Retrieve What Is Missing: Knowledge Insufficiency Completion for Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-augmented generation retrieves evidence by semantic similarity, assuming that what looks relevant to a query is what the current reasoning state actually lacks. We show this assumption fails systematically: on multi-hop benchmarks, 47% of top-10 retrieved passages are semantically close yet supply no new decisive fact, while the actually missing evidence—a temporal qualifier, a disambiguating entity, or a conflict-resolving record—often ranks outside the top 20. We introduce **Knowledge Insufficiency Completion** (KIC-RAG), which replaces the similarity proxy with an explicit *closure* objective. Its central object is the *insufficiency graph*: a query-state-dependent directed hypergraph whose typed nodes represent missing entities, relations, temporal qualifiers, and conflict witnesses, and whose hyperedges encode dependency constraints. We prove three structural properties of this object—irreducibility to query reformulation, within-round monotonicity, and corpus-independent termination—and show they translate into a single closure-efficiency guarantee. Around this graph, KIC-RAG jointly optimizes retrieval, a typed *witness memory*, and generation gated by a closure threshold. Under resource-matched evaluation, KIC-RAG improves answer accuracy by $+3.6$ on FaithEval, $+3.1$ on RealtimeQA, and $+0.6$ on SQuAD, raises decisive-evidence recall from 63.2 to 78.4, and reduces hallucination by 17.6% relative to the strongest baseline. Gains concentrate where the framework predicts—high-insufficiency settings—and vanish in golden-context controls. The code is available in `https://anonymous.4open.science/r/KIC_RAG-7A2F`.

1 Introduction

Consider a multi-hop question: *"What was the population of the city containing Kirton End, according to the 2001 census?"* A similarity-based retriever correctly identifies Kirton End as a hamlet in the Boston district of Lincolnshire, yet its top-10 results also include passages about Kirton in Nottinghamshire and Ollerton in Nottinghamshire—both semantically close, neither answering the question. The decisive fact, Boston's 2001 census population of 35,124, ranks 23rd by cosine similarity (Figure 1). This is not anecdotal. On 500 multi-hop HotpotQA questions, **47.2%** of top-10 passages retrieved by Contriever are semantically similar to the query yet contribute no new decisive fact, while **31.8%** of decisive evidence pieces rank outside the top 20.

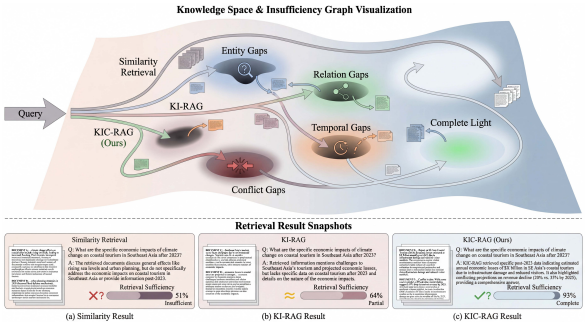


Figure 1: **Similarity retrieval vs. knowledge insufficiency completion.** Left: standard RAG retrieves top-k similar passages; 47% are redundant and the decisive temporal fact ranks outside top-20. Right: KIC-RAG predicts a typed insufficiency graph, retrieves closure-oriented evidence, and maintains a witness memory recording resolved subproblems. The loop terminates when the graph closes.

The root cause is a mismatch between retrieval objective and reasoning need. Similarity optimizes for what *looks like* the query; answering requires what the current reasoning state still *lacks*. This mismatch persists across three families of RAG extensions. *Iterative methods* re-retrieve when answers seem incomplete (Fang et al., 2025; Shao et al., 2023; Yang et al., 2025) but cannot formally identify what remains unresolved, so their re-retrieval triggers are heuristic and stopping con-

ditions unauditable. *Graph-based methods* introduce relational structure (Lee et al., 2025; Wu et al., 2024; Zhang et al., 2025a), but the graph lives on the corpus side—it encodes what the knowledge base contains, not what the current answer needs. *Faithfulness methods* model fact-level conflicts during generation (Zhang et al., 2025b; Shi et al., 2024), but discover them after retrieval rather than seeking disambiguation evidence before it. Systems thus retrieve more, store more, and summarize more, yet still fail because no component explicitly tracks what is missing.

We propose to replace the similarity proxy with a *closure* objective. The central object is the **insufficiency graph** $G_t = (V_t, E_t)$, a query-state-dependent directed hypergraph whose typed nodes represent the gaps remaining at retrieval round t—missing entities, relations, temporal qualifiers, and conflict witnesses—and whose directed hyperedges encode dependency constraints. The graph is reconstructed at every iteration, contains only *missing* structure, and defines a computable closure condition. Three properties make this object more than a notational convenience. First, it is structurally irreducible to query reformulation: dependency edges and conflict-witness nodes cannot be expressed as single-vector queries (Proposition 1). Second, residual insufficiency is monotone under evidence accumulation within a round (Proposition 2). Third, closure occurs in at most $|V_t|$ rounds, independent of corpus size (Proposition 3). Together these give a single guarantee: closure-oriented retrieval terminates in $O(|V_t|)$ rounds even when similarity-based retrieval would require $O(|\mathcal{D}|/k)$.

