CLAIMWEAVER: QUERY-CONDITIONED EVIDENCE ORGANIZATION FOR MULTI-HOP RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) can retrieve relevant evidence yet fail to connect the facts needed to answer a question. This gap is acute in multi-hop question answering, where passages mix answer-bearing facts with duplicates and distractors, while the required dependencies span multiple documents. Free-form notes condense this context, but decisions about which facts to discard, combine, or connect remain implicit in prose. We introduce CLAIMWEAVER, a query-conditioned evidence organization framework over atomic claims. It transforms retrieved passages into source-attributed claims and builds a compact structure in which irrelevant claims are suppressed, redundant claims are consolidated, and complementary claims are linked by directed bridges. This structure persists across retrieval rounds, focuses subsequent search, and provides answer generation with an explicit evidence path. A composite reinforcement-learning objective learns this behavior without gold organization labels. Across seven QA benchmarks, CLAIMWEAVER achieves 50.5 EM and 56.4 F1. Compared with a configuration-matched free-form-note control, it improves average EM by 4.5 points across four multi-hop benchmarks while maintaining average single-hop EM. Human audits assess decomposed-claim faithfulness and pairwise relations, while retrained ablations and fixed-evidence interventions examine the role of organization. These results support query-conditioned claim organization as an effective interface between retrieval and multi-hop reasoning.

1 INTRODUCTION

Retrieval-augmented generation (RAG) grounds large language model (LLM) outputs in external evidence (Lewis et al., 2020; Guu et al., 2020; Gao et al., 2023). Even when relevant evidence is retrieved, its organization can remain a bottleneck: the generator must identify useful facts and determine how they connect before composing an answer. Multi-hop questions make this bottleneck especially important because the answer depends on a sequence of connections distributed across documents (Yang et al., 2018; Trivedi et al., 2023).

Retrieved context is structurally heterogeneous. A single passage can contain several claims with different roles; the same fact may recur across sources; topically plausible passages can support the wrong entity or predicate; and critical evidence may be useful only as an intermediate bridge. These roles also change with the query. A birthplace fact is essential when asking where a director was born and irrelevant when asking which film the director made. Treating all retrieved text uniformly hides these distinctions, allowing repetition and distractors to compete with the evidence path required for the answer.

Existing evidence-processing methods summarize or compress retrieved passages into natural-language notes (Yu et al., 2024; Dai et al., 2025; Xu et al., 2023). These methods reduce context length or highlight answer-supporting information, but encode evidence selection and composition within the generated text. An explicit claim structure makes these decisions separately addressable: which facts to retain, which to consolidate, and which to use in sequence. Corpus-level graphs and hierarchical indices broaden retrieval connectivity (Edge et al., 2024; Sarthi et al., 2024; Gutiérrez et al., 2024), yet the role of each retrieved fact remains specific to the current query. Multi-hop

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.

RAG therefore needs an online organization layer that converts retrieved evidence into a query-conditioned reasoning state while preserving fine-grained provenance.

We introduce CLAIMWEAVER to learn this organization layer. Its central representation is a structure over atomic, source-attributed claims. Claim granularity makes individual facts addressable: distracting claims can be suppressed, repeated claims consolidated without losing provenance, and complementary claims connected through directed bridges. The resulting structure reduces repeated evidence while preserving intermediate links and their sources. It records how the selected facts should be used together, making the evidence path and individual organization decisions inspectable.

CLAIMWEAVER embeds this representation in an agentic retrieve–organize–reason loop. The structure evolves as new evidence arrives and persists as memory for targeted retrieval and answer generation. Because gold organization labels are rarely available, a composite reinforcement learning objective learns the complete behavior from answer correctness, claim grounding, and structural compliance.

Across seven QA benchmarks, CLAIMWEAVER achieves 50.5 EM and 56.4 F1. It improves average EM by 4.5 points over a configuration-matched free-form-note control across four multi-hop benchmarks while maintaining average single-hop EM. Human audits assess decomposed claims and predicted relations. Among the individual operation ablations, removing chaining yields the largest multi-hop drop. Fixed-evidence interventions further show that downstream answers depend on the organization state. These analyses connect the representation to answer behavior through complementary evaluations of accuracy and evidence use.

Our contributions are as follows:

1. We formulate query-conditioned evidence organization over atomic claims, addressing the filtering, redundancy, and dependency structure that free-form evidence representations leave implicit.

2. We introduce an agentic RAG framework that builds a persistent, provenance-aware claim structure and learns to organize evidence without gold organization labels.

3. Across seven QA benchmarks, we show gains concentrated on multi-hop questions and examine the learned organization through human audits, retrained ablations, fixed-evidence interventions, transfer, and cost analysis.

2 RELATED WORK

2.1 RETRIEVAL-AUGMENTED GENERATION

RAG systems combine a retriever with a generator to ground responses in external knowledge (Lewis et al., 2020; Guu et al., 2020; Karpukhin et al., 2020; Ram et al., 2023). Concatenating top-k passages exposes the generator to redundancy and topically related distractors (Shi et al., 2023; Yoran et al., 2024). Adaptive and iterative variants (Asai et al., 2024; Jiang et al., 2023; Trivedi et al., 2023) improve when and how evidence is retrieved. Our focus is the representation maintained between retrieval and answer generation.

