REALHOP: RETHINKING MULTI-HOP REASONING EVALUATION VIA BEHAVIORAL AUDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex questions often require multi-hop reasoning that connects facts distributed across sources or distant regions of a long context through intermediate steps. Benchmarks commonly evaluate this ability with questions built around predefined reasoning chains, treating a correct answer as evidence that the intended composition was used. Yet answer correctness alone leaves open whether success depends on the evidence associated with each intended step: models may instead rely on memorized associations, shorter paths, or partial evidence. We examine this dependence using the Behavioral Necessity Rate (BNR), which measures how often targeted evidence removal prevents answer recovery on initially correct instances. Across five existing benchmarks, panel-mean BNR is only 16.6–48.9%, exposing a substantial gap between annotated structure and observed dependence. Guided by this diagnosis, we introduce REALHOP, a diagnose–construct–verify framework that rebinds entities, factorizes selected relations, adds complete competing paths, and places evidence at traceable locations. Structural and semantic checks precede freezing; behavioral interventions follow. On 790 paired MuSiQue questions, REALHOP raises panel-mean BNR from 27.4% to 94.4% while retaining high Full accuracy. It also yields high BNR on REALHOP-FRAMES and REALHOP-LONGBENCH. On 216 long-context questions, the matched multiple-choice spread across 16 models grows from 13.9 to 59.2 points and persists under repeated open-ended evaluation. Together, these results show that a conceptually coherent chain and a correct final answer do not by themselves establish multi-hop reasoning. Verifying that success depends on every intended hop is therefore as fundamental to multi-hop evaluation as measuring answer accuracy itself.

1 INTRODUCTION

Multi-hop reasoning is essential for meeting complex information needs: it requires models to connect facts distributed across sources or distant regions of a long context and use intermediate reasoning steps to answer questions that no single piece of information can resolve (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022; Bai et al., 2025). To evaluate this capability, researchers construct questions with predefined reasoning chains and supporting evidence, such that information obtained at one step supports subsequent inference and ultimately leads to the answer (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). Answer accuracy is therefore commonly interpreted as evidence that the model carried out the intended composition. This interpretation, however, raises two questions: **Does model success require the evidence associated with each intended step in the predefined chain? And to what extent can success on multi-hop QA benchmarks be taken as evidence of multi-hop reasoning ability?**

These questions matter because models can bypass the predefined chain in several ways: they may recall the answer, follow an unannotated shorter path, or exploit a surface cue (Min et al., 2019; Jiang & Bansal, 2019) (Figure 1). Existing construction techniques reduce such shortcuts through connected composition, false-chain perturbations, and plausible distractors (Trivedi et al., 2022; Serbanescu et al., 2025; Bhuiya et al., 2024), but neither a coherent path nor lower accuracy establishes dependence: noise can make a benchmark harder without making it more diagnostic. A more direct way to test evidence necessity is to intervene on the claimed supports and observe whether model success changes.

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.

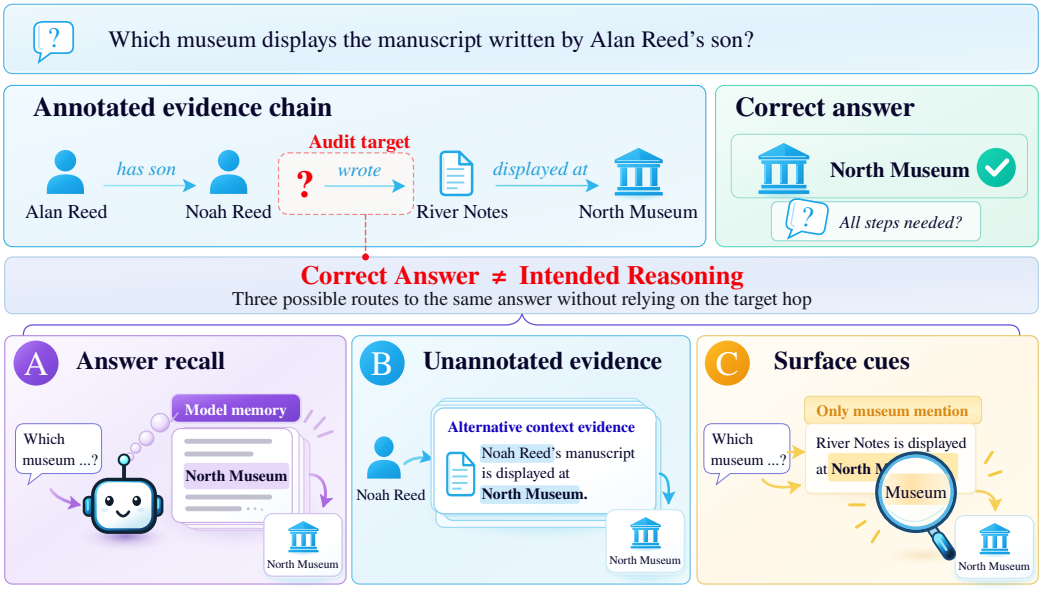


Figure 1: Three bypasses of an annotated gold chain: answer recall, unannotated evidence, and surface cues. Each can preserve the answer after evidence for a gold hop is removed.

We operationalize this idea through behavioral auditing based on targeted evidence deletion. Among questions answered correctly under Full context, the *Behavioral Necessity Rate* (BNR) measures how often removing claimed evidence prevents answer recovery; a matched-background Control separates evidence-specific failure from generic deletion sensitivity. Across MuSiQue, HotpotQA, and 2WikiMultiHopQA, panel-mean BNR is only 16.6–27.4%. Even targeted adversarial constructions designed to make shortcut solving harder yield panel-mean BNRs of only 43.5–48.9% (Serbanescu et al., 2025; Bhuiya et al., 2024). Together, these results show that neither explicit hop annotations nor targeted distractors ensure that successful prediction depends on the intended evidence.

To address this gap, we propose REALHOP, a diagnose–construct–verify framework for building and testing multi-hop questions designed to require the evidence for every reasoning step. It first checks whether models can bypass annotated evidence in existing benchmarks. It then replaces entities consistently, splits selected relations into premises that work together, and adds complete paths to alternative answers that fail a required relation. Each inserted fact is recorded at a known text location. Structural and semantic checks determine which items are retained before the dataset is frozen; subsequent evidence-deletion tests measure whether models can answer with and without the intended support. This separates benchmark construction from behavioral evaluation.

We evaluate this framework in three multi-hop settings. On 790 paired MuSiQue questions, REALHOP-MUSIQUE raises panel-mean BNR from 27.4% to 94.4%, while Full accuracy remains 68.5–83.5%. On source-linked FRAMES (Krishna et al., 2025) questions without native hop annotations, REALHOP-FRAMES reaches 88.9–89.9% BNR. We then scale the construction to 216 long context questions from LongBench v2 (Bai et al., 2025). Under a matched multiple-choice protocol, REALHOP-LONGBENCH expands the 16-model accuracy spread from 13.9 to 59.2 points. This wider separation persists in repeated open-ended evaluation, while targeted deletion yields 96.5–98.1% BNR across three audited models.

Our contributions are threefold:

- **We audit existing multi-hop benchmarks and reveal the gap between annotated reasoning structure and observed evidence dependence.** Annotated chains specify the intended reasoning, yet models can still answer correctly after evidence for a step is removed. We introduce BNR to quantify this gap across three standard benchmarks and two adversarial constructions.

- **We introduce REALHOP, a diagnose–construct–verify framework for rebuilding benchmarks to make each reasoning step necessary for answering.** It combines entity rebinding, evidence rewriting, and complete competing paths, with structural and semantic checks before dataset freezing and behavioral evaluation afterwards.
- **We construct and evaluate three benchmark variants covering standard and long-context multi-hop question answering.** Built from MuSiQue, FRAMES, and LongBench v2, these variants are evaluated through targeted evidence deletion, component ablations, and repeated long-context trials across 16 models.

2 RELATED WORK

Standard multi-hop QA and shortcut-resistant evaluation. Multi-hop QA datasets annotate supporting chains so that answer correctness can be read as evidence of connected reasoning (Welbl et al., 2018; Yang et al., 2018; Talmor & Berant, 2018; Ho et al., 2020; Trivedi et al., 2022). Analyses of HotpotQA showed that compositional questions still admit single-hop solutions, sentence-level cues, and word-matching routes, including adversarial documents that preserve the original answer (Min et al., 2019; Chen & Durrett, 2019; Jiang & Bansal, 2019). Models can also know the individual hops yet fail to compose them (Press et al., 2023).

Construction work then targeted those artifacts: MuSiQue composes connected single-hop questions and filters disconnected reasoning (Trivedi et al., 2022); StrategyQA and MoreHopQA add implicit or generative hops (Geva et al., 2021; Schnitzler et al., 2024); seemingly plausible distractors insert complete but incorrect multi-hop chains (Bhuiya et al., 2024); and FalseCoTQA and CRiT-QA inject knowledge-grounded or hop-anchored counterfactual chains (Serbanescu et al., 2025; Yun et al., 2026). Adversarial distractor methods change the evidence a model sees and typically report the resulting accuracy drop to evaluate robustness. Such methods can make shortcut solving harder, but an accuracy drop under stronger distractors does not establish that successful predictions depend on every intended hop.