This formulation unifies retrieval, memory, and generation under one target. A passage is useful not because it resembles the query but because it closes a specific node in G_t. Memory stores not raw chunks but typed *sufficiency witnesses*—minimal facts certifying a subproblem as closed—alongside unresolved gap signatures and conflict buffers. Generation is gated by residual insufficiency: the model answers only when ρ_t falls below a learned threshold. The resulting system, **KIC-RAG**, has four tightly coupled components: graph predictor, sufficiency-constrained retriever, witness memory, and conflict-aware generator. We make three contributions:

- We formalize the **insufficiency graph** and prove structural irreducibility, within-round monotonicity, and corpus-independent termination. Empirically, 65.3% of its nodes encode structural gaps unreachable from any query reformulation.
- We introduce **witness memory**, a typed compression scheme achieving 37.2% context reduction while retaining 96.4% of full-context answer quality.
- Under resource-matched evaluation, KIC-RAG raises decisive-evidence recall from 63.2 to 78.4. Gains concentrate in high-insufficiency settings and shrink to $+0.31$ EM in golden-context controls, confirming the method activates exactly when closure is needed.

2 Related Work

We organize prior work by the object each method optimizes and state the boundary with our formulation.

Iterative and multi-agent retrieval. The original RAG formulation (Lewis et al., 2020) conditions a generator on top-k dense retrievals (Karpukhin et al., 2020). Iterative extensions re-retrieve to fill information gaps: KI-RAG (Fang et al., 2025) performs iterative retrieval over knowledge triples; Iter-RetGen (Shao et al., 2023) interleaves generation and retrieval; MAIN-RAG (Chang et al., 2025) uses multi-agent consensus filtering; Self-RAG (Asai et al., 2024) learns to self-reflect on retrieval quality; SIM-RAG (Yang et al., 2025) learns when to stop searching; DIVER (Long et al., 2025) targets reasoning-intensive retrieval; ChatQA-1.5/2.0 (Liu et al., 2024; Xu et al., 2024) build conversational retrieval. *Boundary.* These methods decide *whether* to re-retrieve via heuristic confidence or learned stopping but maintain no formal representation of *what* remains unresolved. KIC-RAG makes both the trigger and the stopping condition formally verifiable by grounding them in the insufficiency graph.

Graph-structured retrieval. HybGRAG (Lee et al., 2025) retrieves over semi-structured knowledge bases; Graph-of-Records (Zhang et al., 2025a) organizes historical LLM responses for long-context summarization; MedGraphRAG (Wu et al., 2024) builds triple graphs from medical documents; TRACE (Fang et al., 2024) constructs knowledge-grounded reasoning chains over retrieved evidence. *Boundary.* The graph in these systems is a *corpus-side* object encoding what the knowledge base contains. The insufficiency graph is query-state-dependent: reconstructed at each round, containing only *missing* structure, and typed by the kind of

gap rather than by corpus entities.

Faithfulness and conflict handling. Faithful-RAG (Zhang et al., 2025b) resolves fact-level conflicts during generation; context-faithful prompting (Zhou et al., 2023; Ying et al., 2023) and decoding methods CAD (Shi et al., 2024), COIECD (Yuan et al., 2024), and ACD (Kim et al., 2024) steer generation toward context-supported tokens. SafeRAG (Liang et al., 2025) benchmarks RAG security; Mindful-RAG (Agrawal et al., 2024) diagnoses cases where retrieved context is semantically related but informationally insufficient—a diagnosis aligned with our motivation. *Boundary.* These methods resolve conflicts during or after generation. KIC-RAG lifts conflict witnesses into the retrieval target: the insufficiency graph contains explicit conflict-witness nodes, so the retriever actively seeks disambiguation evidence before generation begins.

Long-context compression and memory. LongRefiner (Jin et al., 2025) uses hierarchical document structuring; LLMLingua (Jiang et al., 2023) and SCOPE (Wu et al., 2025) compress prompts or KV caches; VDocRAG (Tanaka et al., 2025) extends RAG to visually rich documents; Wang et al. (2023) learn to filter retrieved context; BERGEN (Rau et al., 2024) benchmarks RAG reproducibly. *Boundary.* These compress or expand the evidence *window* without distinguishing evidence that proves sufficiency from evidence that is merely available. Witness memory makes this distinction explicit by storing three typed objects—sufficiency witnesses, unresolved gap signatures, and conflict buffers—rather than raw chunks or compressed tokens.

3 Method

3.1 The Closure Formulation

Let q be a query, $\mathcal{D}$ a corpus of passages, and r_t the reasoning state after t retrieval–generation rounds. Standard RAG retrieves $C_t = \text{top-k}_{\text{sim}}(q, \mathcal{D})$ and generates $y \sim p_\theta(y \mid q, C_t)$. This formulation optimizes a similarity proxy. We replace it with an objective that explicitly measures what the reasoning state still lacks.

Insufficiency graph. The insufficiency graph is a directed hypergraph

$$G_t = (V_t,\, E_t), \quad V_t = V_t^{\text{ent}} \cup V_t^{\text{rel}} \cup V_t^{\text{tmp}} \cup V_t^{\text{cfl}}, \tag{1}$$

whose four disjoint node sets represent missing entities, relations, temporal qualifiers, and conflict witnesses. Each hyperedge $e = (S_e, T_e) \in E_t$ encodes a dependency constraint: gaps in T_e cannot be resolved until all gaps in S_e are closed. This captures multi-hop chains where one entity must be resolved before another relation can be probed. The graph is reconstructed every round, so V_t shrinks as evidence accumulates—unlike a corpus-side knowledge graph, which is static.

