-case deletion. A 1,000-resample paired question bootstrap supplies the interaction standard error and 95% interval; the two-sided p-value uses the corresponding normal approximation.

Table 24: Paired link-demand interactions. Each cell reports the per-bin trend, 95% interval, and p-value.

Comparator	Dataset	Citation correctness	Coherence
TypedPath-Open	BioASQ-mech	+1.85 [+0.42,+3.28], $p = 0.012$	+1.54 [+0.18,+2.90], $p = 0.027$
TypedPath-Open	PubMedQA	+0.91 [+0.01,+1.81], $p = 0.048$	+0.74 [-0.15,+1.63], $p = 0.103$
ColMedRAG	BioASQ-mech	+2.43 [+0.84,+4.02], $p = 0.003$	+2.51 [+0.76,+4.26], $p = 0.005$
ColMedRAG	PubMedQA	+1.29 [+0.20,+2.38], $p = 0.021$	+1.12 [+0.06,+2.18], $p = 0.039$

I Expert Verification Usefulness

Across all 150 questions, evaluators select MechTrace 87 times, TypedPath-Open 43 times, and tie 20 times. Conditioning on the 130 non-ties gives a MechTrace preference of 87 cases (66.9%), with Wilson 95% interval [58.4%,74.4%]. Pre-adjudication agreement is $\kappa = 0.71$.

J Removed-Edge Stress Test

The stress set contains 200 mechanistic questions with an adjudicated required bridge and plausible evidence for both endpoint entities. For each question, the intervention removes that bridge and its

Table 25: Preference for verifying a mechanism from the answer and cited evidence. Each cell gives count and percentage among 50 questions; the final column conditions on non-ties.

Dataset	MechTrace	TypedPath-Open	Tie	MechTrace among non-ties
BioASQ-mech	31 (62%)	14 (28%)	5 (10%)	68.9%
PubMedQA	22 (44%)	18 (36%)	10 (20%)	55.0%
DrugMechQA	34 (68%)	11 (22%)	5 (10%)	75.6%

aligned snippets while retaining endpoint passages; the paired intact control restores the unmodified evidence. Annotators blinded to condition labels mark detection when the intervened output explicitly names the held-out relation as unsupported. The signal can originate from MechTrace's validator or from a comparator's generated insufficiency statement; the shared label measures observable detection across interfaces. Escape is the fraction that instead answers with that relation unsupported. False abstention and correctness are measured on all 200 intact counterparts, respectively recording spurious insufficiency signals and ordinary answer utility when a complete witness chain is present. These outcomes jointly measure omission detection, specificity, retained utility, and retrieval-side cost.

Table 26: Held-out missing-edge stress test on 200 mechanistic questions. Detect and escape use intervened evidence; false abstention and correctness use paired intact controls. Binary cells give rate, Wilson 95% interval, and raw count; latency uses question-bootstrap intervals.

Method	Detect ↑	Escape ↓	False abstain ↓	Intact correct ↑	Latency ↓
Self-RAG	30.5 [24.5,37.2] (61)	33.5 [27.3,40.3] (67)	2.5 [1.1,5.7] (5)	55.0 [48.1,61.8] (110)	116 [107,126]
ColMedRAG	36.0 [29.7,42.9] (72)	29.0 [23.1,35.6] (58)	3.0 [1.4,6.4] (6)	59.5 [52.6,66.1] (119)	154 [143,166]
TypedPath-Open	49.0 [42.2,55.9] (98)	23.5 [18.2,30.0] (47)	4.0 [2.0,7.9] (8)	57.0 [50.1,63.7] (114)	178 [166,191]
MechTrace	78.0 [71.8,83.2] (156)	9.0 [5.8,13.8] (18)	6.0 [3.3,10.2] (12)	61.0 [54.1,67.5] (122)	185 [172,199]

Removing one bridge separates a relation-aware insufficiency decision from generic reluctance to answer. MechTrace detects 156 removals while producing 18 escapes; its 12 false abstentions on intact controls quantify the corresponding specificity cost. TypedPath-Open detects 98 and produces 47 escapes, showing that access to identical path construction does not itself preserve the missing relation through the answer interface.

K MECHANISTIC TRACE CASE STUDY

Figure 4 traces the imatinib example at the contract boundary. Each citation identifies an edge–snippet pair; removing the BCR–ABL1-to-pathway pair breaks the legal support chain and activates the localized abstention branch.

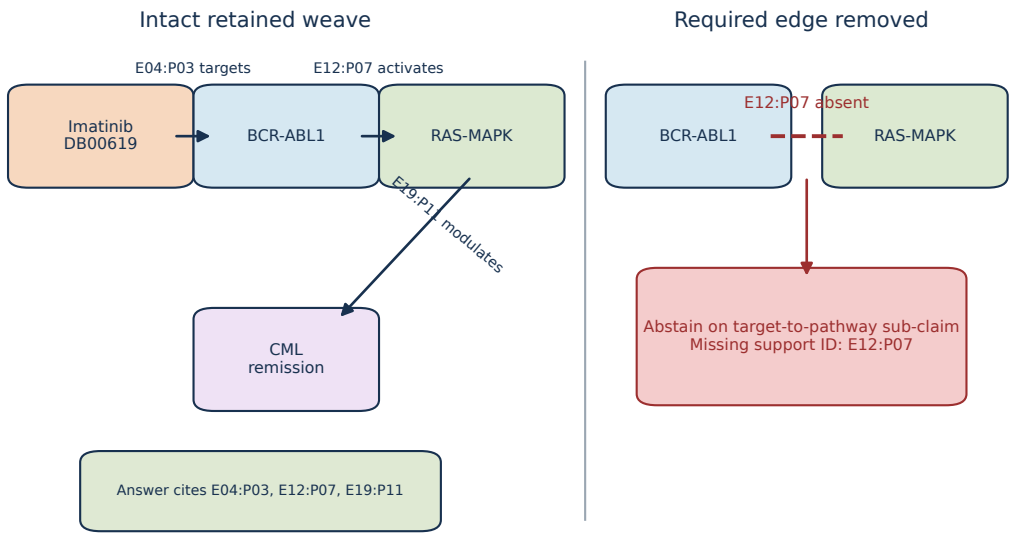

Figure 4: Auditable support trace for "How does imatinib treat CML?" The intact weave exposes each required relation and edge–snippet identifier; the removal branch identifies the missing target-to-pathway support.