2.2 EVIDENCE STRUCTURING IN RAG

Reading notes judge relevance and summarize key evidence (Yu et al., 2024; Dai et al., 2025), while compression methods reduce context length and noise (Xu et al., 2023; Jiang et al., 2024). BRIEF (Li et al., 2025b) uses atomic propositions to construct synthetic supervision for query-aware compression of multi-hop evidence. ClaimWeaver instead exposes filtering, merging, and directed connections as separate decisions in the generated state, making the organization operations individually addressable. Corpus-level graphs and hierarchical indices broaden cross-document retrieval connectivity (Edge et al., 2024; Sarthi et al., 2024; Gutiérrez et al., 2024; 2025). ClaimWeaver complements such indices with query-conditioned operations on currently retrieved claims.

2.3 Claim Decomposition for Verification and Retrieval

Atomic claim decomposition supports factuality evaluation and claim verification by separating text into propositions that can be checked against evidence (Min et al., 2023; Wei et al., 2024; Kamoi et al., 2023). Dense X Retrieval (Chen et al., 2024) studies propositions as retrieval units. Ki-RAG (Fang et al., 2025) decomposes documents into knowledge triples and iteratively retrieves information that fills reasoning gaps. PropRAG (Wang & Han, 2025) guides online retrieval with beam search over proposition paths. ClaimWeaver learns to filter, merge, and bridge the currently retrieved claims within a persistent query-conditioned state. This organization interface integrates evidence selection, redundancy resolution, and directed composition into an outcome-trained policy.

2.4 Reinforcement Learning for RAG

Outcome-based RL improves LLM reasoning (DeepSeek-AI, 2025; Shao et al., 2024) and enables agentic RAG to interleave search and reasoning (Jin et al., 2025; Song et al., 2025). EviNote-RAG (Dai et al., 2025) rewards notes that entail the gold answer, while VERITAS (Xu et al., 2026) rewards turn-level reasoning faithfulness. ClaimWeaver combines the answer signal with claim-grounding and format rewards to learn explicit claim organization without gold structure labels.

3 Method: ClaimWeaver

3.1 Overview

At retrieval round t, the policy π_θ receives $x_t = (q, D_t, M_{t-1})$: the question, current passages, and prior claim memory. It extracts source-attributed claims, predicts relevance and sparse pairwise relations, resolves them into a structure, and writes per-claim usage notes. It then chooses $a_t \in \{\mathrm{search}(u_t), \mathrm{answer}(y_t)\}$. Overlap-connected claims are merged before directed bridges are chained; retained claims without an edge remain isolated. The resulting memory guides both subsequent search and final answering (Figure 1).

Persistent state and environment. Memory serializes surviving claim texts, source IDs, directed links, and per-claim notes in the policy context. Let $\mathcal{C}(M_{t-1})$ denote the claims recorded in earlier memory. The update can be written as

$$\mathcal{C}_t^{\mathrm{new}} \sim \pi_\theta^{\mathrm{dec}}(\cdot \mid q, D_t), \quad \widetilde{\mathcal{C}}_t = \mathcal{C}(M_{t-1}) \cup \mathcal{C}_t^{\mathrm{new}}, \tag{1}$$

$$(V_t, E_t, N_t) \sim \pi_\theta^{\mathrm{org,note}}(\cdot \mid q, \widetilde{\mathcal{C}}_t, M_{t-1}), \quad M_t = \mathrm{serialize}(V_t, E_t, N_t). \tag{2}$$

Here V_t includes source-attributed resolved claims, E_t contains directed bridges, and N_t contains their notes; the superscripts denote stages of the same policy. Old and new claims are considered jointly, so new evidence may revise relevance, merge with an earlier claim, or extend a chain. The model generates search queries; an external frozen retriever returns $D_{t+1} = \mathrm{Retrieve}(u_t)$. Retrieved passages enter the trajectory as environment observations. Appendix B.6 gives a two-round illustration.

3.2 Claim Decomposition

Given a query q and a set of retrieved passages $D = \{d_1, \ldots, d_k\}$, the first step of ClaimWeaver maps the raw text into a set of atomic claims $\mathcal{C} = \{(c_i, s_i)\}_{i=1}^n$, where c_i denotes the textual proposition and $s_i \in D$ specifies its provenance.

An atomic claim must be (i) **atomic**: exactly one factual proposition; (ii) **self-contained**: all pronouns and implicit references disambiguated; (iii) **source-attributed**: linked to its originating passage; and (iv) **verifiable**: independently checkable as true or false, excluding opinions or meta-information. The policy model π_θ is prompted to extract claims satisfying these criteria. Figure 2 illustrates the resulting claims and their organization.