Long-context multi-hop reasoning. Long contexts extend the multi-hop challenge: models must locate facts dispersed across lengthy or multi-document inputs and then combine them along a reasoning chain. LongBench and LongBench v2 include multi-document and reasoning-intensive tasks (Bai et al., 2024; 2025), while Loong places multi-document questions in extended carriers (Wang et al., 2024). FRAMES provides source-linked questions that require combining information across sources, but does not annotate their hop chains (Krishna et al., 2025). Controlled evaluations make the same challenge explicit: RULER includes multi-hop tracing, and BABILong embeds reasoning facts in long background (Hsieh et al., 2024; Kuratov et al., 2024). Synthetic long-context training similarly hides multi-hop questions behind key chains or trajectory-derived distractors (Wang et al., 2026; Lin et al., 2026). These settings test whether models can find and combine information at scale, but neither context length nor the presence of a chain establishes that successful answers depend on every intended hop. Long-context evaluation therefore scales the multi-hop problem without resolving evidence necessity.

Evidence necessity through intervention. Sufficiency asks whether selected evidence is enough; necessity asks whether a correct prediction survives its removal. Deletion tests from interpretability make evidence dependence observable. Extractive rationales treat a sparse subset of the input as sufficient for prediction (Lei et al., 2016); later work asks whether those selections are also comprehensive (Yu et al., 2019; DeYoung et al., 2020). Attention weights and chain-of-thought traces remain unreliable proxies for this dependence (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Jacovi & Goldberg, 2020; Turpin et al., 2023; Lanham et al., 2023). Within multi-hop QA, DiRe combines predictions on complementary support subsets to measure disconnected reasoning (Trivedi et al., 2020). REALHOP directly intervenes on individual hops to measure whether a successful prediction depends on each claimed support, while a matched background deletion separates evidence-specific failure from generic deletion sensitivity. Unlike prior audits, this diagnosis feeds back into construction: complete competing paths make intended relations consequential, and structural, semantic, and behavioral checks verify the result. The resulting diagnose–construct–verify framework applies the same criterion of evidence necessity across standard and long-context multi-hop settings.

Table 1: Panel-mean Full, Drop-one, Control, and BNR on existing multi-hop benchmarks (percent). Means give equal weight to DeepSeek V4 Flash, Gemini 3.6 Flash, and Qwen 3.7 Max. Drop-one and Control are accuracies over all audited questions, whereas BNR conditions on Full-correct ones.

Benchmark	Full	Drop-one	Control	BNR
MuSiQue (Trivedi et al., 2022)	87.4	65.4	87.3	27.4
HotpotQA (Yang et al., 2018)	81.2	69.7	81.5	16.6
2WikiMultiHopQA (Ho et al., 2020)	88.6	74.5	89.1	17.0
FalseCoTQA (Serbanescu et al., 2025)	63.8	36.3	59.9	43.5
Plausible Distractors (Bhuiya et al., 2024)	87.8	44.8	86.4	48.9

3 Problem Formulation and Benchmark Audit

A model can answer correctly without relying on a benchmark's annotated evidence structure. We formalize this gap and audit whether models depend on each claimed evidence target.

3.1 The Conceptual–Behavioral Gap

An annotated evidence graph claims that its steps form a coherent and sufficient structure supporting the answer, but does not establish that a model must use every step. Formally, an instance is $x = (q, C, a^\star, G)$. Let H index the intervention targets associated with the claimed steps or support groups in G, and let $W_h \subseteq C$ denote the evidence assigned to target $h \in H$. We intervene on h by evaluating the unchanged question on $C \setminus W_h$. *Conceptual necessity* holds when $C \setminus W_h$ contains no sufficient path to $a^\star$; *behavioral necessity* holds for a model when it answers correctly on C but not on $C \setminus W_h$. The former does not guarantee the latter: a model may bypass a coherent annotated chain through shortcuts, answer cues, unannotated evidence, or parametric knowledge.

3.2 Audit Protocol and Metrics

The *Behavioral Necessity Rate* (BNR) compares Full with Drop-one, which removes one declared evidence target at a time. For a fixed model, we restrict attention to items with complete Drop evaluations, sample item i uniformly and target h uniformly from H_i, and let F_i indicate Full correctness, D_{ih} Drop correctness, and $\mathcal{I}^+ = \{i : F_i = 1\}$:

$$\begin{aligned} \mathrm{BNR} &= \mathbb{E}[1 - D_{ih} \mid F_i = 1] \\ &= \frac{1}{|\mathcal{I}^+|} \sum_{i \in \mathcal{I}^+} \frac{1}{|H_i|} \sum_{h \in H_i} (1 - D_{ih}). \end{aligned} \tag{1}$$

Thus, BNR is the item-macro proportion of testable targets whose removal prevents recovery of the gold answer, without giving longer chains greater weight. Because BNR is model- and protocol-specific, we report it alongside Full accuracy. Length-matched background Control provides a separate specificity check and is not an input to BNR.

3.3 Audit of Existing Multi-Hop Benchmarks

Multi-hop annotation substantially overstates observed dependence. We audit 790 source MuSiQue, 300 HotpotQA, 300 2WikiMultiHopQA questions, a stratified 300-item FalseCoTQA–MuSiQue sample, and a 300-item Plausible Distractors split of HotpotQA with DeepSeek V4 Flash, Gemini 3.6 Flash, and Qwen 3.7 Max. Sampling and deletion units appear in Appendix A.

Table 1 exposes a consistent gap. Control remains near Full and well above Drop, supporting evidence-specific sensitivity rather than an effect of deletion volume alone. Yet BNR is only 16.6–27.4% on the three standard benchmarks. FalseCoTQA and Plausible Distractors raise it to 43.5% and 48.9%, respectively, but correct answers still survive most gold-hop deletions. These descriptive, unpaired results remain far below the 93.9–95.0% achieved by RealHop-MuSiQue (Section 6.1).

These findings motivate constructions that make every intended binding consequential; Section 4 describes how RealHop pursues this goal in both standard and long-context multi-hop QA.

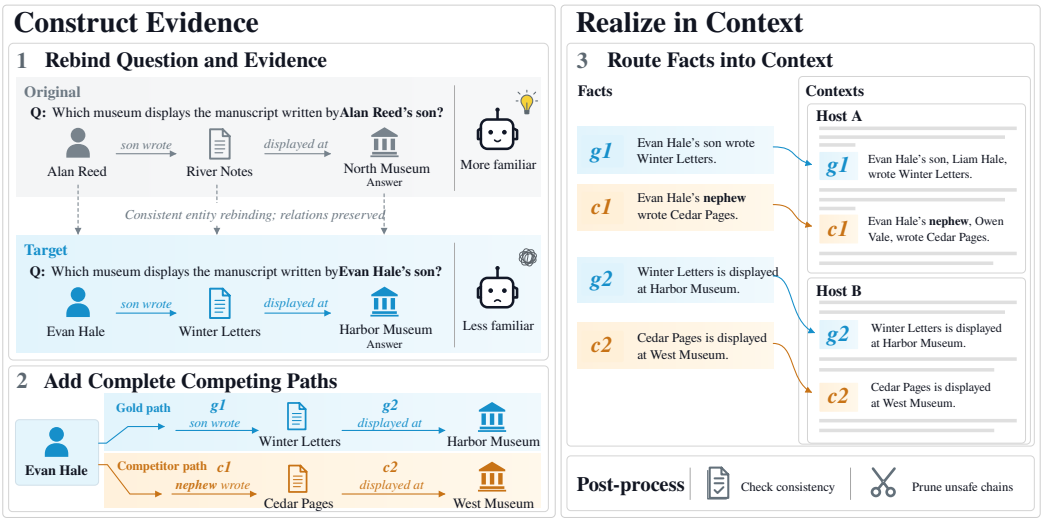


Figure 2: **Overview of the two-stage construct phase of REALHOP.** The framework constructs correct and competing evidence paths, then realizes and validates them in context. Gray marks the original source graph, blue the correct path, and orange the competing path; matching labels link planned facts to their contextual realizations.

4 METHOD

This section details the construct phase of the REALHOP framework, which proceeds in two stages. Given a multi-hop question q, context C, answer $a^\star$, and dependency graph G, it first **constructs evidence**, then **realizes it in context** (Figure 2). Models generate facts and text; separate checks determine sample acceptance.

4.1 CONSTRUCT EVIDENCE

Rebind question and evidence. We replace entities in the question, planned facts, and answer with *virtual counterparts*, preserving relations, entity roles, answer type, and reasoning structure. In the figure, Alan Reed, River Notes, and North Museum become Evan Hale, Winter Letters, and Harbor Museum. This aims to prevent answer recall from prior knowledge and make models reason from the provided context. The question does not reveal intermediate answers, and original passages remain as background rather than undergoing global name replacement. For some relations, we also use several pieces of information (*evidence units*) intended to be combined (Appendix I.1).

Add complete competing paths. For each reasoning step, we add an alternative with the same entity types but a similar relation that does not meet the question's requirement. We then add the subsequent facts needed to reach a different answer. In the figure, Evan's son wrote Winter Letters, displayed at Harbor Museum; his nephew wrote Cedar Pages, displayed at West Museum. Because both manuscripts have museum evidence, a museum mention alone no longer selects one path. The aim is to make the writer's relationship to Evan essential for choosing between the answers.

We use two variants for constructing competitor paths, *Flexible* and *Strict*. Flexible allows the relation mismatch at the target step or, in a separate candidate path, at a later step if this introduces no ambiguity or additional valid answer. Strict allows only a target-step mismatch and keeps the entities in unaffected dependencies unchanged. After problematic paths are removed, at least one complete Strict path remains for each step (Appendix I.2); ablations compare both variants in Section 7.

4.2 REALIZE IN CONTEXT

Route facts into context. A *host* is an existing passage that receives generated evidence. Rewritten gold facts are inserted into the passages containing the original supporting evidence; competing

facts are assigned to model-selected compatible passages. In the figure, g_1, c_1 go to Host A and g_2, c_2 to Host B.

Realization. Models turn the assigned facts and premises into sentences, such as "Evan Hale's son, Liam Hale, wrote Winter Letters." Names and relations must match the planned facts. The original text and its order are preserved, without adding unplanned relations or explicitly ruling out competing answers. We record the exact text expressing each fact and evidence unit.

