IMMUNOGRAPHBENCH: A LARGE-SCALE CONTEXT GRAPH BENCHMARK FOR SCIENTIFIC REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Biomedical knowledge is recorded as conditional claims: a relation between two entities is observed under particular conditions, supported by specific evidence, and associated with particular measurements. A graph interface that exposes only context-free triples can make distinct candidate hypotheses observationally equivalent: they satisfy the same topology, while the observation that distinguishes them is absent. This is an identifiability problem at the reasoning interface, not merely a reduction in text. ImmunoGraphBench makes the problem testable by constructing structurally underdetermined questions where relation-level context breaks the symmetry. We build a biomedical KG of 1.62M triples, 75% of which carry textual evidence or structured metadata, and vary structure and context independently on the same QA instances. Across seven LLMs, adding context to explicit candidate paths raises Medium-level accuracy from 61.35% to 89.56% (+28.20 pp), and the gain persists when the graph is retrieved from the question alone (+14.83 pp). Controls show that the gain depends on relation-matched content, survives evidence-count debiasing, and partially compensates for missing structural support. The results identify relation-level context as the interface signal that restores candidate identifiability beyond topology.

1 INTRODUCTION

Biomedical knowledge is not a collection of context-free facts but a collection of conditional claims. A statement such as "gene g is downregulated in cell type c" is made under a particular condition, supported by particular evidence, and quantified by a particular effect size, and the same entity pair can carry apparently contradictory findings under different conditions (Rosemblat et al., 2019). Immunology is a canonical case: interleukin-6 acts as an anti-inflammatory or a pro-inflammatory signal depending on whether it signals through its membrane-bound or its soluble receptor (Scheller et al., 2011), and many immune findings hold only for a specific cell state, tissue, or stimulation. Reasoning over such knowledge therefore requires identifying which conditional claim is supported, not only which entities are connected; an interface that exposes only edges can leave competing claims observationally indistinguishable.

Knowledge representations for biology have long recognized this. The Biological Expression Language (BEL) attaches citations, evidence text, and experimental context to every statement (OpenBEL Consortium, 2014); INDRA assembles mechanisms together with their supporting evidence from text mining and curated databases (Gyori et al., 2017; Bachman et al., 2023); and SemMedDB links each predication to its source sentence (Kilicoglu et al., 2012). These representations preserve the observations that make a biological claim interpretable. In contrast, the integrated KGs used to ground LLMs, such as Hetionet (Himmelstein et al., 2017), PrimeKG (Chandak et al., 2023), and SPOKE (Morris et al., 2023), commonly present knowledge to the model as triples (h, r, t), and KG-augmented LLM pipelines often serialize these triples into the prompt (Soman et al., 2024; Zhang et al., 2026). The resulting question is therefore not simply how much information is removed, but whether the remaining observation still identifies the correct candidate: *when topology makes competing biomedical hypotheses observationally equivalent, can relation-level context restore their identifiability?*

Testing identifiability requires varying graph structure and edge-aligned context independently on the same instances. Existing benchmarks confound these factors (Section 2), so they do not isolate when topology leaves candidates tied despite names, order, or priors. To address this, we construct *structurally underdetermined* questions in which several candidate paths satisfy the same topology, while edge conditions identify the answer.

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.

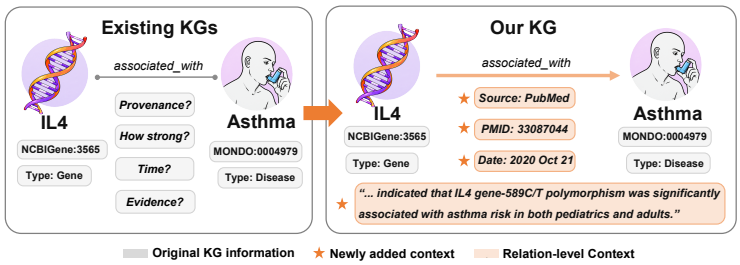


Figure 1: A triple-only KG keeps which entities are related; a context-rich KG also keeps the evidence, conditions, and measurements under which each relation was observed.

For example, FER and ZNF76 share the same structural route from Pegivirus C through inhibition and chromatin, but only the FER record matches the fold change and adjusted p-value in the question (Appendix A.10.2). The triple projection makes the two routes structurally equivalent; relation-level measurements break that equivalence. ImmunoGraphBench implements this test with a 1.62M-triple KG (75% contextualized), three QA levels, and an input-visibility protocol that toggles context on fixed unordered triples or candidate paths, enabling paired comparisons with structure held fixed.

We use the benchmark to ask three questions. **RQ1:** Does relation-level context restore candidate identifiability beyond structure? **RQ2:** Which observations provide the distinguishing signal? **RQ3:** When does that signal matter, and is it used robustly? Across seven LLMs, the answers form a closed behavioral loop. *(1) Context resolves structural ambiguity.* On Medium-level questions, adding context to explicit candidate paths raises accuracy from 61.35% to 89.56%, and the gain persists when the graph is retrieved from the question alone (+14.83 pp). *(2) The signal is relation-specific.* Once structure is given, the largest gains come from structured measurements and empirical relations. *(3) The effect follows the construction.* Context matters most when structure constrains least and survives controls for instance specificity, context quantity, and missing structure.

Our contributions are threefold:

- **An interface-level identifiability test.** We formalize structural equivalence, define the context-based compatibility condition that can distinguish candidates, and estimate the evaluated model's response with paired comparisons across seven LLMs.
- **A context-bearing benchmark resource.** We release a million-scale biomedical KG whose normalized relations carry evidence, provenance, and structured metadata, with a three-level benchmark that measures this information under controlled visibility.
- **Evidence on what is recovered.** Controls for length (three models), instance specificity, context quantity, and missing structure show that the recovered information is relation-matched and complements graph structure rather than restating it.

2 Related Work

Evidence-bearing knowledge representations. The first line of work preserves the information that makes a biological statement interpretable. BEL (OpenBEL Consortium, 2014), INDRA (Gyori et al., 2017; Bachman et al., 2023), and SemMedDB (Kilicoglu et al., 2012) keep evidence, provenance, or experimental context with a statement; hyper-relational KGs such as Wikidata (Vrandečić & Krötzsch, 2014) represent qualifier key–value pairs (Rosso et al., 2020; Galkin et al., 2020). These systems establish that context is representable, but do not measure what an LLM gains from seeing it during graph reasoning. ImmunoGraphBench attaches evidence, provenance, and metadata to the same normalized edge so this channel can be varied independently.

Integrated biomedical KGs and their LLM consumers. The second line supplies graph structure for biomedical reasoning. Hetionet (Himmelstein et al., 2017), PrimeKG (Chandak et al., 2023), and SPOKE (Morris et al., 2023) provide heterogeneous backbones, while later work adds LLM construction, immune annotations, or retrieval (Zhang et al., 2026; Rjoob et al., 2026; He et al., 2026; Feng et al., 2025; Xu et al., 2025). KG-augmented methods retrieve or traverse subgraphs (Sun et al., 2024; Luo et al., 2024; Edge et al., 2024), including biomedical variants (Soman et al., 2024; Xiong et al., 2024), and commonly serialize them as triples. The comparison in Table 1 therefore concerns which relation-level observations survive this interface, rather than graph coverage alone.

Benchmarks and controls for measuring the cost. The third line evaluates question answering, retrieval, or multi-hop reasoning under a fixed representation: BioASQ (Tsatsaronis et al., 2015),

PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021), STaRK (Wu et al., 2024), RiTeK (Huang et al., 2024), BioMedHop (Tan et al., 2026), and AdaPath (Kim et al., 2026). These benchmarks establish end-to-end competence but do not separately identify graph structure and relation-level context. ImmunoGraphBench makes that comparison explicit by toggling edge-aligned context on fixed structures, with controls for length, mismatched content, context volume, and missing structure (Shi et al., 2023; Longpre et al., 2021; Xie et al., 2024; Gururangan et al., 2018). Table 1 locates this intervention at the relation level.

Table 1: What representative graphs expose at each level. Rows are the integrated KGs above, with CoDEx-L as a general-domain reference; the three right-hand columns are the relation-level channels those lines discuss, attached to relations rather than to nodes. Sources: CoDEx-L (Safavi & Koutra, 2020), PrimeKG (Chandak et al., 2023), BioStrataKG (Feng et al., 2025), CardioKG (Rjoob et al., 2026), and ICKG (He et al., 2026). Scale is in millions of triples; ✓ marks explicit support, and – marks the absence of relation-level context at that layer.

	KG			Relation-level context		
KG	**Domain**	**Scale**	**Entity Attributes**	**Structured Metadata**	**Textual Evidence**	**Relation-level Provenance**
CoDEx-L	General	0.61M	✓	–	–	–
PrimeKG	Biomedical	4.05M	✓	–	–	✓
BioStrataKG	Biomedical	0.29M	–	–	–	✓
CardioKG	Biomedical	1.20M	✓	✓	–	–
ICKG	Biomedical	0.12M	–	✓	–	✓
Ours	**Biomedical**	**1.62M**	✓	✓	✓	✓

3 IMMUNOGRAPHBENCH

3.1 CONTEXT-RICH KNOWLEDGE GRAPH

Measuring the cost of flattening places two requirements on the graph: relations have to be comparable across sources, and each relation has to retain the conditions under which it was reported. The KG is built in two stages to meet both (Figure 2): a canonical backbone is integrated from heterogeneous resources, and relation-level context is then attached to 74.98% of edges.

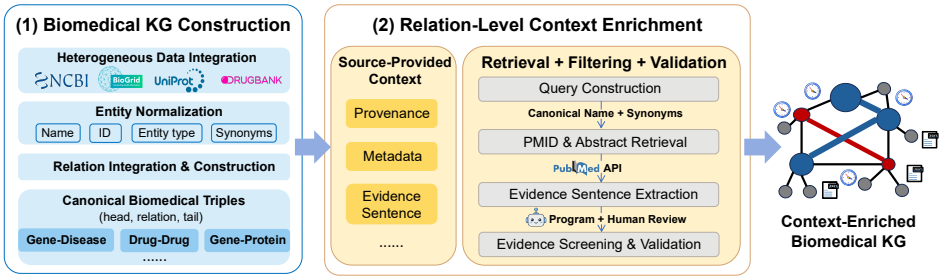


Figure 2: Construction pipeline of the context-rich biomedical knowledge graph.

Graph backbone. Comparability across sources requires one vocabulary for entities and relations. Entity dictionaries assign each entity a canonical name, an authoritative identifier, an entity type, and a synonym set, and relation names and directions are standardized, so that each retained relation becomes a canonical triple. Relations are drawn from NCBI Gene (Brown et al., 2015), BioGRID (Stark et al., 2006), UniProt (The UniProt Consortium, 2025), and DrugBank (Knox et al., 2024), together with ontologies such as the Cell Ontology (Diehl et al., 2016) and Uberon (Mungall et al., 2012), covering cells, genes, proteins, diseases, drugs, anatomy, and biological functions.

Relation-level context. A normalized triple fixes which entities are related but not the evidence, provenance, or conditions under which the relation was reported. We therefore attach to each triple $e = (h, r, t)$ a relation-level context

$$\phi(e) = \mathcal{C}_{h,r,t} = \left(\mathcal{C}^{\text{evi}}_{h,r,t}, \mathcal{C}^{\text{meta}}_{h,r,t}\right), \qquad \mathcal{C}^{\text{evi}}_{h,r,t} = \{(s_i, p_i)\}_{i=1}^{N_e}, \qquad \mathcal{C}^{\text{meta}}_{h,r,t} = \{(k_j, v_j)\}_{j=1}^{N_m},$$

where s_i is an evidence sentence with provenance p_i, and (k_j, v_j) is a structured metadata attribute such as an assay, a cell context, a fold change, or an adjusted p-value. $\mathcal{C}^{\text{evi}}$ records *what was reported*

and where, and $\mathcal{C}^{\text{meta}}$ records *under which conditions and with which effect size*; these fields carry the report-specific information that a triple omits. For relations lacking sufficient textual evidence, additional evidence is retrieved from PubMed abstracts, filtered programmatically, verified by an LLM at the sentence level, and audited by stratified human review (Appendix A.2.1).

KG statistics. The final KG contains 1,618,645 unique normalized triples over 676,279 entities, spanning 15 entity types and 299 relation types, and 1,213,661 (74.98%) of them carry source-provided or literature-derived context, which gives the context layer enough coverage to sample structurally underdetermined questions at scale, including the Medium instances used below.

Structural ambiguity is a property of the graph, not of curation. The Medium level rests on the claim that several graph paths can satisfy one question, so we test that claim on the full graph independently of the benchmark pool. We draw 30,000 answerable structural queries uniformly across three motif families (10,000 each) and count the legal intermediate candidates and the distinct final answers each admits. At least two legal intermediates occur for 3.75% of branching, 0.91% of convergent, and 6.93% of upstream–downstream queries, an equal-family mean of 3.86%, and at least two distinct final answers occur for 75.05% of queries; both conditions hold together for 1,146 of 30,000 queries (3.82%). Structural underdetermination therefore arises in the graph itself, at a rate we quantify rather than assume.

3.2 Structurally Underdetermined Questions

The benchmark has three levels that progressively change what a model must determine. **Easy** questions probe whether a single relation and its context can be read correctly; **Medium** questions make topology alone insufficient by giving several candidates that satisfy the same structure; and **Hard** questions perturb the context or structure of Medium questions while keeping each question and its answer fixed, so any accuracy change is attributable to the perturbation.