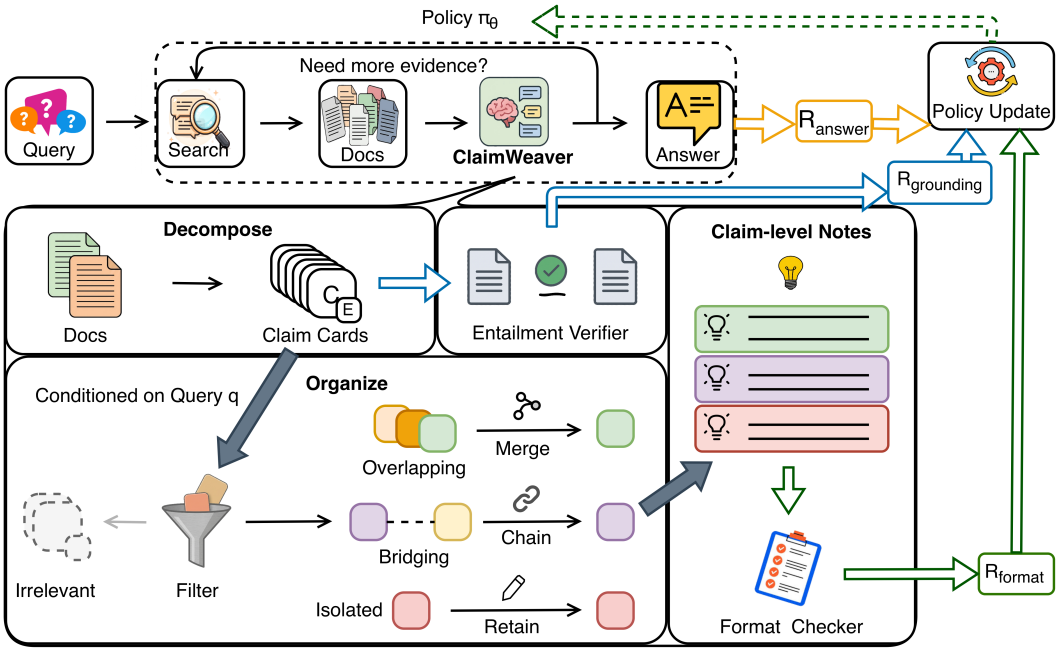


Figure 1: Overview of CLAIMWEAVER. The retrieve–organize–reason loop (top) maintains per-claim notes. The bottom panel summarizes organization operations; Figure 2 shows explicit claim nodes and links. GRPO uses answer, claim-grounding, and format rewards. The entailment verifier runs only during training.

3.3 RELATION-AWARE CLAIM ORGANIZATION

Organization jointly processes old and new claims in Equation 2 through triage, relation discovery, and resolution.

Relevance Triage. Conditioned on the query and the full claim set, the policy is instructed to exclude claims lacking both direct answer-bearing utility and potential value as an intermediate bridge, and to retain uncertain claims. This conservative rule prioritizes preserving possible reasoning links.

Query-Conditioned Relation Discovery. Following triage, the model identifies redundancy and query-dependent connections among the retained claims. For factoid QA, we distill broader discourse relations (Radev, 2000) into two actionable pairwise relations:

- **Overlapping**: Claims that provide semantically equivalent or redundant information with respect to q.
- **Bridging**: A directed pair $c_i \rightarrow c_j$ in which c_i provides an entity or premise for using c_j to answer q.

A relevant claim with no retained pairwise relation is treated as **isolated**, a residual node state. This vocabulary targets factoid composition (Yang et al., 2018; Trivedi et al., 2022) within the broader space of discourse relations (Radev et al., 2004; Zhang et al., 2002). Entity, temporal, and specificity differences must be resolved before claims are treated as redundant. Contradictory sources require separate adjudication because corpus-level inconsistencies can persist across retrieved passages (Semnani et al., 2025).

Figure 2 illustrates a bridge from director identity to birthplace. If the query instead asks who directed *Parasite*, the first claim is answer-bearing and the birthplace claim has no query utility.

Structural Resolution. The model applies four operations: **Exclude** removes non-contributory claims; **Merge** consolidates each overlap-connected component with unioned provenance; **Chain**

Query: What is the birth country of the director of *Parasite*?

(a) Extracted claims and sources

Doc1
c_1: *Parasite* was directed by Bong Joon-ho.
c_2: *Parasite* won four Academy Awards.

Doc2
c_3: Bong Joon-ho was born in South Korea.

Doc3
c_4: Bong Joon-ho directed *Parasite*.

(b) Organized memory and notes

Exclude c_2; merge $c_1, c_4 \rightarrow C_1$.

C_1: *Parasite* was directed by Bong Joon-ho. [Doc1, Doc3]
Entry point; 2 sources.
Use: identify the queried director.

⇓ Chain $C_1 \rightarrow c_3$ via Bong Joon-ho

c_3: Bong Joon-ho was born in South Korea. [Doc2]
Answer-bearing; 1 source; follows C_1.
Use: supply the country after C_1.

Answer: South Korea.