Check the generated content. We first review the proposed facts and relations, then check whether the text expresses them correctly. Automated checks identify missing facts, changes to the original text, and incorrect evidence locations. Semantic review checks names, relations, question meaning, unclear references, and unintended correct answers. Problematic competing paths are removed, and the remaining evidence and paths are checked against the construction requirements.

Once the dataset is fixed, we test whether models answer correctly with the full context (Full) and after removing specified evidence (Drop). Deleting similar-length background text (Control) helps distinguish evidence loss from general text removal (Section 3.2). Two annotators also assess sampled items for quality (Appendix J).

4.3 Adapt Across Settings

RealHop-MuSiQue. MuSiQue provides annotated dependency graphs and supporting passages, which define G and the initial evidence hosts. We can therefore apply the construction process directly to these annotations.

RealHop-FRAMES. FRAMES provides source links but no hop annotations. We reconstruct the linked articles and ask a model to infer a linear reasoning chain with quoted evidence for each step. We then apply the same construction process.

RealHop-LongBench. We start from a fictional graph; the source question guides only task style. The same method constructs gold and competing paths, with evidence placed at separate sentence boundaries across long-document chunks to test retrieval and reasoning over longer contexts. Across the three settings, the adapters vary how reasoning graphs are obtained and evidence is placed while preserving the same objective: making each reasoning step behaviorally necessary.

5 Experimental Setup

Our evaluation asks whether RealHop increases behavioral necessity, which components create the challenge, and whether the long-context construction separates models consistently across trials.

RealHop-MuSiQue and RealHop-FRAMES evaluations. RealHop-MuSiQue contains 790 source–constructed pairs spanning 2–4 hops; the RealHop-FRAMES audit uses a selected cohort of 81 matched source/constructed question pairs with inferred hop spans. Both evaluate DeepSeek V4 Flash, Gemini 3.6 Flash, and Qwen 3.7 Max under Full and targeted Drop and report Full accuracy and BNR. Within each setting, the source and constructed sides use the same solver and scoring protocol. MuSiQue confidence intervals resample paired questions, retaining their interventions and model responses together. FRAMES reports Avg@3 over its recorded trials.

RealHop-LongBench evaluation. RealHop-LongBench, derived from LongBench v2 (Bai et al., 2025), contains 216 open-ended short-answer questions with constructed evidence chains. Each answer is a concise terminal entity or identifier; prompts range from about 11k to 575k tokens. We evaluate 16 models over three trials at each model's highest supported reasoning effort. Exact normalized matches are accepted directly; remaining answers use a frozen short-answer judge (Appendix K). BNR uses the open-ended Full/Drop runs for three models reported in Appendix G, not the three-trial Avg@3 scores. As a format-matched reference, we evaluate both the source questions and RealHop-LongBench once under the official multiple-choice template with the same 16 models; the constructed items use a frozen A–D order whose non-gold choices are planned competitor endpoints (Table 2).

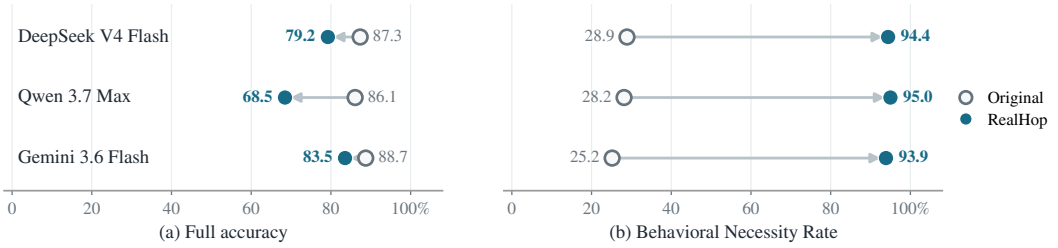


Figure 3: Full accuracy and BNR on source MuSiQue and REALHOP-MUSIQUE. Lower Full accuracy indicates greater difficulty; higher BNR indicates stronger behavioral dependence on the intended evidence.

6 RESULTS

Section 3 identified a substantial gap between annotated multi-hop structure and behavioral necessity; the results below show that REALHOP sharply narrows this gap. We first test evidence necessity through paired source–constructed comparisons in the MuSiQue and FRAMES settings, then scale the same construction objective to long-context multi-hop evaluation, where REALHOP-LONGBENCH is assessed through targeted intervention and repeated model comparisons.

6.1 BEHAVIORAL EFFECTS ON REALHOP-MUSIQUE

Figure 3 summarizes the paired source MuSiQue/REALHOP-MUSIQUE comparison on all 790 questions. Original Full accuracy differs by only 2.6 points across the three models, offering little discrimination. On REALHOP-MUSIQUE, however, the between-model spread widens to 15.0 points, so capability differences that stay compressed on the source items become easier to see.

Despite this increase in difficulty, Full accuracy remains 68.5–83.5%, providing a substantial Full-correct population for BNR evaluation. Within this harder but still answerable setting, BNR rises from 25.2–28.9% on Original to 93.9–95.0% on REALHOP-MUSIQUE. The paired increases are 65.5 points for DeepSeek (95% CI 63.2–67.9), 68.7 for Gemini (66.4–71.0), and 66.8 for Qwen (64.4–69.2); on the common-Full subset, gains remain 66.6–69.7 points (Appendix C). The construction therefore does more than lower accuracy: it sharply increases behavioral dependence on the intended evidence. Its observed BNR also substantially exceeds the levels in our FalseCoTQA and Plausible Distractors audits (Section 3.3), although those use different benchmark samples. Among incorrect Full answers, 64.3–69.5% exactly match a planned competitor endpoint (Appendix D).

For six earlier models, paired accuracy drops by 25.3–47.9 points on model-specific subsets of 305–353 questions, including 44.8 points for GPT-4o (Appendix E). These declines establish increased difficulty, but do not by themselves establish greater reliance on source shortcuts: the subsets differ from the current-model panel, and these runs do not measure BNR.

6.2 REALHOP-FRAMES HOP-LEVEL AUDIT

On the selected 81 matched FRAMES–REALHOP-FRAMES pairs, each trial is binarized and scored separately before averaging; each side's BNR conditions on its own Full-correct subset. Avg@3 source Full is 97.5–98.8%, with BNR only 6.6–8.1%. After construction, BNR rises to 88.9–89.9% while Full remains 68.4–86.0%, so the increase is not obtained by counting unsolved Full questions as evidence-dependent successes. Per-trial scores appear in Appendix I.4.

On the common-Full subset, BNR gains remain 81.8–83.5 points; length-matched background scores exceed Drop (Appendix I.5).

6.3 REALHOP-LONGBENCH ACROSS REPEATED TRIALS

Table 2 compares the two leaderboards. The source leaderboard is comparatively compressed: accuracy under its official protocol spans 58.3–72.2%, and older models sometimes outrank newer members of the same family. Gemini 3.5 Flash leads Gemini 3.6 Flash, and Grok 4.5 leads Grok 4.6.

Table 2: Source ranks use accuracy on the corresponding source questions under the official LongBench v2 multiple-choice template. Trials 1–3 and Avg@3 are open-ended scores on REALHOP-LONGBENCH; the rightmost MCQ column evaluates the same constructed items under that template, run once. Eff. denotes the provider-specific reasoning-effort setting. Arrows give the rank change from the source leaderboard to Avg@3. Spread is the range of displayed scores across models.

		LongBench v2		REALHOP-LONGBENCH						
Model	Eff.	Acc.	Rank	Trial 1	Trial 2	Trial 3	Avg@3	Rank	Change	MCQ
Claude Opus 5	max	70.4	**2**	80.6	85.2	81.9	82.6	**1**	↑ 1	94.4
GPT-5.6 Sol	xhigh	68.1	**3**	63.9	63.9	65.7	64.5	**2**	↑ 1	75.5
Kimi K3	max	65.7	6	52.8	59.7	50.9	54.5	**3**	↑ 3	64.8
Claude Opus 4.8	max	64.8	8	53.7	49.1	55.1	52.6	4	↑ 4	79.6
Grok 4.6	xhigh	65.3	7	43.1	40.3	44.0	42.4	5	↑ 2	60.2
Hy4 Preview	high	64.4	10	39.8	43.1	41.7	41.5	6	↑ 4	52.8
Gemini 3.6 Flash	high	68.1	**3**	41.2	39.8	42.6	41.2	7	↓ 4	50.0
DeepSeek V4 Pro	max	63.9	11	34.3	38.0	40.3	37.5	8	↑ 3	56.9
Qwen 3.8 Max	xhigh	63.9	11	40.3	34.7	34.3	36.4	9	↑ 2	60.2
GLM 5.3	max	64.8	8	31.0	30.1	31.5	30.9	10	↓ 2	53.2
DeepSeek V4 Flash	max	63.0	13	32.4	31.0	27.3	30.2	11	↑ 2	55.6
Gemini 3.5 Flash	high	72.2	**1**	29.2	29.6	29.6	29.5	12	↓ 11	38.9
GLM 5.2	max	61.6	15	25.0	27.8	28.7	27.2	13	↑ 2	50.9
Grok 4.5	xhigh	66.7	5	29.6	24.1	27.8	27.2	13	↓ 8	48.6
Qwen 3.7 Max	xhigh	62.5	14	26.9	24.1	28.2	26.4	15	↓ 1	35.2
Kimi K2.6	thinking	58.3	16	27.3	27.8	23.1	26.1	16	–	49.5
Panel spread		13.9		55.6	61.1	58.8	56.5			59.2

Avg@3 on REALHOP-LONGBENCH instead spans 26.1–82.6%. Across the three runs, 15 of 16 models vary by no more than about six percentage points. The constructed task therefore separates models more widely on long-context multi-hop questions, including within-family comparisons that remain close on the source leaderboard. A same-format check rules out answer format as the sole cause of the wider spread: under the identical MCQ template, REALHOP-LONGBENCH spans 59.2 points versus 13.9 on the source questions. Its MCQ and open-ended Avg@3 rankings correlate strongly (Spearman $\rho = 0.86$, $p < 10^{-4}$).