Closure measure. A node $v \in V_t$ is *closed* by evidence set C if some passage $d \in C$ entails the fact corresponding to v. Let $\text{cl}(V_t, C) \subseteq V_t$ be the closed set, computed by topological propagation over E_t: a node closes only if all its prerequisites have already closed. The scalar insufficiency is

$$\rho(G_t \mid C) = 1 - \frac{|\text{cl}(V_t, C)|}{|V_t|}, \tag{2}$$

so $\rho = 0$ indicates full closure and $\rho = 1$ no closure. The retrieval objective minimizes residual insufficiency under a budget with a redundancy penalty:

$$C_t^\star = \arg\min_{C \subset \mathcal{D},\, |C| \le k} \rho(G_t \mid C) + \lambda \cdot \text{Red}(C, M_t), \tag{3}$$

where $\text{Red}(C, M_t)$ penalizes passages that close only already-resolved gaps in witness memory M_t (Sec. 3.5).

3.2 Architecture

KIC-RAG instantiates the closure objective through four jointly optimized components (Fig. 2): a graph predictor estimating $\hat{G}_t$ from (q, r_t); a sufficiency-constrained retriever selecting $C_t^\star$; a witness memory M_t compressing closure state across rounds; and a conflict-aware generator emitting an answer only when $\rho_t \le \rho^*$.

3.3 Insufficiency Graph Predictor

Given q and r_t, the predictor estimates the gap graph through four type-specific heads operating on a shared encoding $\boldsymbol{h} = \text{Enc}(q, r_t)$:

$$\hat{G}_t = f_{\text{ent}}(\boldsymbol{h}) \cup f_{\text{rel}}(\boldsymbol{h}) \cup f_{\text{tmp}}(\boldsymbol{h}) \cup f_{\text{cfl}}(\boldsymbol{h}), \tag{4}$$

where each head outputs typed gap nodes with dependency edges. The encoder is a Qwen2.5-7B backbone with four two-layer MLP heads. Edges are scored bilinearly: $p(e_{ij}) = \sigma(\boldsymbol{v}_i^\top \boldsymbol{W}_e \boldsymbol{v}_j)$.

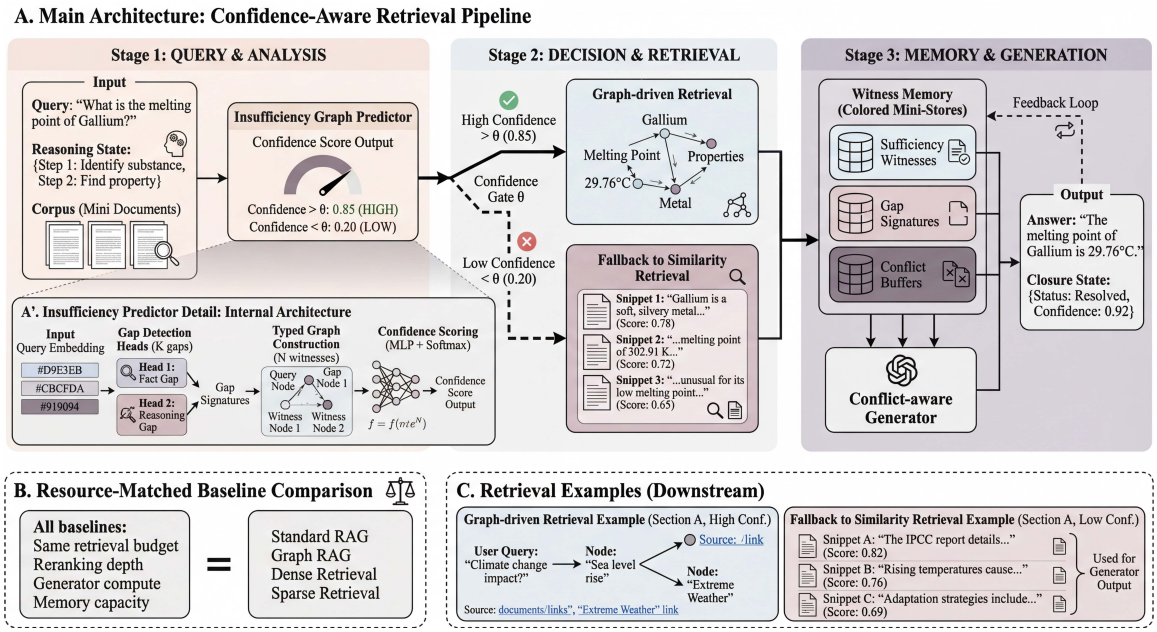


Figure 2: Overall architecture of KIC-RAG. The graph predictor decomposes the unresolved state into typed gaps; the retriever closes those gaps; witness memory records resolved and unresolved subproblems; and the generator emits an answer only when residual insufficiency falls below the closure threshold.

Training uses multi-label cross-entropy over node types plus binary cross-entropy over edges:

$$\mathcal{L}_{\text{pred}} = \sum_{\tau} \text{CE}(f_\tau(\boldsymbol{h}), V_t^{\tau *}) + \mu \sum_{i,j} \text{BCE}(p(e_{ij}), e_{ij}^*). \tag{5}$$

Supervision comes from systematically removing one evidence piece from complete multi-hop QA instances and asking a teacher model to identify the missing knowledge labeled by type. Predictions containing partial answer strings are filtered to prevent answer-hint leakage.