Figure 2: Illustrative organization of extracted claims. The award fact is irrelevant to this query. Equivalent claims c_1 and c_4 merge into C_1 with both sources retained; c_3 keeps its identity. The bridge orders evidence use from director identity to birthplace, as recorded in the usage notes.

projects directed bridges onto the merged nodes; and **Retain** preserves isolated claims. Merging precedes chaining, so a merged node may also participate in a bridge. The format checker sets $R_{\text{format}} = 0$ for self-edges and directed cycles.

3.4 Claim-Level Note Generation

The model serializes surviving claims and their relations as natural-language notes. Each note records the claim's role, distinct source documents, connections, and a usage instruction. Multiple passages from one document count as one source. Roles include entry point, bridge, answer-bearing, supporting, and independent evidence. An incomplete chain can prompt a follow-up search; a complete chain supports answering. Notes persist across rounds and are updated jointly with newly retrieved claims. Figure 2 illustrates the notes; Appendix B gives the full schemas and a two-round trajectory.

3.5 End-to-End RL Training

For each query, the unified policy generates a trajectory τ containing search queries, claims, organization, notes, decisions, and an answer, interleaved with passages returned by the external retriever.

Reward Design. We optimize π_θ with a composite trajectory-level reward for answer correctness, claim grounding, and format compliance:

$$R(\tau) = \lambda_1 R_{\text{answer}} + \lambda_2 R_{\text{grounding}} + \lambda_3 R_{\text{format}}. \tag{3}$$

Here, $R_{\text{answer}} \in \{0, 1\}$ indicates whether the final answer matches the gold answer under exact match, while $R_{\text{grounding}}$ encourages faithful claim decomposition by measuring whether each claim c_i is entailed by its attributed source passage s_i using a frozen Natural Language Inference (NLI) model:

$$R_{\text{grounding}} = \frac{1}{n} \sum_{i=1}^{n} P_{\text{NLI}}(\text{entail} \mid s_i, c_i), \qquad n > 0, \tag{4}$$

where n counts decomposed claims across all retrieval rounds. The frozen verifier is DeBERTa-v3-large fine-tuned on MNLI (He et al., 2023). Averaging over claims can favor selective extraction of high-confidence claims. The answer reward supplies complementary coverage pressure because omitting a required bridge can prevent a correct answer.

R_{format} is a binary reward that checks required markers for search, claim decomposition, relational states, notes, and the final answer. The grounding term supervises source entailment during decomposition; relevance and relation selection are learned from answer outcomes under the grounding and format constraints. Merging and note generation are instructed to preserve the source-supported content.

Optimization via GRPO. We optimize the policy with critic-free GRPO (Shao et al., 2024). For each query, rewards from $G = 8$ sampled trajectories are standardized within the group; a small ϵ stabilizes zero-variance groups. These advantages weight policy-generated tokens, excluding retrieved passages. A KL penalty against the frozen reference policy regularizes the update. We use $(\lambda_1, \lambda_2, \lambda_3) = (1.0, 0.3, 0.1)$ and KL coefficient 0.001. The policy jointly learns retrieval triggers, organization, and answering; Appendix C gives the training configuration.

4 EXPERIMENTS

We evaluate answer quality, organization operations, intermediate structures, transfer, failure modes, and the accuracy–cost trade-off.

4.1 EXPERIMENTAL SETUP

We evaluate four multi-hop benchmarks—HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023)—and three single-hop benchmarks—Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023). Following Search-R1 (Jin et al., 2025), we train on NQ and HotpotQA. Evaluation follows EviNote-RAG (Dai et al., 2025): we sample 500 examples per public evaluation split with seed 42; Bamboogle uses all 125 validation examples. The other five benchmarks are held out during training. We report Exact Match (EM) and token-level F1 with shared answer normalization and alias handling.

Controlled systems use Qwen2.5-7B-Instruct (Yang et al., 2024), E5-base-v2 retrieval (Wang et al., 2022), and shared answer prompts. Default retrieval returns three passages per query. Prompting baselines include Search-o1 (Li et al., 2025a); three graph-inspired local controls use top-50 E5 candidates. EviNote-RAG shares ClaimWeaver's training examples, GRPO schedule, retrieval-turn and context/output limits, decoding, and three training seeds, but retains its native free-form notes and reward. Their scores are averaged over these three seeds. This compares complete systems: reward objectives, realized search traces, and computation differ. Released Search-R1 (Jin et al., 2025), R1-Searcher (Song et al., 2025), and ReARTeR (Sun et al., 2025) retain checkpoint-native backbones and trajectory formats. All systems use greedy decoding. Appendix A specifies the baseline implementations and evaluation protocol.

4.2 MAIN RESULTS AND GENERALIZATION

Table 1 reports every method on all seven benchmarks. ClaimWeaver reaches 50.5 EM and 56.4 F1, compared with 47.9/53.9 for EviNote-RAG. On the four multi-hop datasets, their averages are 45.7/52.6 and 41.2/48.9: gains of 4.5 EM and 3.7 F1, with EM improvements of 4.0–5.0 on each dataset. On single-hop QA, both obtain 56.9 EM; F1 is 61.4 versus 60.6. ClaimWeaver ties EviNote-RAG on NQ EM, gains 0.5 on TriviaQA, and loses 0.4 on PopQA. Its advantage is therefore concentrated on multi-hop QA.