An intervention check confirms that the intended evidence is behaviorally consequential: BNR reaches 97.6% for DeepSeek V4 Flash, 98.1% for Gemini 3.6 Flash, and 96.5% for Qwen 3.7 Max.

7 ANALYSIS

Section 6 establishes the intended behavioral change. We now isolate which parts of the construction method create this challenge.

We isolate the two competitor families on a stratified 80-question sample spanning 2–4 hops. Manual review excludes seven items with ambiguous, incomplete, or invalid questions or answer keys (Appendix F). The remaining 73 items form the cohort for the results below, reducing confounding from identified source-scoring artifacts. All constructed conditions share the question and gold evidence: *Gold-only* contains no competitors, *+Flexible* adds less-constrained relation-level divergences, *+Strict* adds single-break paths with dependency inheritance, and *Full* includes both. The two middle conditions are parallel rather than cumulative: each adds one competitor family to the Gold-only base, while Full combines them. This separates family-specific effects from their joint effect. *Original* denotes the unmodified source questions used as a baseline. Six earlier models characterize accuracy changes. Because their low Full accuracy leaves few Full-correct questions, we estimate BNR with the three current models. Each condition's BNR is conditioned on its own Full-correct subset and is interpreted alongside that condition's Full accuracy.

Difficulty. Table 3 locates the difficulty in structured competition rather than in rewriting the gold path. Mean accuracy is only modestly lower on Gold-only than on Original, then falls as competitor

Table 3: Component ablation on a 73-item MuSiQue cohort (percent). *Original* denotes source questions. Accuracy uses model-specific complete-case subsets; BNR averages hops within each question and is reported only for current models. Mean weights models equally.

			MuSiQue	REALHOP-MUSIQUE			
Metric	Panel	Model	Original	Gold-only	+Flexible	+Strict	Full
Accuracy	Earlier	DeepSeek R1	84.9	72.6	57.5	42.5	37.0
		Doubao 1.5 Pro	74.0	67.1	46.6	30.1	24.7
		GPT-4o	71.2	68.5	31.5	24.7	16.4
		Qwen Max	64.4	58.9	28.8	24.7	21.9
		GLM-4-Plus	58.6	50.0	21.4	20.0	11.4
		Qwen Turbo	39.7	35.6	19.2	9.6	5.5
		Mean	65.5	58.8	34.2	25.3	19.5
	Current	DeepSeek V4 Flash	97.3	94.5	93.2	89.0	83.6
		Gemini 3.6 Flash	93.2	94.5	90.4	91.8	89.0
		Qwen 3.7 Max	94.3	90.0	85.7	82.9	75.7
		Mean	94.9	93.0	89.8	87.9	82.8
BNR	Current	DeepSeek V4 Flash	22.7	56.2	86.5	86.6	91.7
		Gemini 3.6 Flash	18.0	59.1	83.2	86.2	91.0
		Qwen 3.7 Max	29.2	58.3	85.9	87.7	91.4
		Mean	23.3	57.9	85.2	86.8	91.3

families are added; competitor insertion accounts for 85.5% of the earlier-panel mean Original→Full reduction. Strict paths cost more than Flexible competitors overall, and both effects are far larger on the earlier panel than on the current one.

Behavioral necessity. Difficulty alone does not establish the mechanism, so the BNR panel repeats the ablation with targeted all-hop interventions. Rewriting the gold path more than doubles panel-mean BNR. Adding either competitor family then raises BNR by roughly 30 further points; combining both families yields an additional, smaller gain. These comparisons show stronger deletion sensitivity among each condition's Full-correct items, which lower accuracy alone would not establish.

8 LIMITATIONS

BNR measures whether a correct answer survives targeted deletion, not whether the model internally used that hop. The estimate inherits the chosen evidence unit, judge, and trial budget; Drop only indirectly reflects processing, and most reported BNR uses one Full/Drop pair per condition. The constructed items use synthetic entities and planned chains inside source carriers, so they do not cover the full range of natural information needs. On REALHOP-LONGBENCH, the matched MCQ check uses one trial and is not score-equivalent to open-ended Avg@3. FRAMES controls use approximate matching with unverified trial counts; common-Full results are subset-conditional.

9 CONCLUSION

We show that annotated multi-hop structure can remain behaviorally unnecessary even when models answer correctly. REALHOP addresses this gap through a diagnose–construct–verify framework that audits claimed supports, constructs complete competing paths, and validates evidence placement before behavioral evaluation. We quantify behavioral dependence on intended evidence with BNR, which measures how often targeted evidence removal prevents answer recovery among Full-correct items. Panel-mean BNR rises from 27.4% to 94.4% on REALHOP-MUSIQUE, reaches about 89% on REALHOP-FRAMES and above 96% on REALHOP-LONGBENCH, while the long-context evaluation also yields a wider model spread that persists across repeated trials. These results establish evidence necessity as a benchmark design criterion rather than an assumption inferred from annotations or answer accuracy.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, assist with translation, clean and reformat datasets, interpret results, and generate synthetic data in the REALHOP construction pipeline (Section 4). We have not used generative AI tools to support qualitative and thematic data analysis, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, or design or provide feedback on research methodology or experiments. Formulating mathematical claims, providing critical ingredients for proofs, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, create artifacts, and edit the manuscript to improve readability. We have reviewed all AI-assisted work. LLM-generated code was carefully verified and tested by two authors. LLM-polished text was checked at least twice by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The BNR definition and Full/Drop/Control conditions are in Section 3.2; the resampling procedure is in Section 5. The construction procedure and validation criteria are in Section 4 and Appendices I–I.3. Models, trial settings, and scoring rules are in Section 5; external-audit sampling and deletion units are in Appendix A; the frozen solver and judge templates are in Appendix K. Source datasets are public: MuSiQue, HotpotQA, 2WikiMultiHopQA, LongBench v2, and FRAMES. Constructed evaluation sets comprise 790 paired MuSiQue questions, 216 REALHOP-LONGBENCH questions, and 81 FRAMES source–REALHOP-FRAMES pairs, with human spot checks in Appendix J. FRAMES reconstruction and hop-level scoring are in Appendix I.4; LongBench v2 interventions and the random-window probe are in Appendices G and H. FRAMES robustness checks are in Appendix I.5. Anonymized code and executable examples are provided as supplementary material. REALHOP-MUSIQUE will be released publicly upon acceptance.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.183. URL https://aclanthology.org/2025.acl-long.183/.

Neeladri Bhuiya, Viktor Schlegel, and Stefan Winkler. Seemingly plausible distractors in multi-hop reasoning: Are large language models attentive readers? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2514–2528, 2024. URL https://aclanthology.org/2024.emnlp-main.147/.

Jifan Chen and Greg Durrett. Understanding dataset design choices for multi-hop reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4026–4032, 2019. URL https://aclanthology.org/N19-1405/.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. URL https://aclanthology.org/2020.acl-main.408/.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. URL https://aclanthology.org/2021.tacl-1.21/.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020. URL https://aclanthology.org/2020.acl-main.386/.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019. URL https://aclanthology.org/N19-1357/.

Yichen Jiang and Mohit Bansal. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2726–2736, 2019. URL https://aclanthology.org/P19-1262/.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023/.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.243. URL https://aclanthology.org/2025.naacl-long.243/.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.10149.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117, 2016. URL https://aclanthology.org/D16-1011/.

Nianyi Lin, Jiajie Zhang, Lei Hou, and Juanzi Li. LongTraceRL: Learning long-context reasoning from search agent trajectories with rubric rewards. *arXiv preprint arXiv:2605.31584*, 2026. URL https://arxiv.org/abs/2605.31584.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4249–4257, 2019. URL https://aclanthology.org/P19-1416/.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. URL https://aclanthology.org/2023.findings-emnlp.378/.

Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. MoreHopQA: More than multi-hop reasoning. *arXiv preprint arXiv:2406.13397*, 2024. URL https://arxiv.org/abs/2406.13397.

Julien Serbanescu, Mahdiyar Ali Akbar Alavi, Faezeh Ensan, and Fattane Zarrinkalam. FalseCoTQA: Adversarial multi-hop QA via knowledge-grounded false chains of thought. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 160–168. ACM, 2025. doi: 10.1145/3767695.3769494. URL https://doi.org/10.1145/3767695.3769494.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 641–651, 2018. URL https://aclanthology.org/N18-1059/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Is multihop QA in DiRe condition? Measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8846–8863. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.emnlp-main.712/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.04388.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. Leave no document behind: Benchmarking long-context LLMs with extended multi-doc QA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5627–5646. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.322. URL https://aclanthology.org/2024.emnlp-main.322/.

Siyuan Wang, Gaokai Zhang, Li Lyna Zhang, Ning Shang, Fan Yang, Dongyao Chen, and Mao Yang. LoongRL: Reinforcement learning for advanced reasoning over long contexts. In *International Conference on Learning Representations*, 2026. URL https://loongrl.github.io/.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018. URL https://aclanthology.org/Q18-1021/.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. 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. 11–20, 2019. URL https://aclanthology.org/D19-1002/.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. URL https://aclanthology.org/D18-1259/.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. 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. 4094–4103, 2019. URL https://aclanthology.org/D19-1420/.