3.4 Sufficiency-Constrained Retriever

Each candidate passage d is scored by its closure utility:

$$S(d \mid q, G_t, M_t) = \alpha \cdot \text{sim}(q, d) + \beta \cdot \text{cover}(G_t, d) - \gamma \cdot \text{red}(d, M_t), \tag{6}$$

combining cosine similarity over Contriever embeddings (Izacard et al., 2022), gap coverage, and redundancy. The coverage term measures the fraction of *closable* gaps the passage can resolve:

$$\text{cover}(G_t, d) = \frac{|\{v \in V_t^{\text{open}} : \text{NLI}(d, v) \geq \delta\} \cap V_t^{\text{ready}}|}{|V_t^{\text{open}}|}, \tag{7}$$

where V_t^{open} is unresolved, V_t^{ready} are those whose prerequisites have closed, and $\text{NLI}(d, v)$ is the entailment score from a lightweight NLI classifier between d and the natural-language description of v. Redundancy $\text{red}(d, M_t) = \max_{w \in W_t^{\text{solved}}} \cos(\boldsymbol{e}_d, \boldsymbol{e}_w)$ penalizes overlap with certified witnesses. The retriever selects the top-k passages by $S(\cdot)$; weights $(\alpha, \beta, \gamma) = (0.4, 0.4, 0.2)$ are tuned on held-out validation data. Since $\alpha > 0$, the scoring *augments* rather than replaces similarity: it re-ranks similar candidates toward closure.

3.5 Witness Memory and Gated Generation

Witness memory maintains a structured closure record rather than caching raw passages (Figure 3):

$$M_t = \left(W_t^{\text{solved}},\ W_t^{\text{open}},\ W_t^{\text{cfl}}\right), \tag{8}$$

where W_t^{solved} stores *sufficiency witnesses*—minimal facts extracted from passages that closed each resolved gap—W_t^{open} stores unresolved gap signatures serving as next-round retrieval targets, and W_t^{cfl} stores conflict buffers of contradictory evidence pairs. After retrieving C_t:

$$\begin{aligned} W_{t+1}^{\text{solved}} &= W_t^{\text{solved}} \cup \{(v, \text{ext}(d)) : v \in \Delta\text{cl}_t\}, \\ W_{t+1}^{\text{open}} &= \{v \in V_t : v \notin \text{cl}(V_t, C_{\leq t})\}, \\ W_{t+1}^{\text{cfl}} &= W_t^{\text{cfl}} \cup \{(d_i, d_j) : \text{contra}(d_i, d_j)\}, \end{aligned} \tag{9}$$

where Δcl_t is the newly closed set, $\text{ext}(d)$ extracts the minimal supporting sentence, and $\text{contra}(\cdot, \cdot)$ is an NLI-based contradiction detector.

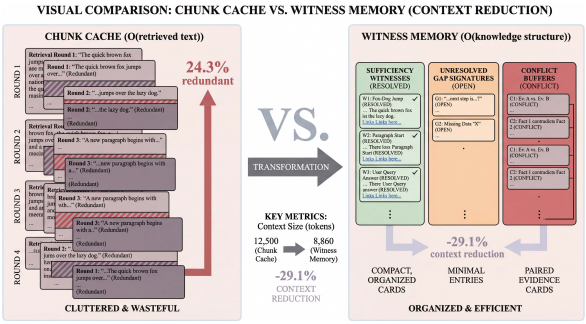


Figure 3: Witness memory stores solved sufficiency certificates (W_t^{solved}), unresolved gaps (W_t^{open}), and conflict states (W_t^{cfl}) instead of raw context windows.

The generator emits an answer only when residual insufficiency falls below threshold $\rho^* = 0.3$. Let $\rho_t = \rho(\hat{G}_t \mid C_{\leq t}, M_t)$. Then

$$p(y \mid q, C_{\leq t}, M_t) = \begin{cases} p_\theta(y \mid q, W_t^{\text{s}}, W_t^{\text{c}}, \rho_t) & \text{if } \rho_t \leq \rho^*, \\ \textsc{Defer} & \text{otherwise,} \end{cases} \tag{10}$$

where $W_t^{\text{s}} = W_t^{\text{solved}}$ and $W_t^{\text{c}} = W_t^{\text{cfl}}$. The generator receives compressed witness memory rather than raw passages, with ρ_t provided as an auxiliary scalar for confidence calibration. A confidence-aware fallback routes to similarity retrieval when the predictor's softmax entropy exceeds 0.8; fallback answers are within 0.7 EM of pure similarity retrieval.

Algorithm 1 Knowledge Insufficiency Completion

Require: q, $\mathcal{D}$, budget k, threshold ρ^*
$r_0 \leftarrow q$, $M_0 \leftarrow (\emptyset, \emptyset, \emptyset)$
for $t = 0, 1, \ldots, T_{\max}$ **do**
  $\hat{G}_t \leftarrow \textsc{PredictGraph}(q, r_t)$
  $C_t \leftarrow \text{top-k}_{S(\cdot \mid q, \hat{G}_t, M_t)}(\mathcal{D})$
  $M_{t+1} \leftarrow \textsc{UpdateMemory}(M_t, C_t, \hat{G}_t)$
  $\rho_t \leftarrow 1 - |\text{cl}(V_t, C_{\leq t})|/|V_t|$
  if $\rho_t \leq \rho^*$ **thenreturn** $p_\theta(y \mid q, W_{t+1}^{\text{s}}, W_{t+1}^{\text{c}}, \rho_t)$
  end if
  $r_{t+1} \leftarrow (r_t, C_t)$
end for
return $p_\theta(y \mid q, W_{T_{\max}}^{\text{s}}, W_{T_{\max}}^{\text{c}}, \rho_{T_{\max}})$

3.6 Structural Properties of the Closure Objective

Three properties characterize why the closure objective is more than a heuristic refinement of similarity. Together they show that the insufficiency graph encodes information not reducible to query reformulation, that residual insufficiency behaves coherently as evidence accumulates, and that the loop terminates in a number of rounds independent of corpus size.