Easy-level QA establishes whether a model can read an individual relation and use its attached context, before any multi-hop ambiguity is introduced. *Triple-structure* questions mask one endpoint of a triple and ask for the tail or the head entity. *Contextual* questions require the relation-level context itself: Evidence-Grounded questions ask for a reported condition (e.g., a treatment, stimulation, or patient subgroup) or for a metadata lookup or comparison, and Evidence Verification questions ask whether the given evidence supports the stated directed relation.

Medium-level QA operationalizes structural underdetermination: the topology must admit at least two candidates, and only the context attached to the edges may separate the gold candidate from the rest. For each of three subtypes—convergent, divergent, and hidden-intermediate, in which the intermediate entity is left unnamed (Figure 3)—we predefine topological templates that share a reasoning pattern but vary in branch number, path length, and relation types. Given such a template τ with a designated candidate position v_c and a candidate subgraph $G_q = (V_q, E_q)$, a candidate is any entity that can occupy v_c in a structural match of τ into G_q:

$$\mathcal{C}_\tau(G_q) = \{C_i \in V_q \mid \exists f : \tau \hookrightarrow G_q,\ f(v_c) = C_i\}\,, \tag{1}$$

where f is a match of τ in G_q that preserves the specified relation directions and structural constraints. Each candidate C_i must extend to a unique answer entity A_i through the same answer-relation type, and different candidates must reach different answers. We retain a candidate group only if $|\mathcal{C}_\tau(G_q)| \geq 2$, i.e., only if the topology alone leaves the answer undetermined. The route through a candidate is its path; the final edge from the candidate to its answer is the *answer-extension edge*, and the remaining edges are *structural-support edges*.

Because topology alone leaves several candidates, the intended decision uses relation-level context to distinguish them. We therefore select as gold a candidate whose context states a condition that separates it from the competing paths, and discard groups in which no candidate has such a condition. The question is generated from that condition and audited. Because topology alone does not identify the answer, and each gold instance is shown with and without the context on the same candidate structure, the paired contrast isolates whether a model *uses* relation-level context to resolve a structural ambiguity.

A formal lens on the ambiguity. Let $H_q = \{h_1, \ldots, h_k\}$ be the candidate paths for question q, let $S(h)$ denote a topology-only signature, and let $Z(h) = \phi(h)$ denote its edge-aligned context. The structural equivalence class is

$$[h]_S = \{h' \in H_q : S(h') = S(h)\}. \tag{2}$$

The triple interface exposes S but not Z; entity names, priors, and order may still provide non-topological tie-breakers. Three elementary propositions make the experimental estimand explicit

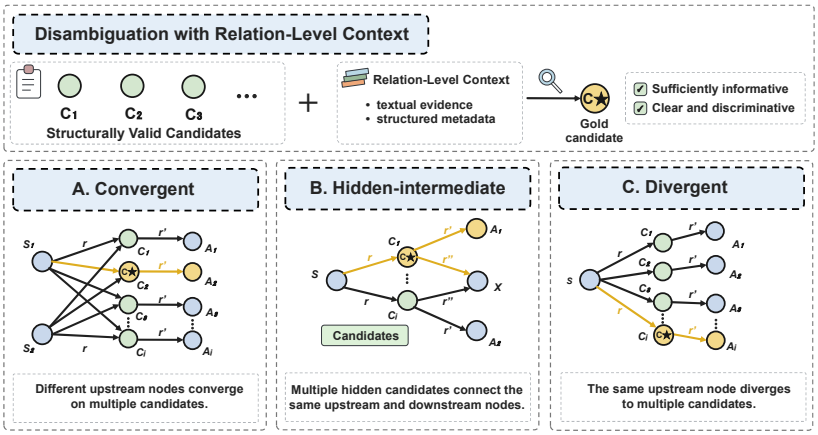


Figure 3: Structural subtypes of Medium-level QA.

(full derivations: Appendix A.1.1). **Proposition 1** is a structural-only reference: if the gold path is uniform in a class of size k_q, any rule observing only (q, S) has accuracy at most $1/k_q$; with prior π_q, the bound is $\max_h \pi_q(h)$. It is not an upper bound on the literal T/P prompts, which retain names and order. **Proposition 2** states that if a compatibility score satisfies

$$m(q, Z(h^\star)) > m(q, Z(h)) \quad \forall h \in [h^\star]_S \setminus \{h^\star\}, \tag{3}$$

then $\arg\max_h m(q, Z(h))$ identifies the gold path; and (iii) for fixed structural input $X \in \{T, P\}$, $\Delta_X = \mathrm{Acc}(X + C) - \mathrm{Acc}(X)$ is the paired empirical response to revealing Z, not a proof of Bayes-optimal inference. These facts explain why the benchmark constructs equivalent topologies, exposes edge-aligned context, and measures the resulting change while testing residual tie-breakers. We control three additional cues during construction. Context is attached to *every* candidate path, so its presence does not reveal the answer. The gold path is never marked, and the settings with and without context expose the same candidate edge set. Every question is audited for mentions of the gold intermediate or the final answer under any alias, and leaking questions are rewritten or discarded, so the answer cannot be read off by string matching. Section 4.2 shows that neither the question alone nor the context alone reaches the accuracy of the combination, so the paired contrast is the quantity of interest.

Hard-level QA applies controlled perturbations to Medium-level instances while preserving the question, candidates, and gold answer; each variant targets one alternative explanation of the context gain (details in Appendix A.3.1). **Context Swap** replaces the context of one eligible gold-path edge with that of another edge sharing its head, relation, and tail types, testing whether models use instance-specific content or merely plausible, type-compatible text. **Context-Quantity Debiasing** removes any advantage of the gold path in the number of informative evidence items, testing whether models simply prefer the path with the most context. **Structural Support Reduction** removes one structural-support edge from the gold path or a competing path while retaining its context as a textual observation, testing whether context can compensate for an incomplete graph.

4 Experiments

The experiments follow the formal argument in Section 3.2. RQ1 tests whether adding Z restores the distinction lost when candidates share S; RQ2 identifies which components of Z supply that signal; and RQ3 tests whether the signal remains useful when structural ambiguity, surface shortcuts, or graph completeness change. All experiments share the input-visibility protocol below.

4.1 Input-Visibility Protocol

The protocol isolates relation-level context by holding the question and the candidate structure fixed while varying what is visible. For each QA instance, Q, T, P, and C denote its question, its candidate subgraph as unordered triples, the same subgraph organized as candidate reasoning paths, and the associated relation-level context. A setting is named by the components shown alongside the question, which is always visible, giving the six standard settings Q, C, T, P, $T+C$, and $P+C$ (Table 6 in Appendix A.4). Here C is a slot whose contents depend on the experimental condition: in the standard setting it holds the original context $\phi(e)$ of every displayed edge e, attached to both

Table 2: Performance (%) and context gains (pp) on Easy and Medium-level QA.

Setting		DeepSeek v4-flash	GLM 5.2	Llama 3.1-8B	MiniMax M3	Qwen 3.5-9B	Qwen 3.6-27B	Qwen 3.8-27B	Average
Easy-level QA	Q	43.88	39.59	44.67	38.74	42.65	44.83	36.71	41.58
	C	64.20	66.19	60.84	63.69	67.42	66.05	67.70	65.16
	T	59.44	63.36	50.05	60.44	58.73	66.27	59.80	59.73
	P	59.47	62.53	63.00	60.33	57.95	68.86	64.60	62.39
	$T + C$	77.74	87.62	76.67	83.60	82.41	87.58	92.65	84.04
	$P + C$	76.74	83.37	75.65	82.80	82.39	85.17	91.05	82.45
Medium-level QA	Q	10.50	9.65	4.25	9.16	5.22	6.85	6.46	7.44
	C	34.89	34.09	29.84	33.98	29.68	34.01	33.89	32.91
	T	62.39	63.10	52.23	60.92	53.85	62.04	62.01	59.51
	P	66.08	65.18	55.13	62.20	55.03	62.56	63.29	61.35
	$T + C$	83.34	89.84	62.96	86.85	71.03	89.59	91.87	82.21
	$P + C$	91.82	94.28	76.99	90.67	84.25	93.65	95.23	89.56
Easy-level QA	Δ_T	18.30	24.26	26.62	23.16	23.68	21.31	32.85	**24.31**
	Δ_P	17.27	20.84	12.65	22.47	24.44	16.31	26.45	**20.06**
Medium-level QA	Δ_T	20.95	26.74	10.73	25.93	17.18	27.55	29.86	**22.71**
	Δ_P	25.74	29.10	21.86	28.47	29.22	31.09	31.94	**28.20**

gold and competing paths, whereas controlled experiments replace it with a filtered, perturbed, or otherwise modified version.

Protocol. Within each experiment, all settings are evaluated on the same QA instances, so every comparison is paired within model and instance. Each experiment uses its own evaluation subset, and its baselines are recomputed on that subset; subset sizes and model coverage are listed in Appendix A.2.3. Medium- and Hard-level answers are scored by alias-aware exact match against the gold answer entity, whereas Easy-level answers, which include free-text condition and verification answers, are scored by an LLM judge (Appendix A.3.2); settings are never mixed across levels, so within-model gains Δ_T and Δ_P are comparable across them.

In this notation, T and P operationalize the topology-only observation S in Proposition 1 while retaining entity names and ordering, whereas $T + C$ and $P + C$ expose the additional context observation Z in Proposition 2. The paired comparisons therefore test whether the distinguishing signal predicted by the formal construction is available to the evaluated model beyond those residual tie-breakers.

4.2 RQ1: Does Context Restore Candidate Identifiability beyond Structure?

RQ1 tests the interface-level prediction that relation-matched context can distinguish candidates that share a structural signature. The standard setting measures the paired response defined in Proposition 3; a random-context control tests whether the response requires relation-matched content rather than prompt length or format; and a retrieval setting tests whether it persists when the graph must first be located from the question.

Base Context-Gain Evaluation. Q and C are baselines for question-only answering and for context without explicit structure. The primary comparisons are the paired settings T versus $T + C$ and P versus $P + C$: both expose the same candidate triples, with T leaving them unordered and P organizing them into candidate paths. This comparison tests whether relation-level context restores a distinction that the triple-only interface leaves unresolved under each structural representation:

$$\Delta_T = \mathrm{Acc}(T + C) - \mathrm{Acc}(T), \qquad \Delta_P = \mathrm{Acc}(P + C) - \mathrm{Acc}(P).$$

The formal result does not impose a universal 50% ceiling: the bound depends on the size and prior of each structural class. Our construction instead makes the competing candidates satisfy the same topology and tests whether the added context changes the model's branch choice, while the position-bias control tests whether the context-free interface resolves the tie with an input-order shortcut.

Table 2 shows that relation-level context improves every model at both levels, with gains between 10.73 and 32.85 pp. On Easy-level QA the mean gain is 24.31 pp over unordered triples and 20.06 pp over candidate paths: single-relation questions leave little structural ambiguity, so path organization adds nothing and slightly lowers accuracy, a 4.25 pp gap in favor of unordered triples.

The Medium level provides the sharpest test, because structure alone leaves the gold relation underdetermined. For all seven models, context on candidate paths yields a larger gain than context

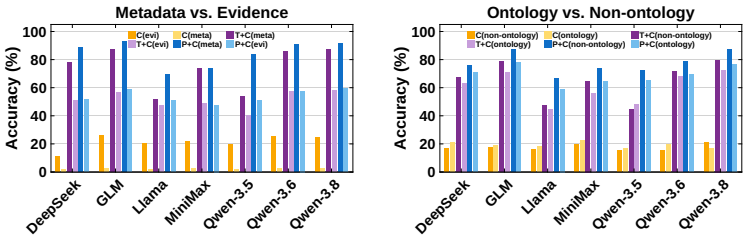


Figure 4: Comparison of context types across models. (a) Metadata vs. evidence. (b) Ontology vs. non-ontology context, with and without structure.

on unordered triples (28.20 versus 22.71 pp on average), so path organization preserves more of the benefit of relation-level context. The Q-only and C-only baselines rule out question knowledge and context alone as sufficient explanations: they reach 7.44% and 32.91%, far below $T + C$ (82.21%) and $P + C$ (89.56%) and below the structure-only settings. The effective input is therefore the combination of the question condition, candidate structure, and edge-aligned context: structure narrows the candidate relations, and context resolves the remaining distinctions. The P baseline also reveals the predicted non-structural tie-breaker. In a position-stratified control over 19,521 two-branch questions, putting the gold path first gives 72.39% accuracy, whereas putting it second gives 50.95%; with context, the corresponding values are 90.77% and 88.52%. The 21.44 pp order gap without context shrinks to 2.25 pp with context, showing position sensitivity.

Random Context Control. Unrelated text of matched length does not reproduce the gain. We replace the context of Medium-level instances with unrelated context of similar length (±5%), preserving volume, template, and slot layout while breaking the correspondence with the displayed relations. For $X \in \{T, P\}$, $\Delta_{\text{random}}(X) = \text{Acc}(X + \text{RandomC}) - \text{Acc}(X)$ measures the effect of adding unrelated text, and $\Delta_{\text{replace}}(X) = \text{Acc}(X + \text{RandomC}) - \text{Acc}(X + C)$ measures the loss from removing relation-matched content. In the three-model control (Llama-3.1-8B, Qwen3.8-27B, GLM-5.2), random context lowers accuracy below the no-context baseline by 11.81 pp for T and 8.59 pp for P, while replacing the original context costs 24.38 and 23.63 pp (Table 7). Within this control, the gain therefore tracks the correspondence between context and relation rather than additional text or a shared prompt format. The random passages are length-matched rather than shorter: the median prompt is 462 tokens against 495 for matched context under $T + C$ and 489 against 522 under $P + C$, and the gain does not grow with the number of tokens that context adds (Appendix A.10). Length is therefore not a sufficient explanation for the gain.