Jungmin Yun, June Hyoung Kwon, and Youngbin Kim. CRiT-QA: Evaluating multi-hop reasoning with counterfactual chains and distractor traps. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 5246–5255. ELRA Language Resource Association, 2026. doi: 10.63317/2jvxj7kwecuo. URL https://aclanthology.org/2026.lrec-1.410/.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, 2024. doi: 10.18653/v1/2024.acl-long.814. URL https://aclanthology.org/2024.acl-long.814/.

APPENDIX CONTENTS

A External Multi-Hop Audit Details **15**

B Matched-Background Specificity **15**

C MuSiQue Common-Full Analysis **16**

D Competitor-Endpoint Error Analysis **16**

E Historical Model Difficulty **16**

F Component Ablation Details **17**

G REALHOP-LONGBENCH Targeted Interventions **18**

H Random-Window Diagnostics **18**

I Evidence Construction Details **19**

I.1 Plans and Traceability 19

I.2 Flexible and Strict Competitor Contracts 20

I.3 Routing, Realization, and Validation Boundaries 20

I.4 FRAMES Adapter and Hop-Level Audit 21

I.5 FRAMES Matched Controls and Sensitivity 22

J Human Spot Checks **24**

J.1 Human Spot Check of REALHOP-MUSIQUE 24

J.2 Human Spot Check of REALHOP-LONGBENCH 24

K Evaluation Prompts **24**

Table 4: Per-model Full, Drop-one, Control, and BNR on the external multi-hop audit (percent).

Benchmark	Model	Full	Drop-one	Control	BNR
MuSiQue	DeepSeek V4 Flash	87.3	64.1	86.9	28.9
	Gemini 3.6 Flash	88.7	68.2	87.8	25.2
	Qwen 3.7 Max	86.1	64.0	87.3	28.2
HotpotQA	DeepSeek V4 Flash	80.7	68.8	81.3	17.2
	Gemini 3.6 Flash	80.7	69.0	81.5	16.9
	Qwen 3.7 Max	82.3	71.2	81.8	15.6
2WikiMultiHopQA	DeepSeek V4 Flash	87.3	71.3	88.1	19.3
	Gemini 3.6 Flash	89.7	75.6	90.0	17.0
	Qwen 3.7 Max	88.7	76.6	89.3	14.8
FalseCoTQA	DeepSeek V4 Flash	63.0	33.8	55.8	46.4
	Gemini 3.6 Flash	64.3	36.3	62.1	44.8
	Qwen 3.7 Max	64.0	38.9	61.9	39.3
Plausible Distractors	DeepSeek V4 Flash	86.7	46.2	85.2	46.7
	Gemini 3.6 Flash	88.3	44.8	85.8	49.2
	Qwen 3.7 Max	88.3	43.5	88.2	50.8

Table 5: Matched-background specificity on aligned MuSiQue support deletions (percent), computed on Full-correct items with valid Drop and Control results.

	Drop survival		Control survival	
Model	Original	REALHOP-MUSIQUE	Original	REALHOP-MUSIQUE
DeepSeek V4 Flash	73.4	5.7	96.6	93.0
Qwen 3.7 Max	74.4	5.8	97.8	89.7
Gemini 3.6 Flash	76.3	6.2	97.2	92.8

A EXTERNAL MULTI-HOP AUDIT DETAILS

The audit covers the 790 source MuSiQue questions and stratified 300-question samples from HotpotQA and 2WikiMultiHopQA. MuSiQue Drop removes one gold paragraph; HotpotQA and 2Wiki Drop remove the supporting sentences associated with one title. We also audit stratified 300-item FalseCoTQA–MuSiQue and 300-item Plausible Distractors splits, the latter with two related named-entity false-chain paragraphs. In both, Drop removes one official gold support, false chains remain in all conditions, and Control deletes a token-matched ordinary distractor. Table 4 gives the per-model results behind Table 1.

B MATCHED-BACKGROUND SPECIFICITY

We retain matched-background deletion as a separate specificity diagnostic. For each target support, the Control removes background spans that occur verbatim in both Original and REALHOP-MUSIQUE, lie outside every gold support, and exclude newly inserted competitor text. The removed token count matches the associated targeted Drop within five tokens or 5%, whichever is larger. Because a Control may span several locations, it matches deletion volume rather than discourse position.

Control survival remains 89.7–93.0% on REALHOP-MUSIQUE, while targeted Drop survival is 5.7–6.2%. Thus the BNR increase is not accompanied by a comparable loss under generic background deletion.

Table 6: MuSiQue BNR on questions with Full correctness on both sides. BNR is percent; the paired increase and its interval are percentage points.

Model	$\lvert\mathcal{I}^{\cap}\rvert$	BNR Original	BNR REALHOP-MUSIQUE	Increase [95% CI]
DeepSeek V4 Flash	559	28.04	94.66	66.6 [64.0, 69.1]
Gemini 3.6 Flash	599	24.36	94.05	69.7 [67.2, 72.2]
Qwen 3.7 Max	482	27.35	94.90	67.5 [64.8, 70.3]

Table 7: Planned competitor-endpoint matches among Full answers on the 790 REALHOP-MUSIQUE questions (percent).

Model	Wrong hit	Share of errors
DeepSeek V4 Flash	14.4	69.5
Gemini 3.6 Flash	11.1	67.7
Qwen 3.7 Max	20.3	64.3

C MUSIQUE COMMON-FULL ANALYSIS

Each side's BNR in Section 6.1 conditions on its own Full-correct questions, so the Original and REALHOP-MUSIQUE populations differ. This check recomputes both sides on the same questions from the saved responses; no additional model calls are made.

Let $\mathcal{I}^{\cap} = \{i : F_i^O = F_i^R = 1\}$, restricted to questions with complete Drop evaluations on both sides. We recompute both sides' BNR on this set with Equation 1, averaging hops within each question and then questions equally.

The intersection retains 559, 599, and 482 of the 790 questions. BNR increases remain 66.6–69.7 points (Table 6), compared with 65.5–68.7 points on each side's own Full-correct population. Restricting to the common set changes the increase by +1.10, +0.98, and +0.74 points, respectively, with paired 95% intervals of $[0.02, 2.22]$, $[0.00, 1.97]$, and $[-0.64, 2.18]$. The result supports persistence on the jointly answerable subset, not equivalence of populations or replacement of the full-cohort results. Intervals use 10,000 paired whole-question bootstrap resamples with percentile endpoints, base seed 20260925, and fixed model offsets; each resampled question retains its Full and Drop results on both sides, and both Full-correct definitions are recomputed within every resample.

D COMPETITOR-ENDPOINT ERROR ANALYSIS

Table 7 reports how often an incorrect Full response on the 790 REALHOP-MUSIQUE questions exactly matches the terminal answer of a planned competitor. Full correctness uses the same responses and denominator as Figure 3.

Exact competitor endpoints account for 64.3–69.5% of incorrect Full answers across models (wrong-hit rates: 11.1–20.3%). Exact matching excludes paraphrases and other branch-induced errors, making this a conservative signature that models select the alternatives introduced by construction.

E HISTORICAL MODEL DIFFICULTY

Table 8 reports the complete paired results summarized in Section 6.1. This earlier model panel measures the difficulty shift under REALHOP-MUSIQUE, not behavioral necessity.

All six models decline: mean paired accuracy falls from 57.9% on source MuSiQue to 18.2% on REALHOP-MUSIQUE, a 39.7-point reduction. DeepSeek R1 remains strongest in both conditions but has the largest decrease. The effect is therefore neither model-specific nor a rank reversal. These runs measure difficulty only, not BNR.

Table 8: Paired accuracy on source MuSiQue and the corresponding REALHOP-MUSIQUE questions for six earlier models (percent).

Model	Paired n	Original	REALHOP-MUSIQUE	Δ
GPT-4o	353	62.6	17.8	-44.8
Qwen Turbo	352	33.8	8.5	-25.3
Qwen Max	352	55.1	15.6	-39.5
GLM-4-Plus	337	51.9	15.1	-36.8
Doubao 1.5 Pro	340	64.7	20.9	-43.8
DeepSeek R1	305	79.3	31.5	-47.9

Table 9: The seven source questions removed from the component ablation.

Question	Reference key	Defect
Who is the child of Mahmoud Mirza's father?	Ahmad Shah Qajar	Self-referential: Mahmoud Mirza is himself a child of that father, and all nine models answer with his name.
What team does the winner of the 2017 BBC African Footballer of the Year play for?	Egypt national football team	Club and national team are both valid; models answer with the club.
What mountain can you see from Portland, in the state that Raven Creek is located in?	Tualatin Mountains	Several mountains are visible from Portland; the key records one.
Bancroft's county borders what county?	Haliburton County	Multiple bordering counties; answers that name the key alongside another county are scored incorrect.
When did Nissan, the Acura Legend maker and the Scion owner open US assembly plants?	1981	Models answer "early 1980s", which the key cannot accept.
When did the luxury division of the employer of Katsuaki Watanabe change the body style of the rx 350?	Sales began worldwide in April 2012	Key is a sentence rather than a date; models answer "March 2012".
How were people from whom new coins were a proclamation of independence by the Somali Muslim Ajuran Empire expelled from the natural boundary between Thailand and Vat Yotkeo's country?	The dynasty regrouped and defeated the Portuguese	Four-hop composition is not interpretable as a single question.

F COMPONENT ABLATION DETAILS

After the removals described below, the stratified sample contains 39 two-hop, 23 three-hop, and 11 four-hop questions. Flexible competitors significantly reduce paired accuracy for GPT-4o (-37.0 points, $p < 10^{-5}$), Qwen Max (-30.1, $p < 10^{-5}$), GLM-4-Plus (-28.6, $p < 10^{-4}$), Doubao (-20.5, $p = 0.008$), and Qwen Turbo (-16.4, $p = 0.008$); the DeepSeek R1 reduction is borderline (-15.1, $p = 0.061$). Positive interaction terms of 9.6–28.7 points indicate overlapping rather than additive failure mechanisms.