TriplePPR-local is the strongest graph-inspired control at 45.8/51.7, 4.7 EM below ClaimWeaver overall. The local controls instantiate individual graph operators on a common candidate set; ClaimWeaver instead updates its evidence through adaptive retrieval. Paired bootstrap tests over matched, seed-averaged example scores against EviNote-RAG (10,000 resamples) give $p \leq 0.003$ on all four multi-hop datasets, meeting the seven-test Bonferroni threshold of 0.007. TriviaQA's unadjusted $p = 0.04$ does not meet that threshold (Appendix L); Appendix E reports seed variance.

Question complexity and model scale. The EM gain over EviNote-RAG increases from 3.8 on 2-hop to 5.2 on three-or-more-hop questions; an exploratory high-distractor subset gains 2.8 (Appendix F). At 3B scale, the seven-dataset scores are 41.2/46.6 versus 38.4/43.6, a 2.8 EM advantage (Appendix I).

Domain and task transfer. Without further training, ClaimWeaver gains 2.8 Strict Accuracy, 3.2 Lenient Accuracy, and 2.7 MRR on the 85 BioASQ 12b factoid questions (Nentidis et al., 2024), using provided PubMed evidence. On overlap-screened ASQA (Stelmakh et al., 2022), it gains 1.4

Method	Multi-hop QA								Single-hop QA						Avg	
	HotpotQA		2Wiki		MuSiQue		Bamboogle		NQ		TriviaQA		PopQA			
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Prompting-based																
Direct	21.6	24.4	22.8	26.4	7.4	8.5	21.6	22.1	13.4	19.8	29.8	32.1	15.0	17.0	18.8	21.5
Naive RAG	33.0	37.1	23.2	24.4	9.4	10.0	24.0	25.4	37.6	42.0	44.2	45.6	45.8	47.9	31.0	33.2
IRCoT-style	39.2	43.5	41.6	49.2	10.2	19.2	11.2	24.5	47.0	51.2	41.8	43.2	28.6	32.2	31.4	37.6
Self-RAG-style	40.2	44.8	39.0	43.2	18.2	19.2	25.6	35.1	46.4	50.8	43.6	45.1	31.4	33.2	34.9	38.8
Search-o1	26.8	33.0	27.2	28.6	14.0	16.8	32.8	35.8	31.0	34.5	56.6	58.9	33.6	36.9	31.7	34.9
Released-policy RL references																
R1-Searcher	45.0	49.8	44.6	49.1	21.4	22.8	17.6	20.1	49.2	53.8	68.8	73.1	41.2	42.7	41.1	44.5
Search-R1	42.0	46.4	24.4	28.0	18.4	27.4	32.0	37.7	50.8	55.0	69.4	75.4	**48.2**	**49.8**	40.7	45.7
ReARTeR	46.4	51.2	53.4	55.4	23.6	29.6	9.6	11.9	50.2	54.5	50.6	54.8	42.2	43.2	39.4	42.9
Graph-inspired local controls																
EntityGraph-local	46.2	52.0	45.8	50.5	24.2	33.4	39.2	48.8	45.8	50.2	59.6	64.8	40.4	43.0	43.0	49.0
ClusterTree-local	47.4	53.0	46.4	51.1	23.0	31.7	40.0	49.4	46.6	50.9	61.2	66.7	41.6	44.1	43.7	49.6
TriplePPR-local	48.6	54.2	50.0	54.4	24.6	34.0	41.6	51.2	48.4	52.6	63.8	69.2	43.8	46.4	45.8	51.7
EviNote-RAG†	49.0	55.7	49.4	53.6	24.0	33.6	42.4	52.8	**52.4**	56.3	70.3	76.5	48.0	49.1	47.9	53.9
CLAIMWEAVER	**54.0**	**59.4**	**54.0**	**58.0**	**28.2**	**37.0**	**46.4**	**56.0**	**52.4**	**56.8**	**70.8**	**77.8**	47.6	49.6	**50.5**	**56.4**

Table 1: Complete results on seven QA benchmarks. Avg. is the unweighted seven-dataset mean. Bold/underline mark the best/second-best scores. †EviNote-RAG shares the training configuration and serving limits, but uses its native free-form notes and reward. Released-policy methods are cross-system references; graph-inspired rows are local controls.

(a) Organization operations

Variant	Avg. EM	Avg. F1	MH EM
Full model	**50.5**	**56.4**	**45.7**
− Filtering	48.1	54.1	42.7
− Merging	49.1	55.4	44.3
− Chaining	47.2	52.9	40.9
− Merge − Chain	45.9	51.6	39.6
− All three	44.0	49.8	36.5

(b) Components and training

Variant	Avg. EM	Avg. F1
Full model	**50.5**	**56.4**
− Notes	47.0	52.6
− Organization	44.0	49.8
− Org − Notes	41.6	46.8
− $R_{\text{grounding}}$	48.6	54.2
Prompt-only (no RL)	25.2	29.8

Table 2: Ablations of organization operations (a) and broader components (b). Avg. covers all seven datasets; MH covers the four multi-hop datasets. Removing all three operations in (a) is the −Organization variant in (b). Per-dataset component results appear in Appendix D.