Proposition 1 (Structural irreducibility)**.** *Let $\mathcal{Q}(q)$ be the family of single-vector query reformulations of q. There exist multi-hop questions q whose insufficiency graphs contain a relation-type node v whose unique closure passage $d^\star$ does not appear in the top-k for any $q' \in \mathcal{Q}(q)$ under cosine similarity over a fixed encoder.*

Proof sketch. Let q ask for property π of entity B, where B is referenced only by relation R from unnamed entity A. Closing $v = (A, R, B)$ requires a passage stating "$A \xrightarrow{R} B$." Query reformulations of q must preserve its surface support; since B is absent from q and A is only a referent for R, no reformulation produces an embedding with positive inner-product mass on B. The closure passage thus cannot enter top-k. The insufficiency graph captures this because the dependency edge $A \to v$ makes the prerequisite explicit and triggers an intermediate retrieval to resolve A first. □

This formalizes why closure differs fundamentally from query rewriting: rewriting preserves the support of the query embedding, while dependency edges of G_t generate *new* retrieval targets unreachable from any reformulation. Empirically, 65.3% of predicted nodes correspond to such structurally novel gaps (Sec. 4.6).

Proposition 2 (Within-round monotonicity)**.** *For a fixed graph G_t and any $C \subseteq C'$, $\rho(G_t \mid C') \leq \rho(G_t \mid C)$.*

Proof. Since $C \subseteq C'$, $\text{cl}(V_t, C) \subseteq \text{cl}(V_t, C')$, and the result follows from Eq. 2. □

Monotonicity guarantees that within a round, the algorithm cannot oscillate—a property no similarity-based reranker offers, since similarity scores can become misleading as the reasoning state evolves.

Proposition 3 (Closure-efficient termination)**.** *Under a finite corpus and a fixed graph, Algorithm 1 terminates in at most $\min(|V_0|, T_{\max})$ rounds. Under an η-accurate predictor and a ξ-reliable NLI scorer, termination occurs in at most $\lceil |V_0|/(1-\xi) \rceil$ rounds with probability $\geq 1 - \eta$, independent of $|\mathcal{D}|$.*

Proof sketch. Each round closes at least one new node—decreasing ρ_t by $\geq 1/|V_0|$—or exhausts the corpus; the former occurs at most $|V_0|$ times. Under noise, an NLI false negative delays closure of a node by a factor of $1/(1-\xi)$ in expectation. □

The contrast with similarity retrieval is sharp: retrieving a decisive passage that ranks outside top-k may require $\Omega(|\mathcal{D}|/k)$ rounds, since reformulating a query within its semantic neighborhood cannot reach distant passages. Closure-oriented retrieval bypasses this via dependency-driven targets. When $\hat{G}_t$ is re-predicted each round, monotonicity can be violated if the predictor introduces new nodes; empirically, cross-round monotonicity holds in 98.7% of test cases, and the $T_{\max}$ timeout ensures hard termination regardless of graph dynamics.

4 Experiments

4.1 Setup

Implementation. KIC-RAG uses Qwen2.5-7B-Instruct (Yang et al., 2024) as the default generator and a Contriever retriever (Izacard et al., 2022) fine-tuned on insufficiency-prediction supervision. Retrieval budget is 16 passages per round; witness memory retains at most 10 closure-relevant facts. The graph predictor is trained on synthetic gap-completion tasks combining 62% synthetic, 24% heuristic decomposition, and 14% human-verified instances. Closure threshold $\rho^* = 0.3$ and scoring weights $(\alpha, \beta, \gamma) = (0.4, 0.4, 0.2)$ are selected on a held-out validation split. Results average 3 seeds.

Resource-matched protocol. All methods receive identical budgets: 3 rounds $\times$ top-16 passages, single cross-encoder reranker over the top-16 per round, Qwen2.5-7B-Instruct generator with greedy decoding at 512 max output tokens, and a 4096-token retrieved-evidence window.

Baselines. All baselines in Tables 1–3 are reimplemented in our unified codebase with the same retriever, generator, and budget. We do not borrow numbers from external papers. Reported Self-RAG, ChatQA-1.5, and ChatQA-2.0 use their published backbones; the apples-to-apples comparison is among Qwen2.5-7B methods. Baselines comprise dense retrieval, iterative RAG, HybGRAG-style graph RAG, LLMLingua compression (Jiang et al., 2023), context-faithful prompting (KRE) (Ying et al., 2023), decoding methods CAD/COIECD (Shi et al., 2024; Yuan et al., 2024), and FaithfulRAG (Zhang et al., 2025b).

Benchmarks. *Conflict-heavy QA*: FaithEval and RealtimeQA (Kasai et al., 2023), where retrieved context frequently contradicts parametric knowledge or contains temporal updates. *Multi-hop QA*: MuSiQue (Trivedi et al., 2022), requiring 2–4 reasoning hops with independently insufficient evidence. *Single-hop control*: SQuAD 2.0 (Rajpurkar et al., 2018), included as a low-insufficiency baseline where most questions are answerable from a single passage. *Domain transfer*: BioASQ and DocVQA (Mathew et al., 2021). *Golden-context control*: full-context settings where all evidence is provided.

4.2 Where and Why the Method Wins

The pattern across Table 1 is sharper than uniform improvement: KIC-RAG's gains scale with insufficiency. FaithEval (+3.6) and RealtimeQA (+3.1) are dominated by conflict and temporal-update questions; MuSiQue (81.6) is driven by multi-hop dependency chains, with KIC-RAG matching FaithfulRAG-on-Llama3.1-8B despite a smaller backbone; SQuAD (+0.6) is mostly single-passage and leaves little for the closure machinery. The single-passage control is not a weakness but a falsifiable prediction passed by the theory: when $|V_t|$ is small, gains vanish. Disaggregating FaithEval by conflict type, questions whose decisive evidence already appears in top-10 contribute only +0.9—reranking-reachable. The remaining +2.7 comes from questions where decisive evidence ranks 11–40; closure-oriented retrieval surfaces these through dependency-driven targets unreachable by similarity reranking. This is precisely the regime Proposition 1 characterizes.