Removed source questions. Seven source items with defective questions or keys are excluded (Table 9). Candidates were items that at most two of nine models answered correctly on source MuSiQue, then checked by hand against the reference key.

BNR follows Equation 1: each Full-correct question with complete Drop evaluations contributes the mean over its gold supports, and questions are averaged. The 73-item sample has 191 gold supports ($39 \times 2 + 23 \times 3 + 11 \times 4$).

Table 10: Targeted-hop intervention results on REALHOP-LONGBENCH (percent).

Model	Full	BNR [95% CI]
DeepSeek V4 Flash	32.4	97.6 [94.3, 100.0]
Gemini 3.6 Flash	41.2	98.1 [95.5, 100.0]
Qwen 3.7 Max	26.9	96.5 [93.1, 99.4]

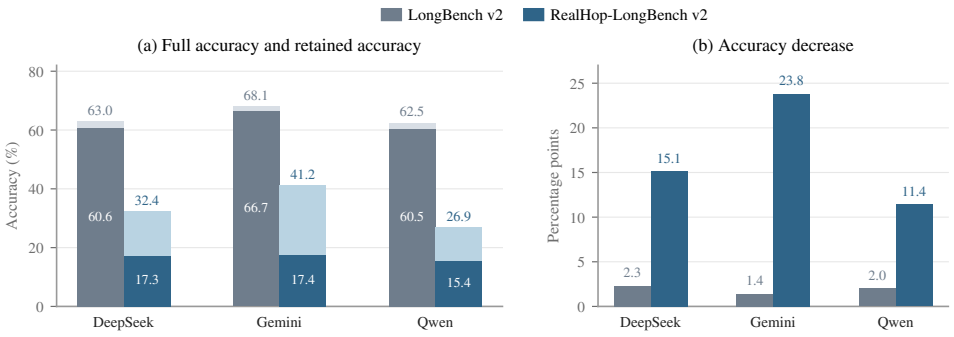


Figure 4: Sensitivity to three frozen circular 20% masks on the matched 216-question panel. **(a)** Full accuracy decomposed into accuracy retained after Mask20 (dark lower segment) and the decrease (light upper segment). **(b)** Mean accuracy decrease. Each Mask20 value averages three frozen masks per question.

G REALHOP-LONGBENCH TARGETED INTERVENTIONS

Full is Trial 1 in Table 2. BNR follows Equation 1: each Full-correct question with complete targeted Drop evaluations contributes the mean over its targeted hops, and questions are averaged.

H RANDOM-WINDOW DIAGNOSTICS

Three frozen contiguous windows each remove 20% of the context. The probe measures spatial sensitivity, not hop necessity. Constructed chains are placed across long carriers, so random deletion can change accuracy.

Protocol. For existing long-context sets, we sample 100 questions each from LongBench v2 (Bai et al., 2025), NarrativeQA (Bai et al., 2024; Kočiský et al., 2018), and InfiniteBench LongBookQA (Zhang et al., 2024). Each Full input is paired with three shared variants that remove a uniformly positioned contiguous 20% window. LongBench v2 uses option matching; the open-ended tasks use a common short-answer judge. Paired effects use complete cases. On the matched 216-question REALHOP-LONGBENCH panel, three frozen circular masks each remove a contiguous 20% token window; accuracy uses the fixed 216-question denominator, and Full is Trial 1 in Table 2. We evaluate Loong (Wang et al., 2024) using the same Full plus three-mask schedule; its open-ended responses use Loong's official 1–100 judge prompt.

On LongBench v2, NarrativeQA, and LongBookQA, panel-mean paired accuracy falls by only 1.9–2.7 points (Table 12). Across the nine model–benchmark pairs, 55.3–83.3% of Full-correct items remain correct under all three masks. Several paired confidence intervals include zero, so the net drop is not a necessity estimate.

On the matched 216 IDs, the same protocol lowers source LongBench v2 accuracy by 1.4–2.3 points, but lowers REALHOP-LONGBENCH accuracy by 11.4–23.8 points (Figure 4, Table 11). This gap is consistent with distributing constructed evidence across the carrier. Which hops are necessary is given by the targeted Drop results in Appendix G.

Table 11: Random 20% mask results on REALHOP-LONGBENCH (percent).

Model	Full	Mask 1	Mask 2	Mask 3	Mask avg.	Accuracy decrease [95% CI]	Full-correct survival
DeepSeek V4 Flash	32.4	18.5	16.7	16.7	17.3	15.1 [8.8, 21.6]	24.3
Gemini 3.6 Flash	41.2	19.4	17.1	15.7	17.4	23.8 [16.8, 30.6]	21.0
Qwen 3.7 Max	26.9	17.6	11.6	17.1	15.4	11.4 [5.2, 17.7]	20.1

Table 12: Full and random-mask accuracy on four long-context audits (percent).

Benchmark	Model	Complete n	Full fixed	Mask fixed	Δ fixed	Δ paired [95% CI]
LongBench v2	DeepSeek	100	73.0	68.7	-4.3	$-4.3\,[-10.0, 1.7]$
	Gemini	100	71.0	69.0	-2.0	$-2.0\,[-6.3, 2.7]$
	Qwen	94	69.0	65.3	-3.7	$-1.8\,[-7.8, 4.3]$
NarrativeQA	DeepSeek	100	57.0	57.0	0.0	$0.0\,[-6.7, 6.7]$
	Gemini	100	54.0	54.3	$+0.3$	$+0.3\,[-5.3, 6.3]$
	Qwen	90	47.0	44.7	-2.3	$-5.9\,[-13.0, 1.1]$
LongBookQA	DeepSeek	98	81.0	76.3	-4.7	$-4.4\,[-10.2, 1.4]$
	Gemini	100	82.0	77.7	-4.3	$-4.3\,[-8.7, 0.0]$
	Qwen	76	58.0	62.7	$+4.7$	$+0.9\,[-4.4, 6.1]$
Loong	DeepSeek	100	81.0	63.3	-17.7	$-17.7\,[-23.3, -12.3]$
	Gemini	100	80.0	63.0	-17.0	$-17.0\,[-23.0, -11.0]$
	Qwen	100	80.0	58.0	-22.0	$-22.0\,[-28.7, -15.7]$

Why the LongBench v2 net drop is small. If a question depended on one indispensable location and mask starts were uniform, each trial would remove that location with probability 20%, and survival under all three masks would be $0.8^3 = 51.2\%$. Observed all-mask survival among Full-correct questions is 75.3% for DeepSeek, 81.7% for Gemini, and 75.4% for Qwen.

Net accuracy also hides two-way flips. Conditional on a Full-correct opportunity, mask trials become incorrect at 12.8%, 8.0%, and 12.3%, still below the 20% single-point prediction. Reverse flips, where an incorrect Full answer becomes correct, cancel 3.7–6.7 points of that one-sided loss. Many supporting spans are therefore not behaviorally unique.

Loong falls by 17.0–22.0 points under the same protocol, so a small Mask20 effect is not a general property of long-context benchmarks.

I EVIDENCE CONSTRUCTION DETAILS

The details below describe the MuSiQue construction procedure and its adaptation to FRAMES.

I.1 PLANS AND TRACEABILITY

The planner emits a structured record containing a shared entity registry, question substitutions, hop dependencies, and one gold realization plan per hop. Each realization plan contains evidence units with stable identifiers, atomic facts, and a composition rule. The planner requires at least $\lceil |H|/2 \rceil$ hardening targets, including the final hop; each selected target must have at least two planned evidence units. For example, rather than directly stating that Winter Letters is displayed at Harbor Museum, the plan uses two complementary premises: "Winter Letters is listed under entry E7 in the exhibition catalogue" and "Entry E7 names Harbor Museum as the display venue." Together they link the manuscript to the museum; neither premise alone establishes that relation. These units support the same hop in the original question graph, so the annotated hop count is unchanged. Each unit is linked to its corresponding text.

Later records keep the link from a graph fact to its text:

- **Competitor plan:** branch and fact identifiers, the protected hop, relation status, entity bindings, and host exclusions.
- **Route plan:** retained fact identifiers and assigned host passages, linked to the compiled fact plan.
- **Realization:** one record per planned source identifier, with realized text spans and, for gold hops, individual evidence-unit spans.

The main MuSiQue evaluation removes a supporting paragraph, not each premise within it.

I.2 Flexible and Strict Competitor Contracts

The *protected hop* h indexes the support-deletion target, not a hop exempt from deletion; the *break hop* b locates the relation mismatch. The full pipeline targets every question hop, independently of the gold premise-decomposition subset and without using behavioral Drop outcomes.

Flexible. The planner proposes a non-entailing relation at $b = h$ with a complete continuation. Where safe, it also proposes a separate path with $b \in \mathrm{Desc}_G(h)$: the protected relation matches, but a later dependency does not. Prefixes that introduce ambiguous referents or another valid answer are rejected. Fact-level pruning may leave fragments.

Strict. The sole mismatch must satisfy $b = h$, using a type-compatible, non-entailing predicate. Only $R_h = \{h\} \cup \mathrm{Desc}_G(h)$ is regenerated; other dependencies inherit exact gold bindings. Descendants preserve required relations but may use alternative entities. Unsafe paths are removed atomically. A change of narrative framing alone is not a break; neither mode labels competitors as wrong.