ROUGE-L, 1.8 Disambig-F1, and 1.5 DR with similar answer length. BioASQ evaluates organization over biomedical evidence, while NQ-derived ASQA evaluates adaptation to long-form answers. Both show positive transfer within these settings (Appendix T).

4.3 CONTRIBUTIONS OF ORGANIZATION AND LEARNING

Table 2 groups two complementary tests: retraining without individual organization operations, and removing broader components. Operation ablations retain decomposition and per-claim notes.

Removing chaining gives the largest individual loss: 3.3 overall and 4.8 multi-hop EM, versus multi-hop losses of 3.0 for filtering and 1.4 for merging. Removing all three loses 9.2 multi-hop EM. Organization without notes exceeds notes without organization by 3.0 overall EM, while removing both loses 8.9. Removing the grounding reward costs 1.8 EM. The prompt-only variant reaches 25.2 EM, highlighting the performance gap between the prompted base model and the trained policy under this interface.

Figure 3 shows that grounding and format rewards stabilize earlier than answer reward. Separately, on 100 fixed HotpotQA validation examples, relation F1 rises from 31.2 to 79.2 and triage precision

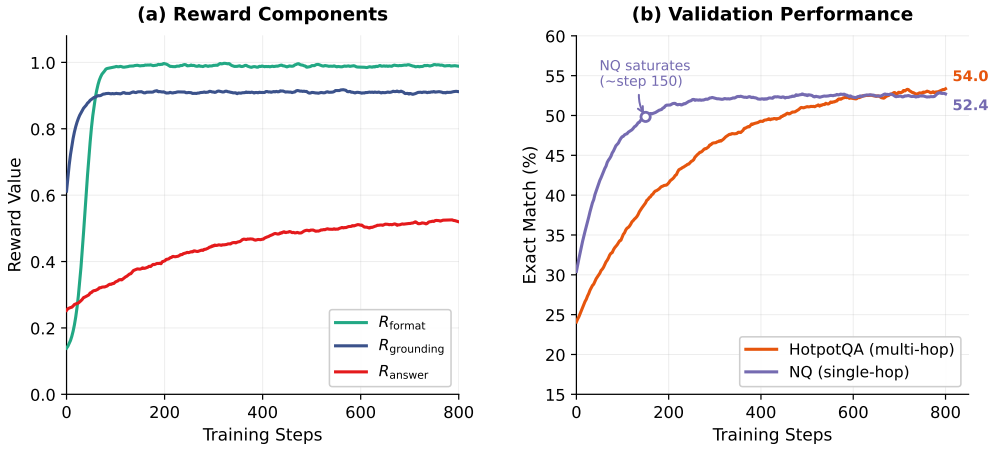


Figure 3: Training dynamics of CLAIMWEAVER. Grounding and format rewards stabilize earlier than answer reward (left). HotpotQA validation EM continues improving after NQ begins to plateau (right).

(a) Human relation audit

Relation	Prec.	Rec.	H-Mean
Overlap	92.0	84.6	88.1
Bridge	82.0	72.8	77.1

Precision: 50 predictions per type.
Recall: 320 candidate pairs from a separate set of queries.

(b) Fixed-evidence interventions

Organization state	EM	Δ EM
Unmodified	54.0	—
Shuffle relations	35.0	−19.0
Reverse bridges	39.0	−15.0
Restore excluded claims	45.0	−9.0
Disable merge	50.0	−4.0

Table 3: Complementary structure diagnostics. (a) H-Mean summarizes precision and recall from different samples. (b) On 100 HotpotQA validation trajectories, passages and search traces are fixed; notes and answers are regenerated after each intervention. Protocols appear in Appendices M and N.

from 52.4 to 86.4 as EM rises from 23.0 to 54.0 (Appendix N). Organization quality and answer accuracy thus improve together under the composite training objective.

Claim grounding and evidence coverage. We track extraction volume alongside grounding scores to characterize the learned claim-extraction behavior. On 200 fixed HotpotQA validation examples, the mean number of pre-triage claims grows from 4.2 at initialization to 7.6 at step 800, while grounding reward rises from 0.61 to 0.91. Counts at the last two checkpoints are 7.8 and 7.6. Higher grounding scores thus accompany a larger average claim set, with extraction volume stabilizing late in training (Appendix R). These counts describe extraction volume; matching claims to supporting facts is needed to measure coverage directly.

4.4 STRUCTURE QUALITY, INTERVENTIONS, AND FAILURES

Three annotators judge 200 decomposed claims per setting by majority vote (average pairwise $\kappa = 0.82$). The full model yields 188/200 faithful claims (94.0%), versus 163/200 (81.5%) without $R_{\text{grounding}}$. These labels assess whether each decomposed claim is entailed by its cited passage (Appendix Q). Table 3 separates relation correctness from downstream sensitivity to changes in the organization state.