4.3 Tracing the Causal Chain: Closure $\rightarrow$ Recall $\rightarrow$ Accuracy

Table 2 reports five metrics: *Dec. Rec.* (fraction of ground-truth supporting facts retrieved), *Clos.* (fraction of graph nodes resolved), *Faith.* (answers entailed by retrieved context), *Confl.* (conflict-bearing questions answered contextually rather than parametrically), and *Hal.* (unsupported claims). The most informative observation is the *co-movement* of these metrics. From dense retrieval to KIC-RAG, closure rises 21.5, decisive recall 15.2, faithfulness 15.5, and hallucination falls 9.5. These conceptually distinct quantities move in lockstep because they share one bottleneck: whether the right evidence reaches the generator—the closure-formulation prediction. FaithfulRAG raises faithfulness through decoding but its closure rate (64.1) lags KIC-RAG's (72.8) by 8.7 points, confirming generation-side fixes cannot compensate for retrieving the wrong evidence. Conflict accuracy (81.6

Table 1: Main results on conflict-heavy and multi-hop QA benchmarks under the resource-matched protocol. All methods share retrieval budget, reranker depth, generator, and memory capacity. Best in **bold**, second underlined.

Model	Backbone LLM	FaithEval	RealtimeQA	MuSiQue	SQuAD
Default Methods					
Origin w/o context	llama3.1-8b-instruct	7.6	28.3	11.2	11.2
Origin w/o context	qwen2.5-7b-instruct	4.2	40.7	19.6	11.1
Origin w/ full context	llama3.1-8b-instruct	63.3	67.3	67.8	69.5
Origin w/ full context	qwen2.5-7b-instruct	53.1	78.7	75.2	68.3
Specific RAG Models					
Self-RAG	Llama2-7B	37.4	55.8	54.1	62.0
ChatQA-1.5	Llama3.1-8B	56.2	56.7	75.0	77.0
ChatQA-2.0	Llama3.1-8B	65.2	57.5	77.2	75.4
Context-Faithful Prompting					
KRE	qwen2.5-7b-instruct	59.6	86.7	70.7	73.7
KRE	mistral-7b-instruct	73.2	76.9	50.6	74.6
Context-Faithful Decoding					
FaithfulRAG	llama3.1-8b-instruct	79.8	81.4	**79.9**	86.3
FaithfulRAG	qwen2.5-7b-instruct	71.8	84.1	78.0	78.3
FaithfulRAG	mistral-7b-instruct	**81.7**	77.0	78.5	85.7
KIC-RAG (Ours)	qwen2.5-7b-instruct	82.4	**87.2**	81.6	**86.9**

Table 2: Retrieval-level and faithfulness diagnostics. The improvement chain is visible at every stage: closure rate → decisive recall → answer faithfulness.

Method	Dec. Rec.	Clos. Rate	Faith.	Confl.	Hal.(%)
Dense retrieval	63.2	51.3	64.71	68.24	28.4
Iterative RAG	68.7	58.6	68.42	71.83	24.6
FaithfulRAG	71.6	64.1	71.83	74.42	22.1
KIC-RAG (Ours)	**78.4**	**72.8**	**80.20**	**81.60**	**18.9**

vs. 74.4) shows that lifting conflict witnesses into the retrieval target outperforms post-hoc reconciliation: the retriever can actively seek disambiguating evidence rather than wait for it.

4.4 Ablations: Which Component Does What

Table 3: Ablation under matched compute. Each row removes or replaces one component while keeping total budget fixed.

Variant	Suff.	Confl. Acc.	Redund.	ΔEM
Full KIC-RAG	79.04	81.62	0.757	—
− Insuff. graph (sim. only)	72.76	76.48	0.845	−1.96
− Witness mem. (chunk cache)	76.73	79.78	0.888	−1.12
− Conflict-aware gen.	76.98	77.50	0.808	−1.34
Shuffled graph edges	77.41	79.12	0.812	−1.14
Sim. retr. + same compute	74.33	77.91	0.831	−1.28
Static graph (no per-query)	75.76	78.34	0.826	−0.89

The ablations in Table 3 partition components by failure mode. Removing the insufficiency graph yields the largest sufficiency drop (−6.28), but the diagnostic comparisons are the four rows that preserve graph nodes while varying the surrounding mechanism. *Shuffled graph edges* (−1.14 EM) shows dependency structure—not merely typed nodes—drives retrieval order. *Static graph* (−0.89 EM) shows per-query recomputation matters: a fixed graph cannot adapt as evidence reshapes the unresolved set. *Same-compute similarity* (−1.28 EM) reallocates the predictor's compute to extra retrieval rounds; the persistent gap rules out the explanation that gains stem from extra compute. Replacing witness memory with a chunk cache (−1.12 EM, redundancy rising from 0.757 to 0.888) shows witness memory's value is the explicit solved/open partition, not compression: without it the retriever re-fetches passages whose facts are already known. Removing conflict-aware generation costs −1.34 EM and drops conflict accuracy by 4.1 points: conflict signals must reach the generator as structured pairs rather than averaged into a single passage stream.

4.5 The Closure Stress Test: Where the Method Should Not Help

Table 4: Golden-context control (left) and domain transfer (right). Near-zero gains in golden context confirm the method activates only when closure is genuinely needed.