Example. Consider kinship, authorship, and display as separate hops in Figure 2, and target authorship. A Strict candidate can inherit Liam's kinship, state that he *translated* another manuscript, and give its display venue: translation does not entail authorship, while display remains matched. Flexible also allows this pattern. A separate Flexible proposal could instead match authorship but replace *displayed* with *stored*. Strict excludes this downstream break for the same target. Flexible still rejects it if the added manuscript makes the question's referent ambiguous. These hypothetical plans illustrate constraints, not extra facts in the figure. No downstream-break option exists at a terminal hop.

Coverage and ablations. Two independent Strict candidates with distinct entities and answers are proposed per hop; at least one complete path per hop must survive pruning. Flexible fragments cannot fill missing Strict coverage. Full combines both families; +Flexible and +Strict add each separately to Gold-only. Removing Strict in an ablation removes its coverage requirement. These are construction contracts, not guarantees of behavioral necessity; routing and acceptance checks follow below.

I.3 Routing, Realization, and Validation Boundaries

A route must respect each fact's explicit forbidden-passage list and account for every retained fact. A host is one dataset passage, not necessarily an entire source document. Flexible plans must include the protected support in each competing fact's forbidden list. In the v8.6.2 backend, Strict plans are validated separately and do not universally enforce that inclusion, so Strict facts may share the protected gold-support passage. A Strict path with at least two generated facts uses at least two hosts, and its near-miss host cannot also contain its generated exact continuation. Base chains with at least three facts also use at least two hosts. These are host-level rules, not a universal token-distance requirement. MuSiQue Drop removes the entire rewritten support passage, including any co-located gold and competitor text. Neither generated nor inherited competitor evidence is therefore guaranteed to survive a deletion; planned-chain completeness is not a post-deletion survival check.

The realizer retains the original title and text in order, adding complete sentences at safe boundaries. It must not strengthen the planned claim's scope or create an unplanned conclusion. Mechanical checks cover source identifiers, planned-unit coverage, and exact span occurrence. If original-text preservation fails, a repair reconstructs the source body and appends recovered additions; anaphora repair can likewise move additions. Generated entities do not inherit unstated properties of the original entities in the retained source text.

Blocking structural checks include graph linkage, entity-registry consistency, executable evidence plans, route coverage, and realization correspondence. The semantic audit rejects material question-contract violations, ambiguous bindings, and unintended complete answer paths. Salience and directional-hardening scores are recorded separately. In v8.6.2, missing unit identifiers or invalid text-span mappings block acceptance, but premise-separation and directional-hardening issues are non-fatal diagnostics. Structural coverage therefore does not certify that every realized premise is complementary or behaviorally necessary. Human spot checks of gold–question match and competitor non-entailment are reported in Appendix J.1 and Appendix J.2.

After realization, competitor pruning is limited to two iterations. Each removal is followed by a coverage check, and items that lose required coverage are excluded. Full/Drop outcomes are not used for retention; behavioral auditing begins only after construction and validation.

I.4 FRAMES Adapter and Hop-Level Audit

FRAMES supplies questions and Wikipedia links, not hop labels. We assemble each article from those links. When tables are kept, HTML rows that overlap the question or the reference answer are flattened and appended in source order. Hop proposal, realization, and source-side evaluation use this same text.

A model then proposes a linear chain: each hop quotes a span in a source passage, and each later hop points to the previous one. Several hops may quote the same article. Checks confirm that the fields are filled and that each quotation occurs in the assembled text. After this conversion, the same MuSiQue construction procedure generates the gold and competitor text. Only entity-answer plans are realized: numbers are not rewritten and derived answers are not recomputed. Two items with stale values are dropped, leaving 81 questions and 208 hops per side. Article bodies are capped at 24,000 characters, plus up to 1,500 characters of selected table rows. This is not the full FRAMES test set.

For the reported BNR analysis, DeepSeek V4 Flash, Gemini 3.6 Flash, and Qwen 3.7 Max answer source and constructed items under Full and every single-hop Drop. The reported analysis uses three trials per condition. DeepSeek V4 Pro scores each response as 1, 0.5, or 0 from the question, reference, and answer, without the context (Appendix K).

Each Drop removes one inferred hop: all recorded evidence-unit spans of that hop on the constructed side, or the quoted evidence span on the source side. Deletion follows the exact span, then removes string-matched endpoint cooccurrences. This cleanup may leave residual paths or remove incidental cooccurrences. A hop's units are deleted jointly, so this audit does not test each split premise independently. Supplementary matched-background and common-Full analyses are reported in Appendix I.5; the background control is separate from the BNR calculation.

BNR recomputation from recorded trials. We recompute Equation 1 from the recorded per-condition trial scores. A score of 1 is correct; scores of 0.5 and 0 are incorrect. We require every expected hop condition, matched to the intervention audit, and all three Full/Drop scores to be present. The available complete panels contain 80, 81, and 79 source–constructed pairs for DeepSeek, Gemini, and Qwen, respectively, covering 205, 208, and 202 hop conditions per side. The remaining 1, 0, and 2 items have incomplete records and are excluded. Only complete three-trial panels are retained.

For each trial t and side s, define $F_i^{s,t} = \mathbf{1}[\mathrm{score}_{i,\mathrm{Full}}^{s,t} = 1]$ and $D_{ih}^{s,t} = \mathbf{1}[\mathrm{score}_{ih,\mathrm{Drop}}^{s,t} = 1]$. On the complete-panel population, we first select $\mathcal{I}_{s,t}^{+} = \{i : F_i^{s,t} = 1\}$, then average $1 - D_{ih}^{s,t}$ over hops within each selected item and over items. Full accuracy uses the complete-pair denominator, whereas BNR uses $|\mathcal{I}_{s,t}^{+}|$. The main summary is the arithmetic mean of the three trial-level Full accuracy and BNR values reported below. No majority vote, best-trial selection, or pre-binarization score average is used. Each side has its own Full-correct subset, so the reported BNR differences are not restricted to items answered correctly on both sides. The same metric is used for MuSiQue, but the intervention, judge configuration, and complete-case population remain distinct.

Table 13: FRAMES BNR recomputed independently for each recorded trial (percent). O/R denote source/constructed items; n^+ is the corresponding Full-correct question count. Complete-panel counts are 80, 81, and 79 throughout. Avg@3 Full and BNR in Section 6.2 are the arithmetic mean of these trial-level metrics.

Model	Trial	n^+		BNR	
		O	R	O	R
DeepSeek V4 Flash	1	79	59	7.81	89.83
	2	79	65	6.75	88.03
	3	79	62	8.44	88.87
Gemini 3.6 Flash	1	80	71	5.21	89.18
	2	79	69	7.59	87.78
	3	80	69	7.08	91.09
Qwen 3.7 Max	1	77	51	6.93	89.87
	2	76	58	8.33	88.62
	3	78	53	9.19	91.19

Table 14: FRAMES BNR on trials with Full correctness on both sides. Counts give $|\mathcal{I}_t^{\cap}|$ for trials 1/2/3; O/R are source/constructed. BNR is percent; the paired increase and its interval are percentage points.

Model	Counts	BNR O	BNR R	Increase [95% CI]
DeepSeek V4 Flash	59/65/61	6.80	88.85	82.1 [75.2, 88.2]
Gemini 3.6 Flash	70/67/68	5.75	89.30	83.5 [77.4, 89.1]
Qwen 3.7 Max	49/56/52	7.79	89.57	81.8 [74.7, 88.2]

I.5 FRAMES Matched Controls and Sensitivity

Two supplementary checks address different confounds: changes in the Full-correct population and nonspecific damage from deleting text. All analyses below reuse saved scores; no additional model calls are made. Records are deduplicated by retaining the latest entry before filtering for completeness. The 80/81/79 complete source–constructed panels are unchanged.

Common-Full subset. For trial t, let $\mathcal{I}_t^{\cap} = \{i : F_i^{O,t} = F_i^{R,t} = 1\}$. We recompute both sides' BNR on this same set, retaining the main paper's rule that only score 1 is correct; a Drop score of 0.5 is therefore a failure. We average hops within each question, questions within each trial, and then the three trial estimates equally. Trial indices align the recorded attempts, not shared random draws. This differs from pooling all retained item–trials or applying a threshold to the three-trial average score.

The intersection retains 72, 71, and 61 distinct questions, contributing 185, 205, and 157 item–trials. The joint Full criterion excludes 55, 38, and 80 item–trials, respectively. BNR increases remain 81.8–83.5 points (Table 14).

Relative to each side's own Full-correct population, restricting to the common set changes the increase by +0.81, +0.83, and +0.03 points, respectively, with paired 95% intervals of $[-1.03, 2.99]$, $[-0.62, 2.62]$, and $[-2.39, 2.54]$. These intervals directly estimate the change in effect, rather than comparing the overlap of separate intervals. The result supports persistence on the jointly answerable subset, not equivalence of populations or replacement of the full-cohort results.

Background control and condition pairing. The separately run background control selects background text using lexical protection of all recorded hop evidence, subjects, answers, and, on the constructed side, competitor spans. Its budget is the summed character length of the target hop's recorded spans, with a 25% tolerance and an attempt to follow their passage distribution; it is not a token-level or exact-position match. This budget can differ from the actual Drop removal, which includes repeated occurrences and relation-closure cleanup. Short-span cases may use a flagged sentence-fragment fallback.

Table 15: Constructed FRAMES background-deletion sensitivity. n/K counts questions/paired hop conditions, not model calls. All score columns are item-macro condition means multiplied by 100; $C - D$ is Control minus Drop, with a paired question-bootstrap interval. Length bands additionally require sentence-aligned deletion.