Retrieval Context-Gain Evaluation. The gain survives when retrieval replaces the provided subgraph. In the retrieval setting, a pipeline identifies entity anchors and relation cues from the question alone and retrieves graph information from the KG (Appendix A.3.3). For each question, the retrieved graph (R) is held fixed and shown with or without its relation-level context ($R + C$), so the gain Δ_R is not confounded by changes in retrieval. Context improves accuracy for all seven models, from 17.70% to 32.53% on average (+14.83 pp), with double-digit gains for six models (Table 8). The paired R versus $R + C$ comparison shows that relation-level context remains useful when graph information is retrieved from the question rather than provided in advance.

Answer to RQ1. On structurally underdetermined questions, the triple interface does not reliably select among candidates that share a topology: P reaches 61.35%, while $P + C$ reaches 89.56% (+28.20 pp). Question-only and context-only inputs reach only 7.44% and 32.91%, so the result is not explained by either source alone. The gain persists when the graph is retrieved from the question rather than provided, and unrelated matched-length text lowers accuracy below the triple-only baseline. Relation-level context therefore restores candidate selection when it is attached to the same structure and matched to the relevant edge.

4.3 RQ2: Which Context Provides the Distinguishing Signal?

RQ2 compares context channels on the same instances while keeping the question, candidates, and structure fixed (Appendix A.2.3). It exposes metadata versus textual evidence, then compares ontology definitions with non-ontology evidence. The second is a specified policy comparison, not a universal ranking.

Structured metadata is the more informative context once candidate relations are explicit. Without structure (C only), evidence yields 21.11% average accuracy and metadata 2.30% (Figure 4a): free text carries direct semantic cues, whereas a fold change or p-value is uninformative

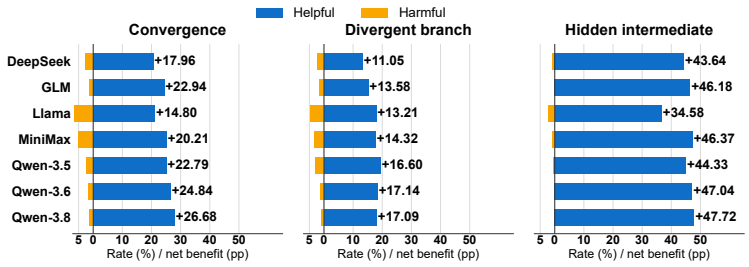


Figure 5: Context effects across Medium-level QA structural subtypes under normal conditions.

without knowing which relation it qualifies. Once structure is shown to the model, the ordering reverses: metadata reaches 73.87% under $T+C$ and 84.22% under $P+C$, versus 51.21% and 53.79% for evidence, because it carries conditions and effect sizes that separate the competing relations; in case (b) of Appendix A.10.2, the fold change and adjusted p-value stated in the question match only the gold record, so the metadata identifies the answer once the candidate records are visible.

Non-ontology evidence yields higher accuracy under the evaluated context-selection policies. The policies perform similarly without structure; with structure, non-ontology evidence exceeds ontology definitions by 4.45 pp under $T+C$ and 8.33 pp under $P+C$ (Figure 4b). Definitions explain endpoint terms, whereas empirical evidence describes particular relations. This paired contrast supports retaining relation-specific observations alongside topology; it does not separate their semantic content from differences in coverage and selection.

Answer to RQ2. With explicit structure, metadata outperforms textual evidence, and the non-ontology evidence policy outperforms the ontology-definition policy. Without structure, textual evidence is more useful than metadata. The utility of each channel therefore depends on how it is connected to candidate relations.

4.4 RQ3: When Does Context Matter, and Is It Used Robustly?

RQ1 and RQ2 establish the average response and the information channel; RQ3 asks whether the response follows the proposed interface mechanism rather than a shortcut. Its three Hard-level variants target relation matching, context quantity, and structural completeness, respectively: Context Swap changes instance-specific content while preserving type compatibility; Context-Quantity Debiasing removes any evidence-count advantage; and Structural Support Reduction removes a gold-specific or competitor-specific support edge while retaining its context. Under normal conditions, we compare the three Medium-level subtypes by paired transitions between the context-free and context-augmented settings: the Helpful rate is the proportion of instances that change from incorrect to correct, the Harmful rate the proportion that change from correct to incorrect, and the net benefit is their difference. We test each paired comparison with an exact two-sided McNemar test (McNemar, 1947) and report 95% normal-approximation intervals for the net benefit (Table 9); all 56 comparisons (seven models, three subtypes plus overall, $T \to T+C$ and $P \to P+C$) are significant at $p < 10^{-35}$ after Bonferroni correction, so the gain holds for each model rather than only on average. Because Medium-level questions share anchors, the per-question test treats correlated items as independent; resampling the anchor clusters leaves every direction and subtype ordering unchanged (Appendix A.10).

Context matters most among questions with the least structural constraint. Hidden-intermediate questions, which structure alone cannot separate, gain the most from context: a 44.96% Helpful rate and a 44.27 pp net benefit, against 21.46 pp for convergent and 14.71 pp for divergent questions (Figure 5). The ordering holds for all seven models, whose hidden-intermediate intervals lie entirely above those of the other two subtypes (Table 9).

Hidden-intermediate questions leave the intermediate entity unnamed while fixing its neighbours, so several candidates remain structurally compatible and only relation-level context separates them. A surface-similarity shortcut would predict the opposite ordering: in the encoder analysis of Appendix A.7, hidden-intermediate gold contexts are the *least* separable (within-instance AUC 0.588, versus 0.880 for convergent and 0.913 for divergent), and the same encoder used as a non-LLM branch selector reaches 82.09% overall but only 58.33% on hidden-intermediate instances. The same ordering holds on all 20,000 Medium instances: a lexical-overlap selector that uses no language model reaches 78.25% overall but only 57.11% on hidden-intermediate questions, near the

49.97% random baseline, where LLMs reach 83.68–95.06% (Appendix A.9); the shortcut offers no account of the subtype where context matters most, so the subtype ordering is not an artifact of context separability.

Perturbed Reasoning Analysis. (1) Context Swap. The random-context control of Section 4.2 tests whether *relevant* context is needed; Context Swap asks whether *instance-specific* context is needed. Donor context shares the head, relation, and tail types, so the swapped passage stays plausible while no longer matching the relation instance; the swap also changes the system prompt (Appendix A.3.2). Accuracy decreases for every model after the swap (Table 10): average accuracy on the 19,985 swap instances falls from 82.21% to 71.38% for $T + C$ and from 89.55% to 77.63% for $P + C$, so the models depend on instance-specific content beyond relation type. The context gain after the swap remains 11.87 and 16.28 pp above the T (59.50%) and P (61.34%) baselines, and C-only accuracy remains above Q (17.95% vs. 7.44%), since only one gold-path edge is swapped (case (c) in Appendix A.10.2). Two placebo checks separate the content effect from the changed prompt and rendering: $P + C$ moves by at most 0.93 pp on instances whose donor context is identical to the original, while the content effect averages 12.02 pp, and swaps that replace a competitor-specific edge cost at most 1.9 pp against 23.83–27.18 pp for gold-path support, so the lost accuracy comes from choosing the wrong branch rather than from copying a donor name (Appendix A.8).

(2) Context-Quantity Debiasing. Removing any advantage of the gold path in the amount of informative evidence leaves accuracy essentially unchanged, from 72.77% to 71.13% under $T + C$ and from 80.58% to 80.21% under $P + C$, with the C-only accuracy moving by 0.22 pp (Table 11): the gain tracks relation-specific content rather than evidence volume. The 3,000 instances split into 1,980 where a competitor naturally carries more evidence and 1,020 constructed by reducing gold-path evidence, and the gain stays positive in both (18.82 and 9.02 pp under $T + C$; 28.94 and 12.52 pp under $P + C$; Appendix A.3.1).

(3) Structural Support Reduction. One structural-support edge is removed from the candidate subgraph while its context is retained as text, simulating an incomplete graph that still records the finding in text. Retaining the context of the removed edge recovers 16.77 and 20.94 pp over the degraded-graph settings T_{deg} and P_{deg}, and the remaining gap of 9.09 and 7.64 pp to the intact $T + C$ and $P + C$ settings is consistent with context and structure being complementary rather than substitutable (Table 12). Shared edges are never removed (Appendix A.3.1). The reported average mixes two opposite deletions: deleting a gold-specific support edge costs 27.45 pp without context but 17.66 pp with it in the P contrast, so context absorbs part of that loss, whereas deleting a competitor-specific edge leaves the gold path as the only complete route (Table 5), and the two deletions therefore act in opposite directions.

Answer to RQ3. Context matters most where structure constrains least: the gain peaks on hidden-intermediate questions, whose candidates are structurally interchangeable. Instance-mismatched but type-matched context lowers accuracy, context compensates for a removed structural-support edge, and removing the evidence-count advantage of the gold path leaves the gain intact, so the recovered content complements graph structure rather than restating it. Error destinations follow: competing-branch errors (answers naming a competing branch) fall from 80.45% under P to 41.03% under $P + C$, and graph-entity errors (answers naming another entity in the graph) rise to 47.58% (Appendix A.10), shifting the errors from branch to entity selection.

5 Conclusion

Knowledge graphs that ground LLMs expose biomedical claims through an interface that may leave competing hypotheses observationally equivalent. On structurally underdetermined, context-resolvable questions, relation-level context restores the intended distinction: accuracy rises from 61.35% to 89.56% over explicit candidate paths and the gain remains +14.83 pp under retrieval. The evidence closes an interface-level behavioral loop: topology creates competing candidates, triple-only inputs remain sensitive to non-structural tie-breakers, relation-matched context changes branch selection, and the remaining errors shift mainly to downstream entity selection.

The conclusion is about representation: triples preserve connectivity but can omit the observation identifying the correct relation instance. Context Swap, quantity controls, hidden-intermediate cases, and gold-specific support removal show that the recovered signal is relation-matched and complements structure rather than merely adding text. The audit bounds the claim: provenance and attachment are often reliable, while relation support remains imperfect. We therefore measure interface-level recovery, not a universal ceiling or a particular internal mechanism.

USE OF AI ASSISTANTS

We used generative AI tools for language editing and as fixed components of the benchmark pipeline: sentence-level verification of literature-derived evidence and scoring of Easy-level free-text answers. These uses are described in Appendix A.2.1 and Appendix A.3.2; they were not used to generate the reported Medium- or Hard-level labels. We reviewed the language edits and manually verified the scientific content and evaluation procedures. We take responsibility for the final content of this work.

REFERENCES

John A. Bachman, Benjamin M. Gyori, and Peter K. Sorger. Automated assembly of molecular mechanisms at scale from text mining and curated databases. *Molecular Systems Biology*, 19(5): e11325, 2023.

Garth R. Brown, Vichet Hem, Kenneth S. Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D. Pruitt, Donna R. Maglott, and Terence D. Murphy. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research*, 43(D1):D36–D42, 2015. doi: 10.1093/nar/gku1055.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023. doi: 10.1038/s41597-023-01960-3. URL https://doi.org/10.1038/s41597-023-01960-3.

Alexander D. Diehl, Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, Ceri E. Van Slyke, Nicole A. Vasilevsky, Melissa A. Haendel, Judith A. Blake, and Christopher J. Mungall. The cell ontology 2016: Enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, 7(1):44, 2016. doi: 10.1186/s13326-016-0088-7.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Yichun Feng, Jiawei Wang, Ruikun He, Lu Zhou, and Yixue Li. A retrieval-augmented knowledge mining method with deep thinking llms for biomedical research and clinical support. *GigaScience*, 14:giaf109, 01 2025. ISSN 2047-217X. doi: 10.1093/gigascience/giaf109. URL https://doi.org/10.1093/gigascience/giaf109.

Mikhail Galkin, Priyansh Trivedi, Gaurav Maheshwari, Ricardo Usbeck, and Jens Lehmann. Message passing for hyper-relational knowledge graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7346–7359, 2020.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, 2018.

Benjamin M. Gyori, John A. Bachman, Kartik Subramanian, Jeremy L. Muhlich, Lucian Galescu, and Peter K. Sorger. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, 13(11):954, 2017.

S. He, Y. Tan, Q. Ye, et al. AI-powered immune cell knowledge graph (ICKG) with granular immune contexts enables immune program interpretation. *npj Artificial Intelligence*, 2:13, 2026. doi: 10.1038/s44387-025-00060-4.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L. Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017.

Jiatan Huang, Mingchen Li, Zonghai Yao, Dawei Li, Yuxin Zhang, Zhichao Yang, Yongkang Xiao, Feiyun Ouyang, Xiaohan Li, Shuo Han, and Hong Yu. RiTeK: A dataset for large language models complex reasoning over textual knowledge graphs in medicine, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William 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.

Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C. Rindflesch. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28 (23):3158–3160, 12 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts591. URL https://doi.org/10.1093/bioinformatics/bts591.

Jun Hyeong Kim, Dongki Kim, Yinhua Piao, and Sung Ju Hwang. Adapath: Query-adaptive path-finding via path-bank for multi-hop implicit biomedical kgqa, 2026. URL https://arxiv.org/abs/2608.30556.

Craig Knox, Mike Wilson, Christen M. Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Lucy Chin, Seth A. Strawbridge, Marysol Garcia-Patino, Ray Kruger, Aadhavya Sivakumaran, Selena Sanford, Rahil Doshi, Nitya Khetarpal, Omolola Fatokun, Daphnee Doucet, Ashley Zubkowski, Dorsa Yahya Rayat, Hayley Jackson, Karxena Harford, Afia Anjum, Mahi Zakir, Fei Wang, Siyang Tian, Brian Lee, Jaanus Liigand, Harrison Peters, Ruo Qi Rachel Wang, Tue Nguyen, Denise So, Matthew Sharp, Rodolfo da Silva, Cyrella Gabriel, Joshua Scantlebury, Marissa Jasinski, David Ackerman, Timothy Jewison, Tanvir Sajed, Vasuk Gautam, and David S. Wishart. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 2024. doi: 10.1093/nar/gkad976.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, 2021.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations*, 2024.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

John H. Morris, Karthik Soman, Rabia E. Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C. Meng, Conrad C. Huang, Gabriel Cerono, Gundolf Schenk, Angela Rizk-Jackson, et al. The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics*, 39(2):btad080, 2023.