Bridges are less reliable than overlap relations. Of the nine incorrect bridges in the precision sample, five reverse direction, three skip an intermediate claim, and one joins unrelated claims. Wrong answers co-occur in 4/5, 2/3, and 0/1 cases, respectively, identifying direction and intermediate-step errors as recurring patterns in this small audit (Appendix O).

Method	Tokens	Sec.	EM	F1
Naive RAG	500	1.2	33.0	37.1
Search-R1	2,800	6.5	42.0	46.4
EviNote-RAG	1,800	4.2	49.0	55.7
CLAIMWEAVER	3,200	7.8	54.0	59.4
EviNote-RAG (5-turn diagnostic)	3,100	—	50.2	56.4

Table 4: HotpotQA accuracy and cost under one serving configuration. Tokens cover the full trajectory, including retrieved passages. The five-turn diagnostic uses the original checkpoint with an extended retrieval horizon; the dash denotes unreported latency. Appendices G and P give cost details and the complete horizon comparison.

Under fixed passages and search traces, reversing bridges lowers EM by 15 points, restoring excluded claims by 9, and disabling merge by 4. Their ordering agrees with the retrained operation ablations: chaining has the largest effect, followed by filtering and merging. The two protocols capture complementary behavior: adaptation after removing an operation and downstream sensitivity when an already learned state is changed.

Failure localization. Among 100 inspected incorrect HotpotQA answers, the earliest failure is retrieval in 41%, organization in 28%, generation in 22%, and decomposition in 9%. Organization failures comprise useful claims being filtered out (12 cases), missed bridges (11), and wrong merges (5). Thus, 23 of the 28 organization failures involve losing useful evidence or failing to connect it, motivating conservative retention and more reliable bridge discovery (Appendix H).

4.5 ACCURACY–COST TRADE-OFF

On HotpotQA, ClaimWeaver uses 1,400 more trajectory tokens (+77.8%) and 3.6 more seconds (+85.7%) per query than EviNote-RAG for +5.0 EM/+3.7 F1 (Table 4). Giving EviNote-RAG five turns brings its token count to 3,100 and leaves a 3.8 EM/3.0 F1 gap. The five-turn setting extends inference for the original checkpoint, so the comparison measures the benefit of additional search under that trained policy.

Mean retrieval turns are 1.8 on single-hop and 2.1 on multi-hop QA; 12% of multi-hop questions reach the three-turn limit. The NLI verifier runs only during training and contributes no inference calls. Of the 3,200 HotpotQA trajectory tokens, approximately 600 are used for decomposition, 500 for organization, and 400 for notes. Together these stages account for about 47% of the trajectory, making intermediate evidence processing a substantial part of the measured cost (Appendix G).

5 LIMITATIONS

The evaluation focuses on English factoid QA; temporal, causal, and conflicting evidence may require richer relation types. Human grounding labels cover decomposed claims, leaving final-memory and path faithfulness for further evaluation. System comparisons combine organization, reward design, adaptive retrieval, and inference cost. Isolating these factors and comparing retrained directed and undirected policies require additional controlled experiments.

6 CONCLUSION

CLAIMWEAVER organizes retrieved evidence into query-conditioned claims and directed connections, with gains concentrated on multi-hop QA. Retrained ablations and fixed-evidence interventions identify chaining as the most consequential individual operation in the evaluated settings. Human audits characterize claim and relation quality, while BioASQ and ASQA extend the evaluation to biomedical evidence and long-form answers. Together, the results support persistent claim organization as an inspectable interface for learning how to use retrieved evidence in multi-hop reasoning.

Ethics Statement

This work studies evidence organization on existing QA benchmarks. Incorrect claims or connections can produce unsupported answers, even when intermediate sources are inspectable. Biomedical transfer is evaluated as benchmark QA. The human annotation protocol is described in Appendix S.

Reproducibility Statement

Section 3 specifies the organization interface and reward. Appendices A–C describe the retrieval setup, trajectory schemas, and training hyperparameters. Appendices L, M, and S describe statistical tests and human evaluation.

AI Use Statement

Generative language models are components of the proposed method and are used to generate search queries, claims, organization decisions, notes, and answers as described in Section 3. During manuscript preparation, OpenAI Codex assisted with language editing, LaTeX formatting, figure refinement, literature checking, arithmetic consistency checks, and discussion of experimental analyses. The authors checked AI-assisted numerical corrections against their experimental results. The authors are responsible for the final text, scientific claims, references, and research artifacts.

References

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

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://aclanthology.org/2024.emnlp-main.845/.

Yuqin Dai, Guoqing Wang, Yuan Wang, Kairan Dou, Kaichen Zhou, Zhanwei Zhang, Shuo Yang, Fei Tang, Jun Yin, Pengyu Zeng, Zhenzhe Ying, Can Yi, Changhua Meng, Yuchen Zhou, Yongliang Shen, and Shuai Lu. EviNote-RAG: Enhancing RAG models via answer-supportive evidence notes. *arXiv preprint arXiv:2509.00877*, 2025. URL https://arxiv.org/abs/2509.00877.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. URL https://arxiv.org/abs/2404.16130.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. 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)*, pp. 18969–18985, 2025. doi: 10.18653/v1/2025.acl-long.929. URL https://aclanthology.org/2025.acl-long.929/.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. URL https://arxiv.org/abs/2312.10997.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract.html.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

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*, 2020. URL https://aclanthology.org/2020.coling-main.580/.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.91/.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.495/.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025. URL https://arxiv.org/abs/2503.09516.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017. URL https://aclanthology.org/P17-1147/.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world entailment for claims in wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.470/.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. URL https://aclanthology.org/Q19-1026/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230-Abstract.html.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5420–5438, 2025a. URL https://aclanthology.org/2025.emnlp-main.276/.