Model	Filter	n/K	Full	Control	Drop	$C - D$ [95% CI]
DeepSeek V4 Flash	All feasible	80/204	77.50	78.09	9.81	68.3 [60.9, 75.6]
	$\pm 25\%$	80/191	77.50	77.74	7.92	69.8 [62.7, 76.8]
	$\pm 10\%$	56/78	74.40	77.38	9.33	68.1 [57.9, 77.7]
	$\pm 5\%$	30/38	74.44	80.56	12.78	67.8 [53.3, 81.1]
Gemini 3.6 Flash	All feasible	81/207	86.01	83.82	9.37	74.5 [66.9, 81.6]
	$\pm 25\%$	81/194	86.01	84.40	7.72	76.7 [69.2, 83.8]
	$\pm 10\%$	57/80	85.38	83.92	10.04	73.9 [63.5, 83.8]
	$\pm 5\%$	30/38	84.44	83.15	13.33	69.8 [53.9, 84.8]
Qwen 3.7 Max	All feasible	79/201	68.57	67.98	8.09	59.9 [50.8, 68.6]
	$\pm 25\%$	79/188	68.57	68.11	6.19	61.9 [53.2, 70.5]
	$\pm 10\%$	55/77	68.79	68.08	8.38	59.7 [47.8, 70.8]
	$\pm 5\%$	30/38	71.67	68.33	7.78	60.6 [42.8, 77.2]

We align Control and Drop by sample ID and the explicit hop index. Control means are stored as a list: we recover its condition names using the producer's lexicographic order of feasible conditions, requiring exact per-item agreement between metadata, score-list length, and missing-condition counts. We do not shift list positions across infeasible conditions. Every selected comparison uses the same question–hop conditions for Full, Control, and Drop, averaging condition means within a question and then questions equally. These are mean judge scores on a 0/0.5/1 scale, not binary accuracy or BNR. Full and Drop have three recorded scores per condition; control trial counts are not committed, so their three-trial completeness cannot be verified or reconstructed from an average score.

Constructed-side control coverage is 204/205, 207/208, and 201/202 conditions; all feasible constructed-side deletions are sentence-aligned. Original-side coverage is 198/205, 201/208, and 196/202, with nine feasible fragment deletions per model; these are excluded from sentence-only sensitivity checks. Missing controls are omitted together with their paired Drop conditions, not counted as failures. On all feasible constructed-side pairs, Control–Full mean-score differences are +0.59, -2.19, and -0.58 points, with paired 95% intervals [-3.68, 5.14], [-4.42, -0.13], and [-5.51, 4.23]. Thus we do not claim that background deletion has exactly zero effect, even though its damage is much smaller than the targeted-Drop effect.

Actual-length sensitivity. Using deletion metadata alone, we additionally require a whole-sentence control and $|L^K_{ih}/L^D_{ih} - 1| \leq \epsilon$, where L^K_{ih} is the control's removed character count and L^D_{ih} is the actual Full–Drop character difference, including closure. We report all three nested bands, $\epsilon \in \{0.25, 0.10, 0.05\}$, without selecting conditions by model scores. This is an offline, post-hoc robustness check, not a new set of model runs.

All three constructed-side Control–Drop intervals remain above zero in every band (Table 15). Even the 5% band retains 38 conditions from 30 questions per model and a 60.6–69.8-point gap. The tighter subsets change the question population and do not establish exact matching of passage position, deletion geometry, or semantic relevance. Together with the common-Full check, the results support increased measured evidence dependence that cannot be attributed solely to differing answerable populations or nonspecific text deletion; they do not establish universal necessity of every evidence unit.

Uncertainty and reproducibility. All supplementary intervals use 10,000 whole-question bootstrap resamples and percentile endpoints, with base seed 20260925 and fixed model/cohort offsets. For BNR, each resampled question retains all hops, trials, both sides, and both Full-correct definitions; the intersection and Avg@3 estimates are recomputed within every resample. For background scores, each side and length cohort is resampled separately, keeping its Full/Control/Drop triple paired. These intervals reflect question-sampling variability conditional on the saved scores, not

independent uncertainty in model trials, judges, or generated contexts. No multiplicity correction or equivalence test is claimed. The standard-library script `offline_matched_analysis.py` records input and script hashes, per-condition matches, exclusion lists, all original- and constructed-side sensitivity results, and bootstrap settings in its JSON output. This is independent of the MuSiQue background-control experiment in Appendix B.

J Human Spot Checks

After adjudication, spot checks accepted 98/100 frozen MuSiQue items and 50/50 LongBench items. Protocol and disagreements are below.

J.1 Human Spot Check of RealHop-MuSiQue

After freezing the 790-item evaluation set, we sampled 100 RealHop-MuSiQue instances, hop-stratified at random with quotas 54/31/15, matching the 430/244/116 split of 2-, 3-, and 4-hop items. Annotators received the question, labeled answer, realized context, gold evidence, and inserted competitor sentences; model outputs were withheld.

For each item, annotators judged whether the written gold evidence matches the question and supports the labeled answer, and whether each inserted competitor relation fails to entail the required gold relation, so those sentences do not yield that answer. Labeled answers are newly coined names bound to the constructed entities; retained source text concerns the original entities and cannot recover these answers.

Two annotators labeled all 100 items. They marked 4 and 5 items as questionable, and agreed on 95/100: both accepted 93 items, both rejected 2, and disagreed on 5. A third annotator judged the disagreements correct. After that, two items (2/100) have gold-chain errors from a pronoun with the wrong antecedent; the competitor sentences were not judged to give the gold answer.

J.2 Human Spot Check of RealHop-LongBench

We applied the same protocol to 50 randomly sampled RealHop-LongBench instances from the frozen 216-item set. Constructed answers are new names absent from the unmodified carrier text, so residual carrier passages cannot recover them.

Two annotators labeled all 50 items. They marked 1 and 0 items as questionable, and agreed on 49/50: both accepted 49 items and disagreed on 1. A third annotator judged the disagreement correct. After that, all 50 items were accepted.

The higher acceptance rate is consistent with a difference in realization. RealHop-MuSiQue inserts complete sentences while retaining source text in order, so a later demonstrative can bind to an intervening source entity; the two residual errors in the MuSiQue sample are of this kind. RealHop-LongBench rewrites the local neighborhood of each insertion, so gold and competitor relations are realized with explicit antecedents rather than cross-sentence pronouns into unmodified carrier text.

K Evaluation Prompts

The scores in the paper use the frozen templates below. The judge sees the question, reference answers, and model response, and does not see the document. For RealHop-MuSiQue, HotpotQA, 2WikiMultiHopQA, and RealHop-LongBench, only verdict `correct` counts as correct. `partial` is not treated as correct. FRAMES uses the same judge template; its recorded `correct`/`partial`/`incorrect` verdicts map to 1/0.5/0. The reported FRAMES Full accuracy and BNR binarize those recorded scores, counting only 1 as correct (Appendix I.4).

Multi-hop solver. RealHop-MuSiQue, source MuSiQue, HotpotQA, and 2WikiMultiHopQA use the official LongBench v1 HotpotQA template. FRAMES uses this template for the reported Full and Drop conditions.

```
Answer the question based on the given passages. Only give me the answer
     and do not output any other words.

The following are given passages.
{context}

Answer the question based on the given passages. Only give me the answer
     and do not output any other words.

Question: {input}
Answer:
```

Long-context open-ended solver. REALHOP-LONGBENCH Full, Drop, and random-window trials use:

```
Please read the following text and answer the question below.


{context}


Question: {question}

Give only the concise answer. Do not provide an explanation or list
    alternatives.
```

Normalized exact matches are accepted without a judge call; remaining responses use the short-answer judge below.

Long-context multiple-choice solver. Matched source questions and the REALHOP-LONGBENCH MCQ evaluation use the same official LongBench v2 0-shot multiple-choice template. The constructed evaluation keeps a frozen A–D option order, with the gold terminal and three planned competitor terminals as the four choices.

```
Please read the following text and answer the question below.

$DOC$

What is the correct answer to this question: $Q$
Choices:
(A) $C_A$
(B) $C_B$
(C) $C_C$
(D) $C_D$

Format your response as follows: "The correct answer is (insert answer
    here)".
```

Short-answer judge. The short-answer judge uses the following system message:

```
You are a calibrated grader for short-answer multi-hop QA. Grade the
    response relative to the QUESTION, not as a standalone sentence. The
     reference answer may be more verbose than the minimum answer
    actually requested by the question.
Reply with a SINGLE LINE of JSON in this exact format and NOTHING ELSE:
{"verdict": "<correct|partial|incorrect>", "reason": "<one short
    sentence>"}
Grading rules:
  - correct: the response fully fills the answer slot requested by the
    question and is semantically equivalent to a reference. Allow
    paraphrases, explicit aliases, harmless generic prefixes or suffixes
```

```
    , and omission of participants or roles already unambiguously fixed
    by the question.
  - A pronoun or short predicate can be complete when the question
    itself uniquely supplies its participant. For example, for 'How were
     X expelled?', 'defeated' can be complete; do not require repeating
    X or the actor unless the question asks who.
  - Preserve answer-defining distinctions not supplied by the question.
    A wrong entity, location, date, direction, number, role, scope, or
    predicate is incorrect.
  - Do not accept a merely related detail in place of the requested slot
    . For example, if asked what role or category someone won for, a
    work title alone is partial or incorrect.
  - Extra information is allowed unless it contradicts or changes the
    answer.
  - partial: the response gives a genuinely correct part of the
    requested answer but omits a detail needed to fill the requested
    slot.
  - incorrect: the response is wrong, unsupported, contradictory,
    refuses to answer, or contains only topical overlap.
  - Judge only from the ordinary semantics of the question, references,
    and response. Do not assume hidden source context or an unstated
    alias.
Do not output anything other than the JSON line.
```

User message:

```
Question:
{question}

Reference answer(s) (any one suffices):
{references}

Model response:
{response}

Now grade the model response.
```