Christopher J. Mungall, Carlo Torniai, Georgios V. Gkoutos, Suzanna E. Lewis, and Melissa A. Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13(1):R5, 2012. doi: 10.1186/gb-2012-13-1-r5.

OpenBEL Consortium. BEL specification, version 2.0. https://openbel.org/language/version_2.0/bel_specification_version_2.0.html, 2014.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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. 3982–3992, 2019.

Khaled Rjoob, Kathryn A. McGurk, Sean L. Zheng, Lara Curran, Mahmoud Ibrahim, Lingyao Zeng, Vladislav Kim, Shamin Tahasildar, Soodeh Kalaie, Deva S. Senevirathne, Parisa Gifani, Vladimir Losev, Jin Zheng, Wenjia Bai, Antonio de Marvao, James S. Ware, Christian Bender,

and Declan P. O'Regan. A multimodal vision knowledge graph of cardiovascular disease. *Nature Cardiovascular Research*, 5(1):18–33, 2026. doi: 10.1038/s44161-025-00757-4.

Graciela Rosemblat, Marcelo Fiszman, Dongwook Shin, and Halil Kilicoglu. Towards a characterization of apparent contradictions in the biomedical literature using context analysis. *Journal of Biomedical Informatics*, 98:103275, 2019. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2019.103275. URL https://www.sciencedirect.com/science/article/pii/S1532046419301947.

Paolo Rosso, Dingqi Yang, and Philippe Cudré-Mauroux. Beyond triplets: hyper-relational knowledge graph embedding for link prediction. In *Proceedings of The Web Conference 2020*, pp. 1885–1896, 2020.

Tara Safavi and Danai Koutra. CoDEx: A Comprehensive Knowledge Graph Completion Benchmark. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8328–8350, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.669. URL https://aclanthology.org/2020.emnlp-main.669/.

Eric Sayers. A general introduction to the e-utilities. In *Entrez Programming Utilities Help*. National Center for Biotechnology Information (US), Bethesda, MD, 2009. Updated November 17, 2022.

Jürgen Scheller, Athena Chalaris, Dirk Schmidt-Arras, and Stefan Rose-John. The pro- and anti-inflammatory properties of the cytokine interleukin-6. *Biochimica et Biophysica Acta (BBA) – Molecular Cell Research*, 1813(5):878–888, 2011.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 31210–31227, 2023.

Karthik Soman, Peter W. Rose, John H. Morris, Rabia E. Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, et al. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(9): btae560, 2024.

Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34 (suppl_1):D535–D539, 2006. doi: 10.1093/nar/gkj109.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*, 2024.

Xingyu Tan, Shiyuan Liu, Xiaoyang Wang, Qing Liu, Xiwei Xu, Xin Yuan, Liming Zhu, and Wenjie Zhang. Weaving multi-source evidence for biomedical reasoning: The biomedhop benchmark and bioweave framework, 2026. URL https://arxiv.org/abs/2606.16211.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025. doi: 10.1093/nar/gkae1010.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138, 2015.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis N. Ioannidis, Karthik Subbian, James Zou, and Jure Leskovec. STaRK: Benchmarking LLM retrieval on textual and relational knowledge bases. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024.

Ran Xu, Patrick Jiang, Linhao Luo, Cao Xiao, Adam Cross, Shirui Pan, Jimeng Sun, and Carl Yang. A survey on unifying large language models and knowledge graphs for biomedicine and healthcare. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 6195–6205, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736556. URL https://doi.org/10.1145/3711896.3736556.

Duzhen Zhang, Zixiao Wang, Zhong-Zhi Li, Yahan Yu, Shuncheng Jia, Jiahua Dong, Haotian Xu, Xing Wu, Yingying Zhang, Tielin Zhang, Jie Yang, Xiuying Chen, and Le Song. Medkgent: a large language model agent framework for constructing temporally evolving medical knowledge graph. *npj Digital Medicine*, 08 2026. ISSN 2398-6352. doi: 10.1038/s41746-026-03058-7. URL https://doi.org/10.1038/s41746-026-03058-7.

A Appendix

A.1 Foundations and Identifiability

A.1.1 Formal Proofs for Structural Non-Identifiability

This appendix expands the proof sketches in Section 3.2. The statements concern the structural signature S, not the complete serialized prompt: entity names and order can provide additional priors, which is why the empirical protocol measures rather than assumes a chance ceiling.

Proof of Proposition 1. Fix q and a known finite structural class E of size k_q. In the reference experiment, sample the gold member Y after fixing the observed pair (q, s), with $\Pr(Y = h \mid q, s, E) = \pi_q(h)$; the predictor receives no further information about this draw. Structural equality alone does not imply this conditional distribution. Write $a_h = \Pr(\hat{h} = h \mid q, s, E)$, with $a_h \geq 0$ and $\sum_{h \in E} a_h \leq 1$. Predictor randomization is independent of Y conditional on the observation. Under the uniform conditional distribution,

$$\Pr(\hat{h} = h^\star \mid q, s) = \sum_{h \in E} \frac{1}{k_q} a_h = \frac{1}{k_q} \sum_{h \in E} a_h \leq \frac{1}{k_q}. \tag{4}$$

The bound is achieved by putting all output mass on members of E, so it is tight. For a general prior π_q, the same calculation gives $\sum_{h \in E} \pi_q(h) a_h$. A linear function over the probability simplex is maximized at a vertex, so the best strategy puts all mass on $\arg\max_h \pi_q(h)$ and obtains $\max_h \pi_q(h)$. This proves that structural equivalence removes information about the gold member; it does not assert that real prompts eliminate priors or names.

Proof of Proposition 2. Add the edge-aligned context and suppose the compatibility score obeys Eq. 3. The candidate set is finite, so an argmax exists. The strict inequality says that every competitor has a smaller score than $h^\star$, hence the argmax set is the singleton $\{h^\star\}$. The rule therefore returns the gold path. If the inequality is not strict, context may still help, but this proposition no longer guarantees identification; this is the regime measured by the benchmark rather than assumed away.

Proof of Proposition 3. Let I_i^X and I_i^{X+C} be the correctness indicators for model m on the same instance i under the two interfaces. The paired accuracy difference is

$$\Delta_X = \frac{1}{n} \sum_{i=1}^{n} \left(I_i^{X+C} - I_i^X \right). \tag{5}$$

Because the question, candidate graph, ordering, model, decoding rule, and scorer are held fixed within each pair, the summand changes only when the revealed context changes the model output or its validity. It can be decomposed into helpful transitions $X : \text{wrong} \to X + C : \text{correct}$ minus harmful transitions $X : \text{correct} \to X + C : \text{wrong}$, divided by n. Thus Δ_X is an estimand for the evaluated model's response to the context interface. It is not a Bayes-optimality proof, and it does not identify an internal neural mechanism. The random-context, swap, quantity, and structural-reduction controls test the principal alternative explanations for this observed difference.

A.2 Dataset and Benchmark Construction

A.2.1 KG Construction Details

For relations lacking sufficient textual evidence, we retrieve additional evidence from PubMed through the NCBI Entrez E-utilities (Sayers, 2009). Each triple is verbalized into a natural-language query using the relation together with the names and synonyms of head and tail entities. Candidate abstracts mentioning both endpoint entities are retained. Before semantic processing, candidate records undergo programmatic filtering to verify relation presence, entity matching, and duplication. To further verify semantic correctness, an LLM is then used to select the specific evidence sentence. Finally, stratified human review is conducted across relation types, entity types, sources, and evidence types to verify that the extracted evidence refers to the correct entities and genuinely supports the corresponding relation.

A.2.2 Manual Context Audit

We manually audited 300 records sampled with seed 20260926 from 24 eligible source files. The sample was stratified by source file, relation-frequency tier, context mode, and relation; it was therefore designed to cover heterogeneous records rather than estimate prevalence in the full KG. Two annotators independently judged six binary items. The 112 "Unclear" cells from Annotator 1 were adjudicated, while "N/A" judgments were excluded from the relevant denominator.

Table 3: Manual audit of 300 context records. Pass rates are computed over the applicable records; agreement is raw exact agreement. The sample is stratified and is not a prevalence-weighted estimate for the full KG.

Audit item	n	**A1 pass**	**A2 pass**	**Agree / κ**
Correct head/tail entities	300	96.0%	96.3%	93.7% / 0.141
Evidence supports relation	300	67.0%	70.0%	81.0% / 0.560
Relation direction correct	262	88.9%	75.2%	85.5% / 0.523
Condition/measurement attached	120	99.2%	100.0%	99.2% / 0.000
Provenance correct	300	89.7%	94.7%	95.0% / 0.657
Overall valid context record	300	66.3%	68.3%	82.0% / 0.591

Table 4: QA pools and evaluation subsets. "Derived" subsets are built from the Medium-level pool.

Level	**Subset**	**#QA**	**Reported in**
Easy	triple-structure	13,229	Table 2
	contextual	16,341	
	evidence verification	20,430	
	total	50,000	
Medium	convergent	6,710	Table 2, Fig. 5
	hidden-intermediate	7,597	
	divergent	5,693	
	total	20,000	
Derived	random-context control	10,189	Table 7
	graph retrieval	2,000	Table 8
	metadata vs. evidence	5,074	Fig. 4(a)
	ontology vs. non-ontology	1,420	Fig. 4(b)
Hard	Context Swap	19,985	Table 10
	Context-Quantity Debiasing	3,000	Table 11
	Structural Support Reduction	10,000	Table 12

The audit identifies a specific quality boundary: provenance and entity attachment are usually correct, but relation support is weaker, with pass rates of 67.0% and 70.0%. Records containing metadata or provenance without relation-specific text can remain traceable while being insufficient to establish the asserted relation. The 99.2% raw agreement for condition or measurement attachment should not be read as strong kappa evidence, because one annotator supplied no negative labels. These results motivate treating context as an evaluated interface signal rather than assuming that every context record is equally evidential.

A.2.3 Benchmark Statistics and Evaluation Subsets

Table 4 summarizes the QA pools and the evaluation subsets used in each experiment. All Hard-level variants and the retrieval subset are derived from the 20,000 Medium-level instances; each derived instance shares the question, candidate set, and gold answer of its Medium-level source. Every experiment compares input-visibility settings on a fixed set of QA instances, so all within-experiment differences are paired.

Per-model coverage and scoring of the Easy level are given in Appendix A.3.2.

The Medium-level main experiment therefore comprises 120,000 model inputs per model (20,000 instances $\times$ six settings). The random-context control is run on three representative models (Llama-3.1-8B, Qwen3.8-27B, and GLM-5.2), and all other experiments are run on all seven models. The Q2 subsets are restricted to instances in which both context types being compared are present, so that each variant removes one type while holding the question, candidates, and structure fixed.

A.2.4 Input Rendering and Instance Validation

Candidate set. Each Medium-level instance contains the gold candidate together with up to seven real, structurally complete competing candidates. Candidates are required to have pairwise-distinct final answer identifiers and a unique single-answer extension, so that selecting the correct candidate and producing the correct answer are equivalent, and no two candidates lead to the same answer.

T **versus** P**.** The unordered-triple setting T presents the *identical* edge set as the path setting P, in a fixed shuffled order. P groups the same edges into candidate branches, each shown with its full structural support. The gold branch is not marked or placed in a fixed position, so neither representation reveals the answer through layout. Because T and P share the same edges, the difference between Δ_T and Δ_P reflects only the organization of structure, not its content.

Context slot. In the $+C$ settings, the relation-level context $\phi(e)$ of every displayed edge e is attached to that edge, for both gold and competing branches. Each context block lists the evidence sentences of $\mathcal{C}^{\text{evi}}$ with their evidence rationales, followed by the metadata of $\mathcal{C}^{\text{meta}}$ as key–value pairs. In the C-only setting, context passages are listed without the graph, and the prompt states explicitly that their order does not indicate correctness. The graph content is identical between X and $X + C$ for $X \in \{T, P\}$, and the T and P orders are fixed per instance by hashing the QA identifier, so the same order is used with and without context.

Automated validation. Every released instance passes a set of automated checks, including (i) context is present for every candidate, (ii) T and P contain the same edge set, (iii) every candidate has a unique single-answer extension, and (iv) the question does not mention the gold intermediate or the final answer under any alias. Leak detection uses token-boundary matching over all aliases; a leaking question is minimally rewritten and re-audited, and only rewrites that keep all anchor entities and the same answer are retained. Of the 20,000 final Medium-level instances, 2,105 carry a repaired question.

A.3 Evaluation Protocol and Perturbations

A.3.1 Hard-Level Construction Details

Context Swap. For each Medium-level instance, one eligible target edge on the gold branch is selected, and its relation-level context is replaced by the context of a donor edge that shares the same head-entity type, relation type, tail-entity type, and context channel (evidence or metadata). All other edges and contexts are left unchanged, and source instances flagged by the leakage audit are excluded. Type matching ensures that the swapped context remains superficially plausible (e.g., an interaction passage is replaced by another interaction passage), which makes this perturbation harder to detect than random replacement. The unswapped rows of Table 10 are computed on the same 19,985 source instances as the swapped rows, so each swap effect compares matched instances and is not confounded by subset composition.