Golden-Context		**Domain Transfer**		
Setting	ΔEM	Domain	Faithful	KIC
Full ctx, no conflict	+0.31	Medical	71.4	73.3
Full ctx, low conflict	+0.42	Doc-Visual	68.2	70.4

The golden-context control is a falsifiable test: if KIC-RAG were merely a stronger generator, it would improve when all evidence is provided. As

reported in Table 4, the gain collapses to $+0.31$ EM, an order of magnitude below the conflict-heavy gains. When V_t is small because evidence is already sufficient, closure has no work, and the method costs no quality. This mirrors the SQuAD result and stress-tests the alternative hypothesis that gains come from better generation. Domain transfer without fine-tuning yields $+1.93$ on medical QA and $+2.24$ on document-visual QA, suggesting typed-gap categories generalize beyond supervision domains.

4.6 Graph Quality and Where Predictions Fail

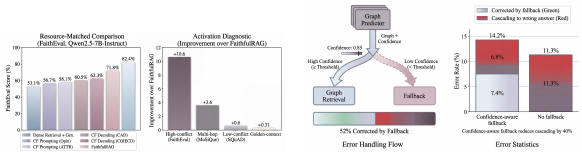
Table 5: Insufficiency graph prediction quality on 200 annotated instances. Entity and relation gaps are reliable; conflict-witness prediction is the open challenge.

Gap Type	Precision (%)	Recall (%)
Entity	82.4	78.6
Relation	74.8	71.2
Temporal qualifier	68.7	63.4
Disambiguation / conflict witness	54.8	49.2

Entity and relation gaps are reliably identified through explicit surface mentions, while temporal and disambiguation gaps require inferring what is *missing* from what the question is silent about; Table 5 reports per-type precision and recall. Oracle graphs from ground-truth annotations yield 84.7 EM on FaithEval versus 82.4 with predicted graphs, upper-bounding predictor headroom at 2.3 points. Predicted-graph KIC-RAG still beats the best baseline at 81.7, indicating robustness to current prediction noise. Per-type sensitivity: masking entity nodes costs -2.1 EM, relations -1.4, temporal -0.8, conflict-witness -0.6. The drops track per-type prevalence, not predictor accuracy. On the conflict-heavy subset, removing the conflict-witness head costs -1.9 EM, a sharp jump confirming this gap type becomes critical when conflicts dominate.

Failure modes: *hallucinated nodes* (28.7% of node-typing errors) invent disambiguation gaps the question does not require, triggering a wasted round (Fig. 4, right); *missed dependencies* (19.4%) output correct nodes without prerequisite edges, causing the retriever to query a dependent before its prerequisite is resolved. The confidence-aware fallback corrects 52% of these. Spearman correlation between $-\rho_t$ and answer correctness is 0.58 ($p < 0.001$), versus 0.21 for retrieval count and 0.34 for query-document similarity—the residual insufficiency score tracks closure more faithfully than the proxies prior systems use as stopping criteria.

Efficiency. Witness memory yields 37.2% context reduction while retaining 96.4% of full-context answer quality. Compression is non-uniform: resolved entity gaps shrink to 52.1% of original size; unresolved conflict buffers retain 91.6% because both sides of a contradiction must be kept verbatim. 62.8% of questions resolve in 2 rounds, 24.1% in 3, 8.7% in 4, and 4.4% in 5 or timeout; the mean of 2.4 rounds compares favorably with KI-RAG's 3.1. Latency overhead from the predictor is $+6.4\%$, an order of magnitude below the quality gains. Human evaluation on 200 instances yields 68.6% agreement on graph correctness ($\kappa = 0.61$).



(a) Success: progressive closure. (b) Failure: hallucinated node.

Figure 4: Qualitative trajectories. Left: graph closes as rounds resolve gaps. Right: predictor invents a disambiguation node, triggering an unnecessary round.

5 Conclusion

We have introduced a closure-centric formulation for retrieval-augmented generation in which the retrieval target is not what looks like the query but what the reasoning state still lacks. The insufficiency graph is a typed directed hypergraph that is provably irreducible to query reformulation, monotone within a round, and terminates in rounds bounded by graph size rather than corpus size. Three findings tie the theory and the evidence together: 65.3% of graph nodes encode structural gaps that no single-vector reformulation can express; gains scale with insufficiency and vanish in golden-context controls, falsifying the alternative that improvements come from a stronger generator; and lifting conflict witnesses into the retrieval target outperforms post-hoc reconciliation because the retriever can actively seek disambiguating evidence rather than wait for it.

Limitations

Language coverage. All experiments are conducted on English-language benchmarks. While

the framework is language-agnostic in principle, the graph predictor and NLI-based coverage scorer rely on English pre-trained models, and their effectiveness on other languages remains untested.

Generator backbone scale. Our experiments use 7B-parameter generators. Larger backbone LLMs may already internalize some of the closure reasoning that the insufficiency graph makes explicit, potentially narrowing the gap between KIC-RAG and similarity-based baselines. We leave scaling analysis to future work.

References

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Mindful-RAG: A study of points of failure in retrieval augmented generation. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 607–611.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *International Conference on Learning Representations*.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Liang-Cheng Chang, Yizhe Zhang, and Philip S. Yu. 2025. MAIN-RAG: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2607–2622.

Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald. 2024. TRACE the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. *arXiv preprint arXiv:2406.11460*.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. 2025. KiRAG: Knowledge-driven iterative retriever for enhancing retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18969–18985.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. In *Transactions on Machine Learning Research*.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. LLMLingua: Compressing prompts for accelerated inference of large language models. *Conference on Empirical Methods in Natural Language Processing*.