Context-Quantity Debiasing. This variant removes the potential shortcut that the gold path carries more context than its competitors. Context quantity is measured as the number of informative evidence items on a branch (items of at least 25 characters that are not null-like). The variant combines instances in which some competing branch naturally carries more such items than the gold branch with constructed instances in which the evidence on the gold answer-extension edge is removed until the advantage disappears; constructed instances whose gold branch would be left with no informative evidence are dropped. Because the reduction only removes context, any remaining context gain cannot be explained by the gold path carrying more text. Of the 3,000 instances, 1,980 fall in the natural group and 1,020 were constructed by reduction. The gain is positive in both groups, at 18.82 and 9.02 pp under $T + C$ and 28.94 and 12.52 pp under $P + C$.

Structural Support Reduction. One informative, non-shared structural-support edge is removed from the candidate subgraph. In the $+C$ settings, its relation-level context is retained and shown as an additional retrieved contextual observation, so the information about the missing relation is still available in textual form while the graph no longer contains the edge. Half of the 10,000 instances remove a gold-specific support edge and half remove a support edge specific to a competing candidate. Edges shared by the gold and competing paths are never removed, so the perturbation never deletes the gold path entirely and never makes it the only remaining option. This design evaluates whether context can compensate for missing structure rather than whether the answer becomes structurally trivial. The removed edge is always drawn from a structural-support path and never from an answer-extension edge, and only edges with informative context (a non-empty evidence sentence or an informative metadata field such as a definition, fold change, p-value, or interaction score) are eligible. Every candidate branch, including its answer-extension edge to the candidate entity, is kept, so each reduced instance offers exactly the same candidate set as its Medium-level source, with the gold candidate and all competitors still present. The 10,000 instances are drawn from the Context Swap source pool, with the removed-edge role balanced overall (5,000 gold-specific and 5,000 competitor-specific removals across the three subtypes). Table 5 separates the two roles, which the merged average cancels. Removing gold-specific support costs 27.45 pp without context and 17.66 pp with it, so context absorbs 9.78 pp of that loss [7.18, 12.87]. Removing competitor-specific sup-

Table 5: Structural Support Reduction by removed-edge role. Accuracy and the fraction of answers naming a competing branch, in percent; intact values use the same 10,000 Medium-level identifiers.

	Remove gold-specific edge		Remove competitor-specific edge	
Condition	**Acc.**	**Comp. (%)**	**Acc.**	**Comp. (%)**
T	61.86	26.05	56.21	31.52
T_{deg}	31.24	52.50	80.34	9.43
$T + C$	83.21	7.60	80.09	10.01
$T_{\text{deg}} + C$	61.85	26.17	83.26	6.52
P	64.29	27.56	57.74	33.80
P_{deg}	36.84	53.96	83.45	10.42
$P + C$	90.14	3.81	87.33	5.89
$P_{\text{deg}} + C$	72.48	19.11	89.70	4.22

port makes the gold path the only complete route and therefore raises accuracy, from 57.74% to 83.45% without context; with context the same removal changes accuracy by only 2.37 pp, because context already supplies much of the missing signal.

A.3.2 Evaluation Protocol

Prompting. Models are instructed to answer using only the supplied information together with generally known biomedical knowledge, and to return one best-supported answer. The output is a JSON object {`"answer": [...], "reason": ...`} containing a single direct final answer and a short rationale. In the Context Swap variant, the system prompt additionally states that the supplied context may be mismatched or misleading, so that models are not implicitly told to trust every passage. The warning favors discounting the swapped passage, so the context gain that survives the swap is a conservative estimate. Importantly, models must *generate* the answer entity rather than select an option label: the candidate branches are shown as graph content, not as a lettered choice list, and the answer must name the entity at the end of the chosen branch.

Decoding. All models are queried with greedy decoding (temperature 0), JSON-mode output, and reasoning or "thinking" modes disabled, so that differences between settings are not confounded by variable-length hidden reasoning. The output budget is fixed within each experiment and ranges from 180 to 500 tokens across experiments. Open-weight models (Llama-3.1-8B, Qwen3.5-9B, Qwen3.6-27B, Qwen3.8-27B) are served locally through an OpenAI-compatible inference server, and the remaining models are accessed through their providers' APIs.

Scoring. Medium- and Hard-level predictions are scored by alias-aware exact match against the gold answer entity. Both prediction and aliases are normalized with Unicode NFKC normalization and case folding, and all non-alphanumeric characters are removed. The accepted alias set of an entity includes its canonical name, the names attached to its authoritative identifier, and, for diseases, the MONDO preferred label and EXACT synonyms; BROAD, NARROW, and RELATED synonyms are excluded, so that a more general or more specific disease is not credited. Malformed outputs are counted as incorrect rather than discarded, so that a setting cannot gain accuracy by failing on difficult inputs. Easy-level answers, which include free-text condition and verification answers that do not reduce to a single entity, are scored with an LLM judge.

Easy-level coverage and judging. The Easy level serves as a per-relation diagnostic; the central comparisons of the paper rest on the Medium and Hard levels, where every model is scored on every instance by deterministic alias matching. Five models answered all 50,000 Easy instances in each setting (with at most 496 missing responses in any setting, for Qwen3.8-27B), while GLM-5.2 and MiniMax-M3 were run on subsets of about 12.5k instances per setting (12,499–12,500 and 12,568–12,959, respectively). Judge labels cover a subset of these responses; under T and $T + C$, the scored denominators are 24,880–24,992 for DeepSeek-v4-flash, Llama-3.1-8B, Qwen3.5-9B, and Qwen3.6-27B, 12,499 and 12,500 for GLM-5.2, 12,959 and 12,677 for MiniMax-M3, and 2,000 for Qwen3.8-27B. Labels for DeepSeek-v4-flash, Llama-3.1-8B, Qwen3.5-9B, and Qwen3.8-27B come from an LLM judge with answer-type-specific instructions, and labels for GLM-5.2, MiniMax-M3, and Qwen3.6-27B come from post-hoc revised rule-plus-semantic scoring; in all cases entity answers are credited for aliases and synonyms. Because all settings of a given model are scored by the same procedure, the within-model gains Δ_T and Δ_P do not depend on differences between judges, and Easy-level accuracies are compared within rather than across models.

Table 6: Unified input-visibility settings used across experiments.

Setting	LLM-visible information
Q	Question only
C	Question + Relation-Level Context
T	Question + Unordered Candidate Triples
P	Question + Candidate Reasoning Paths
$T + C$	Question + Unordered Candidate Triples + Relation-Level Context
$P + C$	Question + Candidate Reasoning Paths + Relation-Level Context

Paired analysis. Helpful and harmful transitions are computed separately for each model on the instances for which both compared settings returned a technically valid response (19,935–20,000 of the 20,000 Medium-level instances, depending on model and comparison). Restricting instead to the 19,926 instances that are valid for every model and setting changes each subtype average by at most 0.05 pp. The net benefit is $(\#\text{helpful} - \#\text{harmful})/n$. Because each model is compared with itself on identical inputs except for the context slot, differences in parametric knowledge or potential pretraining exposure affect both arms of each comparison equally. Significance is assessed with an exact two-sided McNemar test, i.e., a binomial test with success probability 0.5 on the discordant pairs, $\min(\#\text{helpful}, \#\text{harmful}) \sim \text{Bin}(\#\text{helpful} + \#\text{harmful}, 0.5)$. The 95% interval for the net benefit $\hat{d}$ is the paired normal approximation $\hat{d} \pm 1.96\sqrt{(\hat{\pi}_{\text{disc}} - \hat{d}^2)/n}$, where $\hat{\pi}_{\text{disc}}$ is the discordant fraction. Across the 56 comparisons, the smallest number of discordant pairs is 862 and the largest unadjusted p-value is 1.3×10^{-37}; a Bonferroni correction over the 56 tests leaves every comparison at $p < 10^{-35}$.

A.3.3 Retrieval Setting Details

The retrieval pipeline receives only the question; the gold candidate and answer are used exclusively for offline scoring. Entity anchors are identified as the longest non-overlapping alias n-grams ($n \leq 8$) in the question, and a rule-based parser maps the relational cues attached to each anchor and to the answer slot onto relation groups of the KG. When a question contains two or more anchors, their candidate sets are intersected, with graded relaxation if the intersection is empty. To keep the prompt bounded, retrieval is capped at 4 anchors, 20 candidates, 5 answers per candidate, 50 branches, one answer-extension hop, and 8 support edges, and a strict answer-relation filter removes branches whose final relation does not match the parsed answer relation. The 2,000 evaluated questions are balanced across the three Medium-level subtypes (667/667/666). In the $R + C$ setting, the context records of each retrieved edge are attached, subject to a global budget of 56,000 characters beyond which records are compacted. Because R and $R + C$ share exactly the same retrieved graph, Δ_R is not confounded by changes in retrieval.

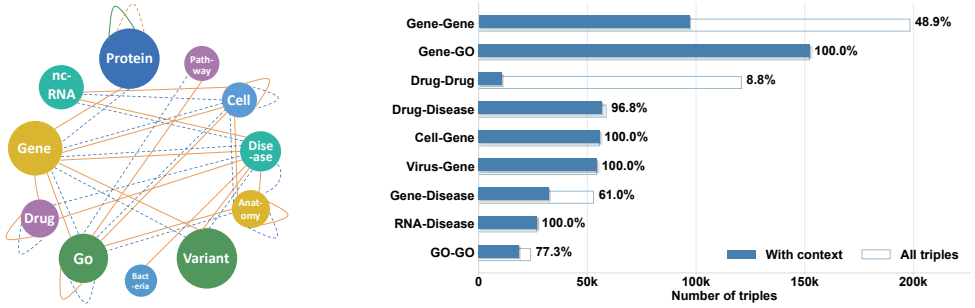
A.3.4 Context-Type Comparison Details

Metadata versus evidence. An instance is eligible if its candidate subgraph contains at least one edge with usable metadata and at least one edge with usable evidence. In the metadata variant, only structured metadata fields (definition-type and numeric fields) are shown; in the evidence variant, only evidence items with a sentence or rationale are shown. The graph and its presentation order are unchanged, so the two variants differ only in the context channel.

Ontology versus non-ontology. Relations are labelled as ontology relations if they originate from ontology resources (cell and anatomy ontologies, Uberon, and Gene Ontology hierarchies), and as non-ontology relations otherwise (e.g., protein interactions, gene–disease, drug–disease, pathway, variant, and expression relations). For ontology edges, the context consists of the authoritative definitions of the head and tail entities; for non-ontology edges, it consists of the single most informative evidence item of the edge. This operationalization reflects the nature of the two relation families: ontology relations are defined by the meaning of their endpoint terms, whereas non-ontology relations are supported by experimental or literature evidence about the specific pair.

A.4 Detailed Results and Case Evidence

The tables below report the per-model results underlying the main-text analyses; every average quoted in the main text is computed from them.



(a) Entity and relation schema. **(b)** Triple distribution and context coverage.

Figure 6: Statistics of the constructed biomedical knowledge graph. (a) Entity and relation schema. (b) Triple distribution and context coverage.

Table 7: Effects of random context on Medium-level QA. Accuracy is reported in percentage (%), and differences are measured in percentage points (pp).

	Triple-based					Path-based				
Model	T	$T+\text{RandomC}$	$T+\text{C}$	$\Delta_{\text{random}}(T)$	$\Delta_{\text{replace}}(T)$	P	$P+\text{RandomC}$	$P+\text{C}$	$\Delta_{\text{random}}(P)$	$\Delta_{\text{replace}}(P)$
Llama-3.1	56.68	28.37	63.71	-28.31	-35.34	60.70	40.18	73.69	-20.52	-33.51
Qwen-3.8	72.05	64.91	89.02	-7.14	-24.11	74.15	68.44	92.66	-5.71	-24.22
GLM-5.2	73.44	73.47	87.15	0.03	-13.68	77.40	77.85	91.02	0.45	-13.17
Average	**67.39**	**55.58**	**79.96**	**-11.81**	**-24.38**	**70.75**	**62.16**	**85.79**	**-8.59**	**-23.63**

Table 8: Performance (%) and context gains (pp) under graph retrieval.

Setting	**DeepSeek-v4-flash**	**GLM-5.2**	**Llama-3.1-8B**	**MiniMax-M3**	**Qwen3.5-9B**	**Qwen3.6-27B**	**Qwen3.8-27B**	**Average**
Retrieval R	17.30	21.90	12.50	18.65	14.70	18.90	19.95	17.70
Retrieval $R+C$	35.05	39.10	15.30	35.65	27.45	37.30	37.85	32.53
Δ_R	17.75	17.20	2.80	17.00	12.75	18.40	17.90	**14.83**

Table 9: 95% confidence intervals of the net benefit of context, Helpful rate minus Harmful rate, across Medium-level QA structural subtypes (P vs. $P+C$).

Model	**Overall**	**Convergence**	**Divergent**	**Hidden-intermediate**
DeepSeek	[25.08, 26.41]	[16.89, 19.03]	[10.07, 12.03]	[42.49, 44.78]
GLM	[28.44, 29.76]	[21.85, 24.02]	[12.57, 14.58]	[45.04, 47.31]
Llama	[21.16, 22.56]	[13.59, 16.00]	[12.02, 14.40]	[33.41, 35.74]
MiniMax	[27.76, 29.18]	[18.99, 21.43]	[13.18, 15.45]	[45.21, 47.53]
Qwen3.5	[28.53, 29.89]	[21.66, 23.91]	[15.46, 17.74]	[43.19, 45.48]
Qwen3.6	[30.42, 31.75]	[23.72, 25.97]	[16.09, 18.20]	[45.92, 48.17]
Qwen3.8	[31.27, 32.61]	[25.55, 27.80]	[16.05, 18.13]	[46.59, 48.85]

Table 10: Performance (%) under normal and context-swap conditions on the same 19,985 instances.

Setting	**DeepSeek-v4-flash**	**GLM-5.2**	**Llama-3.1-8B**	**MiniMax-M3**	**Qwen3.5-9B**	**Qwen3.6-27B**	**Qwen3.8-27B**	**Average**
C	34.91	34.11	29.85	34.00	29.69	34.03	33.91	32.93
C_{Swap}	18.80	20.42	14.11	18.75	14.73	19.69	19.17	17.95
$T+C$	83.34	89.84	62.96	86.85	71.03	89.59	91.86	82.21
$T+C_{\text{Swap}}$	70.96	83.25	49.70	73.66	57.25	82.07	82.74	71.38
T	62.39	63.09	52.22	60.91	53.87	62.03	62.00	59.50
$P+C$	91.82	94.28	76.98	90.67	84.25	93.65	95.23	89.55
$P+C_{\text{Swap}}$	76.14	85.68	62.67	80.70	69.08	83.92	85.19	77.63
P	66.07	65.16	55.12	62.19	55.03	62.55	63.29	61.34

Table 11: Performance under reduced gold-path context quantity.

Setting	DeepSeek-v4-flash	GLM-5.2	Llama-3.1-8B	MiniMax-M3	Qwen3.5-9B	Qwen3.6-27B	Qwen3.8-27B	Average
C	23.97	25.08	18.17	25.23	20.67	24.90	26.60	23.52
C_{disadv}	24.37	26.73	16.43	26.01	20.30	24.33	24.93	23.30
$T + C$	70.33	82.63	52.70	78.73	57.10	81.50	86.37	72.77
$T + C_{\text{disadv}}$	69.70	82.20	49.07	73.08	54.57	83.33	85.93	71.13
$P + C$	81.30	88.27	61.97	84.27	73.20	85.40	89.67	80.58
$P + C_{\text{disadv}}$	81.33	89.80	58.27	84.24	71.43	86.83	89.60	80.21

Table 12: Performance under graph structure degradation.

Setting	DeepSeek-v4-flash	GLM-5.2	Llama-3.1-8B	MiniMax-M3	Qwen3.5-9B	Qwen3.6-27B	Qwen3.8-27B	Avg.
T_{deg}	55.37	60.00	47.48	57.03	51.70	60.40	58.54	55.79
P_{deg}	60.15	64.13	53.77	60.89	56.18	62.81	63.09	60.15
$T + C$	82.33	89.41	62.60	86.14	70.51	88.99	91.56	81.65
$T_{\text{deg}} + C$	71.91	80.59	56.14	73.00	65.33	79.54	81.38	72.56
$P + C$	91.26	93.70	75.81	89.92	83.02	92.78	94.63	88.73
$P_{\text{deg}} + C$	79.30	89.59	64.88	81.56	75.53	86.73	90.03	81.09

A.5 Relation-wise Analysis of Context Gains

The aggregate gains of Section 4.2 do not show whether context acts uniformly across relation types. We therefore decompose the context gain by relation and model, and ask which relations benefit most, whether the ordering is consistent across models, and how the contribution changes when the same edges are organized as explicit candidate paths.

For each relation type r and model m, we compute the context gains as

$$\Delta_T(r, m) = \text{Acc}_{r,m}(T + C) - \text{Acc}_{r,m}(T), \tag{6}$$

$$\Delta_P(r, m) = \text{Acc}_{r,m}(P + C) - \text{Acc}_{r,m}(P), \tag{7}$$

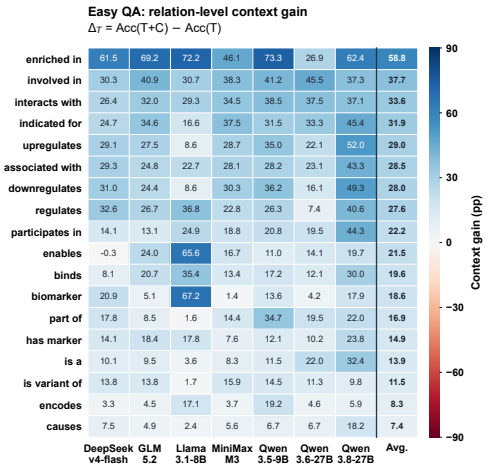
Since an Easy-level question is centered on a single target relation rather than a multi-hop supporting path, the relation-wise decomposition for Easy-level QA reports Δ_T; the aggregate Easy-level Δ_P is given in Table 2. The resulting relation-by-model gain matrices for Easy-level QA and Medium-level QA are presented below.

Across both Easy and Medium-level QA, relation-level context improves performance across the vast majority of relation types, but the magnitude of the improvement varies substantially by relation. In Easy-level QA, for example, *enriched in* achieves an average gain of 58.8 pp, whereas *encodes* and *causes* improve by only 8.3 and 7.4 pp. Medium-level QA shows a similarly structured distribution, with several regulatory, association, and participation relations receiving large gains, while *is a* remains close to zero. These results show that the value of context depends strongly on the semantics represented by the relation rather than acting as a uniform source of additional information.

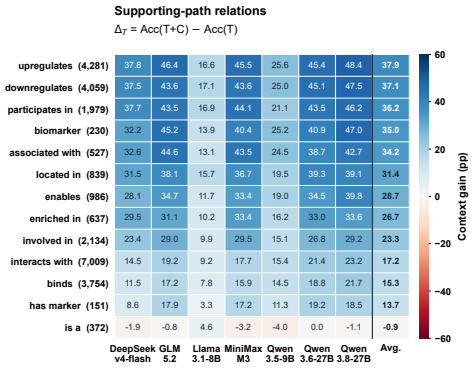
Easy-level QA isolates the contribution of context at the level of a single target relation. The largest gain is observed for *enriched in* (58.8 pp), followed by *involved in* (37.7 pp), *interacts with* (33.6 pp), and *indicated for* (31.9 pp). These relations frequently depend on instance-specific biological information that is not represented by the relation label alone. For example, an *enriched in* edge identifies a gene–cell association, but the structural triple does not encode the expression strength, experimental background, or other quantitative information associated with that particular relation instance.

In contrast, *encodes* and *causes* obtain much smaller average gains. Their structural relations already impose a more specific semantic constraint on the connected entities, leaving less unresolved information for relation-level context to supply. The Easy-level QA results therefore show that the benefit of context increases when the discrete relation label compresses a larger amount of instance-specific information.

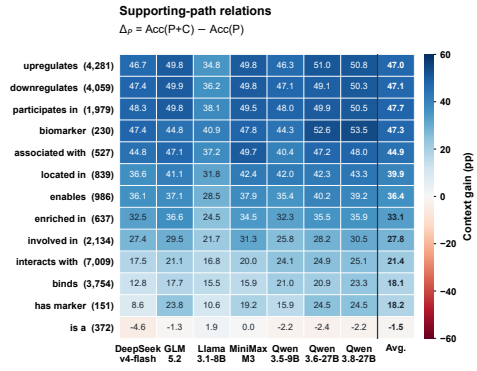
Medium-level QA reveals a different role of relation-level context because the model must distinguish among multiple structurally valid reasoning routes. A particularly consistent result is that the path-based context gain Δ_P exceeds the triple-based gain Δ_T for 12 of the 13 analyzed relations. For example, the average gain for *biomarker* increases from 35.0 to 47.3 pp, *participates in* from 36.2 to 47.7 pp, and *associated with* from 34.2 to 44.9 pp when moving from the triple to the path setting. Similar increases are observed for *upregulates* ($37.9 \rightarrow 47.0$ pp) and *downregulates* ($37.1 \rightarrow 47.1$ pp).



(a) Easy QA: Triple setting (Δ_T).



(b) Medium QA: Triple setting (Δ_T).



(c) Medium QA: Path setting (Δ_P).

Figure 7: Relation-wise context gains across models for Easy and Medium-level QA. Panel (a) shows the triple-based gains for Easy-level QA, while Panels (b) and (c) show the triple- and path-based gains for Medium-level QA, respectively. Each cell reports the change in accuracy after adding relation-level context, measured in percentage points.

Organizing the same edges as explicit candidate paths raises the gain rather than dissipating it: the path identifies the structurally plausible reasoning routes, and the associated context supplies the instance-specific evidence that distinguishes among them. The larger gains under the path setting therefore indicate that context becomes more useful when it can be aligned with an explicitly represented reasoning route.

The main exception is *is a*, whose average gain remains close to zero and is slightly negative in both settings (-0.9 pp for Δ_T and -1.5 pp for Δ_P). As a definitional ontology relation, *is a* already expresses most of its relevant semantics through the graph structure itself, which matches the ontology versus non-ontology contrast of Section 4.3. Additional context therefore adds no discriminative information for this relation.

Across models, the vast majority of relation types consistently benefit from relation-level context, showing that the observed gains are not specific to individual model families. However, the magnitude of the gain differs considerably across models of varying scales. Under the Medium-level triple setting, for example, the gain for *upregulates* is 16.6 pp for Llama-3.1-8B, compared with 45.4 and 48.4 pp for Qwen3.6-27B and Qwen3.8-27B. The same pattern appears for *downregulates*, with gains of 17.1, 45.1, and 47.5 pp, respectively. The two larger Qwen models therefore extract substantially more benefit than Llama-3.1-8B from the same contextual information in these multi-step reasoning cases. The relationship is not strictly determined by parameter count, however: several Easy-level relations show different ordering across models. The results indicate that context utilization depends on both model capacity and the specific semantic and reasoning demands of the task.

Table 13: Paired sample sizes used for the relation-wise helpful–harmful decomposition. The same paired QA instances are used to compute both helpful and harmful transition rates.

Relation	**Paired** n
participates in	1,979
biomarker	230
downregulates	4,059
upregulates	4,281
associated with	527
located in	839
enables	986
enriched in	637
involved in	2,134
interacts with	7,009
has marker	151
binds	3,754
is a	372

Overall, the relation-wise results show that relation-level context contributes most when the graph relation fails to capture much of the instance-specific semantics. In Medium-level QA, this contribution becomes even stronger when the context can be combined with an explicit reasoning path, consistent with the complementary roles of graph structure and relation-level context.

A.6 Helpful–Harmful Decomposition across Relation Types

The relation-wise gains reported above describe the net change in accuracy, but do not distinguish how that change is produced. A positive gain can arise because context corrects previously incorrect predictions, because it introduces few new errors, or through a combination of both effects. We therefore perform a paired decomposition of context-induced prediction changes for each supporting-path relation.

For each relation r, model m, and structural input setting $X \in \{T, P\}$, we define

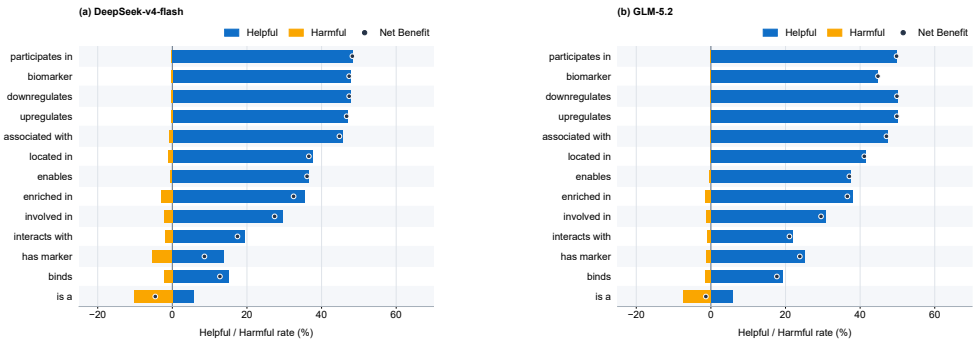
$$H^{+}(r, m; X) = \frac{\#\{\text{wrong}_X \rightarrow \text{correct}_{X+C}\}}{N_{r,m;X}}, \tag{8}$$

$$H^{-}(r, m; X) = \frac{\#\{\text{correct}_X \rightarrow \text{wrong}_{X+C}\}}{N_{r,m;X}}, \tag{9}$$

where H^{+} and H^{-} denote the helpful and harmful transition rates, respectively. The corresponding net benefit is

$$B_{\text{net}}(r, m; X) = H^{+}(r, m; X) - H^{-}(r, m; X). \tag{10}$$

Helpful and harmful transitions are reported as percentages of paired QA instances, while net benefit is reported in percentage points. Because the comparison is performed on the same paired instances, the net benefit directly corresponds to the context-induced accuracy change for that relation. The decomposition is reported separately for $T \rightarrow T + C$ and $P \rightarrow P + C$.