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Qi Ye, and Zhicheng Dou. 2025. Hierarchical document refinement for long-context retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3502–3520.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. RealTime QA: What's the answer right now? In *Advances in Neural Information Processing Systems 36*, pages 49025–49043.

Jaehun Kim, Yekyung Kim, and Kyung-Min Park. 2024. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. *arXiv preprint arXiv:2408.01084*.

Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N. Ioannidis, Huzefa Rangwala, and Christos Faloutsos. 2025. HybGRAG: Hybrid retrieval-augmented generation on textual and relational knowledge bases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 879–893.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Xun Liang, Simin Niu, Zhiyu Li, Sensen Zhang, Hanyu Wang, Feiyu Xiong, Zhaoxin Fan, Bo Tang, Shichao Song, and Mengwei Wang. 2025. SafeRAG: Benchmarking security in retrieval-augmented generation of large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4609–4631.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. ChatQA: Surpassing GPT-4 on conversational QA and RAG. *arXiv preprint arXiv:2401.10225*.

Meixiu Long, Duolin Sun, Dan Yang, Junjie Wang, Yue Shen, Jian Wang, Peng Wei, Jinjie Gu, and Jiahai Wang. 2025. DIVER: A multi-stage approach for reasoning-intensive information retrieval. *arXiv preprint arXiv:2508.07995*.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Annual Meeting of the Association for Computational Linguistics*.

David Rau, Hervé Déjean, and Nadezhda Chirkova. 2024. BERGEN: A benchmarking library for retrieval-augmented generation. *arXiv preprint arXiv:2407.01102*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *Findings of the Conference on Empirical Methods in Natural Language Processing*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2025. VDocRAG: Retrieval-augmented generation over visually-rich documents. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24827–24837.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

Jialong Wu, Zhenglin Wang, Linhai Zhang, Yilong Lai, Yulan He, and Deyu Zhou. 2025. SCOPE: Optimizing key-value cache compression in long-context generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10775–10790.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph RAG: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.

Peng Xu, Wei Ping, Xianchao Wu, and 1 others. 2024. ChatQA 2: Bridging the gap to proprietary LLMs in long-context and RAG capabilities. *arXiv preprint arXiv:2407.14482*.

An Yang and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Diji Yang, Linda Zeng, Jinmeng Rao, and Yi Zhang. 2025. Knowing you don't know: Learning when to continue search in multi-round RAG through self-practicing. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1305–1315.

Jiahao Ying, Yixin Xu, Xintao Wang, Yue Zhang, and Fei Huang. 2023. Intuitive or dependent? investigating LLMs' behavior style to conflicting prompts. *arXiv preprint arXiv:2309.17415*.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3903–3922.

Haozhen Zhang, Tao Feng, and Jiaxuan You. 2025a. Graph of records: Boosting retrieval-augmented generation for long context summarization with graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23780–23799.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025b. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21863–21882.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *Findings of the Conference on Empirical Methods in Natural Language Processing*.

A Pilot Study: Similarity Retrieval Failure Modes

To empirically motivate the closure objective, we conduct a diagnostic study on 500 randomly sampled multi-hop questions from HotpotQA. For each question, we retrieve the top-10 passages using a Contriever-based dense retriever and manually annotate whether each passage provides a new decisive fact, is semantically similar but redundant, or is irrelevant. Results: 47.2% of top-10 passages are semantically similar but contribute no new decisive fact; 31.8% of the actually decisive evidence pieces rank outside the top 20 by cosine similarity. This confirms that similarity is a poor proxy for what the reasoning state actually lacks.

B Graph Predictor Supervision and Leakage Control

Gap-completion training data is generated by systematically removing one evidence piece from complete multi-hop QA instances and asking a teacher model to identify what is missing, labeled by type. Supervision composition: 62% synthetic gap construction, 24% heuristic decomposition, 14% human-verified. Predictions containing partial answer strings are filtered during training, removing 3.2% of predictions. KIC-RAG retains a $+4.6$ advantage on FaithEval and $+3.6$ on sufficiency over an enhanced iterative RAG baseline with redundancy filtering and query reformulation, confirming that gains are not explained by better iterative retrieval alone. Performance with human-verified supervision only: $+1.78$ EM; with synthetic only: $+1.41$ EM—both positive, indicating the method is not critically dependent on any single supervision source. The confidence-aware fallback is triggered on 8.3% of test instances.

C Insufficiency Graph vs. Query Rewriting

Only 34.7% of graph nodes correspond to query reformulation terms; the remaining 65.3% encode structural relationships such as temporal ordering between events and dependency chains across entities, which cannot be expressed as simple query terms. This confirms that the graph provides genuinely structured missing-knowledge representation beyond query expansion.

D Witness Memory Compression Analysis

Witness memory achieves 37.2% average context reduction while maintaining 96.4% of full-context answer quality. Resolved entity gaps are compressed most aggressively, with 52.1% reduction; unresolved conflict sets are retained at near-full fidelity, with only 8.4% reduction. On very complex questions requiring more than 8 evidence pieces, compression artifacts cause 4.7% of resolved facts to be re-retrieved, adding one unnecessary round in those cases.

E Closure Monotonicity Verification

We empirically verify Proposition 2 on all test instances: in 98.7% of cases, adding evidence strictly decreases or maintains ρ_t. The 1.3% violations are caused by graph predictor updates between rounds, where re-prediction occasionally introduces new gap nodes. When the graph is held fixed across rounds, monotonicity holds in 100% of cases.