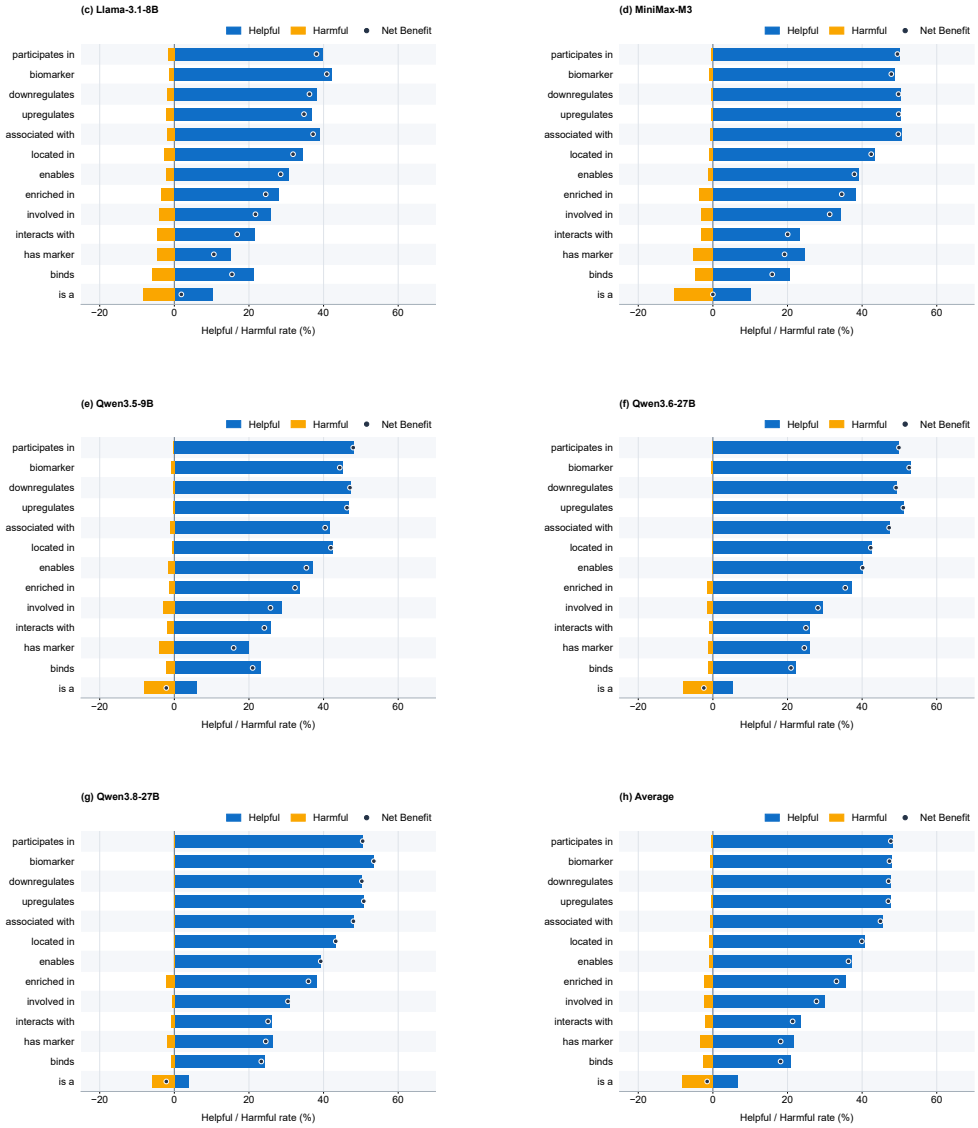


Figure 8: Relation-wise decomposition of context-induced prediction changes under the path setting ($P \rightarrow P + C$) across seven models and their average.

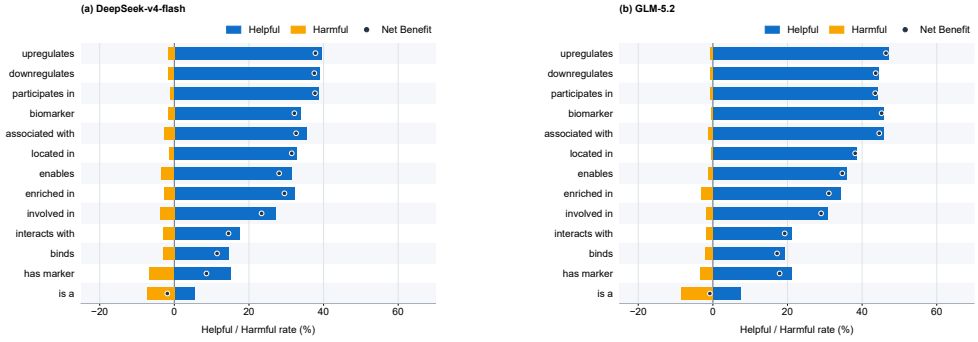




Figure 9: Relation-wise decomposition of context-induced prediction changes under the triple setting ($T \rightarrow T + C$) across seven models and their average.

For high-gain relations, helpful transitions clearly outnumber harmful transitions, showing that context improves performance mainly by correcting initially incorrect predictions. This pattern is particularly pronounced for relations such as *upregulates* and *downregulates*, which achieve net gains of 37.9 and 37.1 pp, respectively. The clearest exception is *is a*: its helpful and harmful transitions are of comparable magnitude, causing the two effects to largely cancel and resulting in a slightly negative net effect of -0.9 pp, which accounts for the negative context gain observed for *is a* in Appendix A.5.

Extending this analysis to the Path setting, we find that explicitly exposing reasoning paths consistently amplifies the helpful rates across all models, while harmful transitions do not increase correspondingly. This suggests that structured paths act not only as reasoning routes but also as a scaffold that helps models exploit contextual evidence more effectively, without a corresponding increase in context-induced errors.

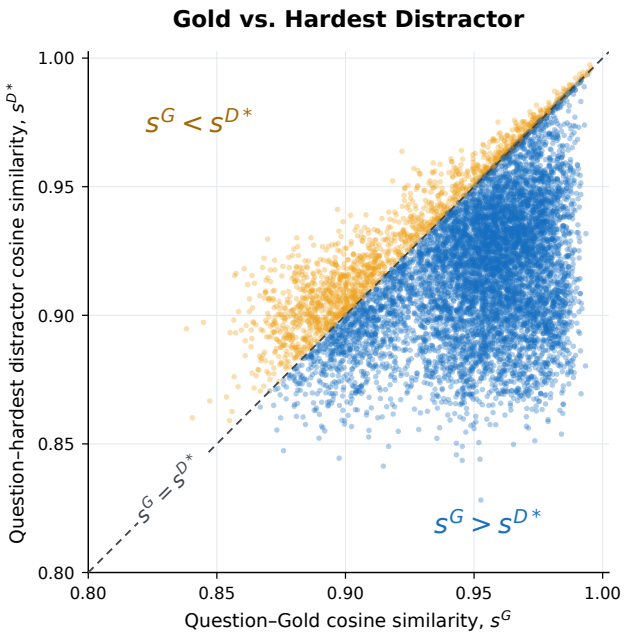


Figure 10: Question–context cosine similarity for the correct candidate and the hardest distractor across QA instances. Each point represents one QA instance.

A.7 Correct–Distractor Context Discriminability

The previous analyses examine how context affects model reasoning across different relation types. Further evaluation assesses whether the gold candidate's context aligns better with the question than competing contexts. This analysis is performed directly in the embedding space and is independent of model prediction outcomes.

For each candidate, the deduplicated evidence sentences of its complete branch, including the answer-extension edge, are concatenated into a single text sequence and encoded into a context embedding, while the question is encoded separately in the same embedding space. We use the biomedical sentence encoder S-PubMedBert-MS-MARCO, a PubMedBERT model (Gu et al., 2021) fine-tuned for query–passage retrieval in the Sentence-BERT framework (Reimers & Gurevych, 2019), with a maximum sequence length of 350 tokens (87 gold contexts exceed this length and are truncated). An instance is eligible if its question is nonempty, its gold candidate has evidence text, and at least one competing candidate has evidence text; 8,616 Medium-level instances with 9,196 distractors satisfy this condition. Because the encoder is trained for retrieval, this analysis approximates how easily a retrieval-style similarity heuristic could identify the gold context. For each QA instance i, we compute the semantic similarity between the question Q_i and the context of the correct candidate G_i. Among all distractor candidates $\{D_{ij}\}$ in the same QA instance, the one with the highest question–context similarity is selected as the hardest distractor. Accordingly, we compute

$$s_i^G = \mathrm{sim}(Q_i, G_i), \tag{11}$$

$$s_i^{D^*} = \max_j \mathrm{sim}(Q_i, D_{ij}), \tag{12}$$

where s_i^G denotes the question–context similarity for the correct candidate, and $s_i^{D^*}$ denotes that of the hardest distractor. s_i^G and $s_i^{D^*}$ are compared for each QA instance. A point below the diagonal $s_i^G = s_i^{D^*}$ indicates that the gold context is more similar to the question than every available distractor context, whereas a point above the diagonal indicates that at least one distractor is more semantically aligned with the question. The resulting similarity comparison across QA instances is shown in Fig. 10.

Table 14: Question–context discriminability of the gold context by Medium-level subtype (95% bootstrap CIs over QA instances). Margin is $s^G - s^{D^*}$.

Subtype	#QA	#Distr.	Top-1 (%)	Macro AUC	Mean margin
Convergent	3,155	3,183	86.94 [85.74, 88.08]	0.880 [0.868, 0.890]	0.0282
Divergent	3,630	4,161	89.86 [88.87, 90.83]	0.913 [0.904, 0.922]	0.0412
Hidden-intermediate	1,831	1,852	58.33 [56.09, 60.62]	0.588 [0.566, 0.610]	0.0057
Overall	8,616	9,196	82.09 [81.28, 82.89]	0.832 [0.824, 0.839]	0.0289

Across 8,616 QA instances with 9,196 distractor candidates, the correct candidate's context ranks top-1 in similarity for 82.09% of cases, exceeding all available distractor contexts. The within-QA macro AUC reaches 0.8318, computed as the average fraction of distractor contexts ranked below the correct candidate within each QA instance, with ties assigned a weight of 0.5. Top-1 is strict, i.e., ties with a distractor count as failures. Confidence intervals are obtained from 10,000 bootstrap replicates that resample QA instances.

Interpretation. A high overall top-1 rate is expected by construction, since the gold candidate is selected for having context that is sufficient and discriminative for the question (Section 3.2). The more informative finding is the variation across subtypes. For hidden-intermediate questions, the gold context is barely more similar to the question than the hardest distractor (mean margin 0.0057, AUC 0.588, close to chance), because both candidates share the same upstream and downstream anchors and their passages describe closely related biology. If models relied on surface similarity between question and context, hidden-intermediate questions should therefore benefit the *least* from context. The opposite is observed: they obtain the largest net benefit of all subtypes (44.27 pp, versus 21.46 pp for convergent and 14.71 pp for divergent questions; Section 4.4). The context gain is thus largest precisely where a similarity heuristic is least effective, which is consistent with models using the relational content of the context, such as which specific partners, conditions, or measurements are reported, rather than its overall resemblance to the question. Case (a) in Appendix A.10.2 illustrates this mechanism.

A.8 Context Swap Placebo Checks

The Context Swap condition differs from the unswapped condition in two ways besides the swapped passage: it uses a system prompt that warns that the supplied context may be mismatched or misleading, and its instances were rendered separately. Two controls that require no additional model calls separate the content effect from the prompt and rendering effect.

The first control uses the 2,806 of the 19,985 swapped instances whose donor context is identical to the original after normalization, for example repeated metadata values such as {`"reviewed": true`}. These null swaps keep the warning prompt and the rendering while changing no content, so the accuracy change they induce measures the prompt and rendering effect. The remaining 17,179 informative swaps additionally change the content. On the five larger models the null-swap effect under $P + C$ ranges from -0.93 to $+0.39$ pp, while the difference-in-differences estimate of the content effect averages 15.25 pp under C, 5.96 pp under $T + C$, and 12.02 pp under $P + C$. The null-swap instances are not ones that ignore context: their context gain under $T + C - T$ exceeds that of the informative group in every model, for example 33.25 versus 29.31 pp for Qwen3.8-27B. The two smaller models do show non-negligible null-swap effects under $T + C$ (14.43 pp for Llama-3.1-8B and 11.83 pp for Qwen3.5-9B), so their $T + C$ swap drop largely reflects the prompt and the rendering; their $P + C$ content effects remain 8.85 and 10.31 pp. Anchor-cluster intervals preserve the contrast between replacing a gold-path edge and a competitor-specific edge: 25.51 [23.94, 27.52] pp under C, 15.26 [14.36, 16.52] pp under $T + C$, and 19.84 [18.59, 21.47] pp under $P + C$.

The second control splits the informative swaps by the role of the replaced edge (Table 15). Replacing a competitor-specific edge leaves accuracy within 1.91 pp of the unswapped condition in every setting, even though these instances share the warning prompt and the rendering with all other swapped instances; the drop therefore follows the content of the replaced passage rather than the experimental apparatus. Replacing gold-path support costs 23.83 to 27.18 pp. The lost accuracy reflects a wrong branch decision rather than copying the donor: under $T + C$ and $P + C$, the tail name of the donor edge appears in at most 1.6% of answers. A pattern audit of the rationales finds an explicit statement that the context is inconsistent or misleading in less than 1.7% of informative swaps for five of the seven models, and at most 5.62% for the largest, Qwen3.8-27B under $P + C$, so models rarely flag the mismatch even when the prompt invites it.

Table 15: Context Swap effect by role of the replaced edge. Drop is the same-instance accuracy difference between the unswapped and the swapped condition, in pp; negative values mean the swap raises accuracy. Gold-support (selector) restricts to edges that a lexical selector used to identify the branch.

Replaced edge	**#**	C	$T+C$	$P+C$
distractor-extension	3,600	-0.15	1.29	0.03
distractor-support	4,047	-1.03	1.91	-0.01
gold-extension	3,683	14.98	4.36	5.41
gold-support	5,258	25.80	23.83	27.18
gold-support (selector edge)	3,294	34.44	19.29	23.44

Table 16: Branch-selection accuracy of non-LLM selectors over all 20,000 Medium-level instances, in percent, with the average LLM accuracy under $P+C$ as a reference. Numeric match is evaluated on the 11,779 instances whose question contains a number.

Selector	**Convergent**	**Divergent**	**Hidden-int.**	**All**
random	49.95	47.66	49.97	49.30
position (Path A)	50.25	47.88	49.49	49.29
evidence count	53.55	44.68	51.93	50.41
metadata word overlap	53.25	51.66	52.84	52.64
numeric match	59.98	60.09	50.54	56.42
longest context	73.45	77.12	54.53	67.31
rationale only	84.08	88.61	53.01	73.57
evidence sentence only	88.19	94.25	54.13	76.98
BM25	89.21	94.30	54.67	77.54
lexical overlap	88.72	94.11	57.11	78.25
IDF-weighted overlap	88.84	94.18	57.27	78.37
logistic regression	89.17	94.42	56.76	78.35
LLM $P+C$ (reference)	87.49	84.64	95.06	89.55

A.9 Non-LLM Shortcut Selectors on the Full Medium Set

The encoder control of Appendix A.7 covers 8,616 instances with a distractor. We repeat the same question on all 20,000 Medium instances using selectors that never call a language model. Each selector scores every candidate branch from the question and the branch text and picks the highest-scoring branch, with ties broken uniformly at random. The lexical score is the number of shared content words after removing stop words, pure numbers, and the words of displayed entity names, which serve as identifiers rather than as evidence. The logistic selector combines the same features under an anchor-grouped five-fold split.

Position, evidence count, and metadata word overlap stay close to the 49.30% baseline, at 49.29%, 50.41%, and 52.64%; numeric match and longest context reach higher values, 56.42% and 67.31%, still far below the text-overlap selectors, which reach 88.2 to 94.3% on convergent and divergent questions and match or exceed the LLM accuracy under $P+C$ (87.49% and 84.64%); on those two subtypes surface matching alone identifies the branch, and the remaining model errors there are dominated by hop errors rather than branch errors (Appendix A.10). The picture inverts on hidden-intermediate questions, the subtype where context matters most. There every shortcut stays at or below 57.27%, close to the 49.97% random baseline, while the LLMs reach 83.68 to 95.06%. Layering the model results on the selectors gives the same ordering: where no selector picks a unique branch the context gain is 60.97 [52.00, 67.26] pp, and where the combined selector is wrong it is 32.26 [28.68, 33.84] pp, so the gain does not shrink on the instances that surface matching cannot solve.

A.10 Cluster-Robust Inference, Prompt Length, and Agreement

Cluster-robust intervals. Medium-level questions share anchor entities, so questions are not independent and the per-question McNemar test overstates precision. Resampling the 4,281 anchor clusters with a Poisson bootstrap (1,000 replicates) gives 95% quantile intervals for the average paired difference across models (Table 17). The overall $P+C-P$ interval widens by 10.4 times, from [27.79, 28.67] to [23.05, 32.27], and the width of the overall $T+C-T$ interval grows by

Table 17: Paired differences in accuracy (pp) with per-question and anchor-cluster 95% intervals. Clusters are the first head entity of the gold path (4,281 clusters); the bootstrap resamples clusters with Poisson weights.

Comparison	Point	Per-question	Anchor-cluster
$P + C - P$, overall	28.20	[27.79, 28.67]	[23.05, 32.27]
$T + C - T$, overall	22.71	–	[18.69, 25.99]
$P + C - P$, convergent	21.46	–	[19.19, 23.61]
$P + C - P$, divergent	14.71	–	[13.42, 16.05]
$P + C - P$, hidden-intermediate	44.27	–	[41.06, 45.61]
$T + C - T$, random context	24.38	[23.62, 25.10]	–
$P + C - P$, random context	23.63	[22.85, 24.33]	–
Swap, $P + C$ drop	11.93	–	[11.49, 12.48]
Structural, $T + C - T_{\text{deg}} + C$	9.10	–	[5.50, 12.29]
Structural, $P + C - P_{\text{deg}} + C$	7.64	–	[5.03, 9.84]
Structural, $P - P_{\text{deg}}$	0.86	–	$[-0.48, 2.50]$
Metadata − evidence, C	-18.81	–	$[-20.44, -16.03]$
Metadata − evidence, T	22.65	–	[17.01, 25.70]
Metadata − evidence, P	30.43	–	[22.71, 34.46]
Non-ontology − ontology, C	-1.64	–	$[-2.93, -0.18]$
Non-ontology − ontology, T	4.45	–	[3.41, 5.56]
Non-ontology − ontology, P	8.33	–	[7.14, 9.62]
Quantity, natural $T + C$	18.82	–	[14.96, 21.99]
Quantity, natural $P + C$	28.94	–	[23.46, 33.12]
Quantity, constructed $T + C$	9.02	–	[7.42, 10.58]
Quantity, constructed $P + C$	12.52	–	[10.66, 14.42]

a factor of 9.3; after splitting by subtype the width ratio is 1.9 to 3.5 times. Every interval for the main contrast stays strictly positive, the subtype ordering is unchanged, and the smallest per-model lower bound for $P + C - P$ is 17.76 pp. The interval for $P - P_{\text{deg}}$ contains zero, which is expected because that contrast averages two opposite perturbations; Appendix A.3.1 separates the two roles.

The same resampling changes the reading of the RQ2 comparisons once structure is removed. Under C only, metadata minus evidence is -18.81 pp $[-20.44, -16.03]$ and non-ontology minus ontology is -1.64 pp $[-2.93, -0.18]$, both the reverse of their signs under T and P; the reversal is therefore not an artifact of correlated items.

Prompt length. Median prompt length is 220 tokens for Q, 412 for C, 315 for T, 493 for $T + C$, 346 for P, and 523 for $P + C$. The random passages used in the length control are not shorter: the median is 462 tokens against 495 for matched context under $T + C$, and 489 against 522 under $P + C$. Binning $P + C - P$ by the number of tokens that context adds gives gains of 33.75 [29.01, 36.69], 26.02 [21.18, 29.48], 26.02 [22.08, 29.18], and 19.29 [17.16, 21.50] pp from the first to the fourth quartile, so the gain falls rather than rises as the context grows longer.

Invalid outputs. The fraction of inputs with no extractable answer is 1.54 to 3.28% for Llama-3.1-8B, 0.60 to 1.65% for Qwen3.5-9B, at most 0.41% for MiniMax-M3, and at most 0.08% for the remaining models. Restricting both arms to instances that produced a valid answer, the average gain is 28.34 pp [23.18, 32.39] for $P + C - P$ and 22.77 pp [18.75, 26.04] for $T + C - T$, so counting malformed outputs as incorrect does not drive the result.

Agreement. Fleiss kappa across the seven models is 0.484 for Q, 0.762 for C, 0.402 for T, 0.329 for $T + C$, 0.394 for P, and 0.383 for $P + C$. The context-only setting is the most consistent across models, and the fraction of instances failed by at least four models falls from 35.69% under P to 5.88% under $P + C$.

Position bias. Branch order is fixed per instance by hashing the QA identifier. On the 19,521 two-branch questions, P reaches 72.39% when the gold path is rendered first and 50.95% when it is rendered second, so without context the task is close to a random choice plus a position bias; adding context narrows the gap to 2.3 pp (90.77% versus 88.52%).

Error destinations. Adding context changes not only the number but also the kind of error (Table 18). Under P, 80.45% of errors name the answer of a competing branch and only 16.16% name another graph entity. Under $P + C$, competing-branch errors fall to 41.03% and errors that name another graph entity rise to 47.58%, so the remaining failures are concentrated at the hop rather than

Table 18: Destination of incorrect answers, as a percentage of all incorrect answers, averaged over the seven models and restricted to the first answer. Competitor means the final answer of a competing branch; graph-entity means another entity that occurs in the shown graph; off-graph means an entity outside the graph.

Condition	Competitor	Graph-entity	Off-graph	No answer
Q	1.29	7.71	90.92	0.08
C	5.89	23.44	70.52	0.15
T	72.50	24.86	2.63	0.01
$T + C$	49.09	43.84	6.93	0.14
P	80.45	16.16	3.37	0.01
$P + C$	41.03	47.58	10.93	0.46

Table 19: Evidence coverage for eight high-coverage relation types.

Relation	Usable contexts	Unique evidence texts	Evidence coverage
biomarker	13,544	12,107	99.27%
indicated_for	16,440	14,387	64.88%
genomic_alterations	10,965	10,329	99.45%
is_a	23,592	15,534	97.23%
part_of	6,519	5,890	95.18%
develops_from	992	945	94.12%
positively_regulates	975	974	100.00%
negatively_regulates	986	985	100.00%
Total	74,013		

at the branch decision; the shift is largest for divergent questions, where graph-entity errors reach 61.30%.

A.10.1 Coverage and Dispersion of Relation-Level Evidence

To characterize the textual evidence layer beyond the aggregate coverage in Figure 6, we audit evidence coverage per relation type. Evidence coverage is the fraction of unique head–relation–tail edges of a relation with at least one nonempty evidence sentence, so multiple evidence records on the same edge do not inflate coverage. Table 19 lists eight high-coverage relations (at least 900 usable evidence contexts and at least 50% coverage) that span clinical and disease association, ontology and development, and regulation.

Relations whose evidence is sparse at the edge level, such as *interacts_with* (13.28%), *binds* (3.93%), *enables* (20.73%), and *associated_with* (44.20%), are typically complemented in the KG by structured metadata (e.g., BioGRID experimental systems and source databases), so their edges can carry context even without an evidence sentence; the overall context coverage of 74.98% counts both channels. These statistics are descriptive properties of the evidence layer and are independent of any model prediction.

A.10.2 Case Studies

Three instances from the evaluated sets illustrate the mechanisms discussed in Sections 4.3 and 4.4; their labels follow the main-text scoring, and the cases are not used to estimate effect sizes.

(a) Evidence resolves a hidden intermediate. Instance `medium_hidden_intermediate_04603` asks: *"Two genes are structurally linked to TRAF2 via an 'interacts with' relationship and subsequently linked to 'tumor necrosis factor receptor activity' via an 'enables' relationship. One of these genes is associated with a protein that interacts with TRAF1, TRAF2, and TRAF3, and induces NF-kappaB activation via TRAF2. Which protein is encoded by the gene associated with this protein?"* Both TNFRSF18 and TNFRSF4 satisfy the full topology (TRAF2 $\rightarrow$ gene $\rightarrow$ receptor activity, plus an *encodes* extension), so structure alone cannot separate them, and neither candidate name appears in the question. The context of TRAF2–*interacts_with*–TNFRSF18 contains the sentence "AITR associates with TRAF1 (TNF receptor-associated factor 1), TRAF2, and TRAF3, and induces nuclear factor (NF)-kappaB activation via TRAF2" (PMID:10037686), which matches every condition in the question, whereas the corresponding TNFRSF4 passage ("Ox40 interacted with TRAF3 and TRAF2", PMID:9418902)

matches only part of it. Qwen3.8-27B, Qwen3.6-27B, and GLM-5.2 all answer TNFRSF4's protein under both T and P, and all switch to the gold answer, tumor necrosis factor receptor superfamily member 18, under both $T + C$ and $P + C$. The rationale of GLM-5.2 under P attributes the TRAF1/2/3 clue to TNFRSF4 from parametric memory, whereas under $P + C$ it attributes it to TNFRSF18 in line with the supplied evidence. This is exactly the hidden-intermediate situation in which question–context similarity is least discriminative (Table 14), yet the specific conjunction of reported partners decides the answer.

(b) Numerical metadata distinguishes structurally identical candidates. Instance `medium_hidden_intermediate_10655` asks: *"Pegivirus C is reported to downregulate a specific gene that is located in chromatin. Experimental data for this gene in peripheral blood mononuclear cells shows a log2 fold change of approximately* -1.25 *and an adjusted p-value of 0.0115. Which cell type is this gene enriched in?"* Both FER and ZNF76 are downregulated by Pegivirus C and *located_in* chromatin, but their answer extensions lead to different cell types (basophil and gamma-delta T cell). Both differential-expression records refer to peripheral blood mononuclear cells in the same study (GSE131504, via VExD), so the textual setting does not separate them either. Only the numerical metadata does: the FER record reports $\log_2 \mathrm{FC} = -1.247935$ and adjusted $p = 0.01148045$, whereas the ZNF76 record reports -1.412464 and 0.00332607. GLM-5.2 answers gamma-delta T cell under T and P and the gold answer, basophil (CL:0000767), under $T + C$ and $P + C$, with the rationale "FER matches the given log2 fold change (-1.25) and adjusted p-value (0.0115), is located in chromatin, and is enriched in basophils." This case illustrates the RQ2 finding that metadata provides precise discriminative signals once structure has narrowed the candidate space.

(c) A mismatched passage redirects the answer. Instance `medium_convergence_04928` asks: *"EP300 and RBPJ share multiple interaction partners. One such partner is reported to colocalize in vivo and form a complex with the transcriptional coactivator p300. Which protein is encoded by this shared interaction partner?"* POLB and NFATC1 are both shared interaction partners of EP300 and RBPJ. The context of EP300–*interacts_with*–POLB states "Polbeta colocalized in vivo and formed a complex with the transcriptional coactivator p300" (PMID:12453427), which matches the question, while NFATC1 carries passages describing other p300 interactions. In the corresponding Context Swap instance (`hard_context_swap_04292`), only this passage is replaced by a type-matched donor passage from RANBP9–*interacts_with*–CDKN1A ("Mechanistically, RanBPM protein physically interacts with p21, …", PMID:37676377); the question, the edge set, and the gold answer are unchanged. Qwen3.8-27B, Qwen3.6-27B, and GLM-5.2 answer DNA polymerase beta under both $T + C$ and $P + C$ on the original instance, and all switch to the NFATC1 protein under both settings after the swap. The models do not copy an entity from the donor passage; instead, the Qwen3.8-27B rationale states that "POLB is not associated with p300 in the provided context" and selects the remaining p300-related alternative. This shows that the models read and weigh the edge-specific context rather than relying on the presence of any interaction-like text, which is why type-matched swapping is a stronger perturbation than random replacement.

B Data Availability and Licensing

The benchmark QA datasets and lightweight evaluation code are publicly available at `https://huggingface.co/datasets/ImmunoGraph-Bench/ImmunoGraphBench2027`. The release retains available source identifiers and citations. Third-party content remains subject to the original providers' terms; this release does not grant additional rights to such content.