Yuankai Li, Jia-Chen Gu, Di Wu, Kai-Wei Chang, and Nanyun Peng. BRIEF: Bridging retrieval and inference for multi-hop reasoning via compression. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5464–5485, 2025b. doi: 10.18653/v1/2025.findings-naacl.301. URL https://aclanthology.org/2025.findings-naacl.301/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://aclanthology.org/2023.acl-long.546/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.741/.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, and Georgios Paliouras. Overview of BioASQ tasks 12b and synergy12 in CLEF 2024. In *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, 2024. URL https://ceur-ws.org/Vol-3740/paper-01.pdf.

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/.

Dragomir Radev. A common theory of information fusion from multiple text sources step one: Cross-document structure. In *1st SIGdial Workshop on Discourse and Dialogue*, pp. 74–83, 2000. URL https://aclanthology.org/W00-1009/.

Dragomir Radev, Jahna Otterbacher, and Zhu Zhang. CST bank: A corpus for the study of cross-document structural relationships. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004. URL https://aclanthology.org/L04-1239/.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. URL https://aclanthology.org/2023.tacl-1.75/.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/pdf?id=GN921JHCRw.

Sina Semnani, Jirayu Burapacheep, Arpandeep Khatua, Thanawan Atchariyachanvanit, Zheng Wang, and Monica Lam. Detecting corpus-level knowledge inconsistencies in Wikipedia with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34839–34866, 2025. URL https://aclanthology.org/2025.emnlp-main.1765/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024. URL https://arxiv.org/abs/2409.19256.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-Searcher: Incentivizing the search capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025. URL https://arxiv.org/abs/2503.05592.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8273–8288. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.566. URL https://aclanthology.org/2022.emnlp-main.566/.

Zhongxiang Sun, Qipeng Wang, Weijie Yu, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Yang Song, and Han Li. ReARTeR: Retrieval-augmented reasoning with trustworthy process rewarding. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1251–1261, 2025. URL https://arxiv.org/abs/2501.07861.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://aclanthology.org/2023.acl-long.557/.

Jingjin Wang and Jiawei Han. PropRAG: Guiding retrieval with beam search over proposition paths. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6212–6227, 2025. URL https://aclanthology.org/2025.emnlp-main.317/.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. URL https://arxiv.org/abs/2212.03533.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/937ae0e83eb08d2cb8627fe1def8c751-Abstract-Conference.html.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*, 2023. URL https://arxiv.org/abs/2310.04408.

Zhichao Xu, Zongyu Wu, Yun Zhou, Aosong Feng, Kang Zhou, Sangmin Woo, Kiran Ramnath, Yijun Tian, Xuan Qi, Weikang Qiu, Lin Lee Cheong, and Haibo Ding. Beyond correctness: Rewarding faithful reasoning in retrieval-augmented generation. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=mZ0gGlXelF.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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/.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/pdf?id=Ez6fkqltwV.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://aclanthology.org/2024.emnlp-main.813/.

Zhu Zhang, Sasha Blair-Goldensohn, and Dragomir R. Radev. Towards CST-enhanced summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2002. URL https://cdn.aaai.org/AAAI/2002/AAAI02-067.pdf.

A Detailed Experimental Setup

Datasets. We evaluate on seven QA benchmarks spanning different reasoning requirements. **Multi-hop QA**: HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023) require composing multiple evidence pieces across documents. **Single-hop QA**: Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017) measure behavior on direct lookup questions. PopQA (Mallen et al., 2023) further tests robustness on rare entities where parametric memory is less reliable. Following prior work (Jin et al., 2025; Dai et al., 2025), we train on the combined training sets of NQ and HotpotQA, while the remaining five datasets serve as held-out cross-dataset evaluations. We randomly sample 500 examples from each test or validation set for evaluation using a fixed random seed (seed = 42); for Bamboogle, we use the full validation set of 125 examples. This sample size follows the evaluation protocol of EviNote-RAG (Dai et al., 2025). We assess significance with paired bootstrap resampling over evaluation examples.

Baselines and Evaluation Protocol. We compare against methods spanning multiple RAG paradigms:

- **No retrieval**: Direct prompting (closed-book).
- **Standard RAG**: Naive RAG (top-k concatenated).
- **Adaptive RAG**: Self-RAG-style (Asai et al., 2024), IRCoT-style (Trivedi et al., 2023).
- **Prompt-based agentic**: Search-o1 (Li et al., 2025a).
- **Configuration-matched free-form-note control**: EviNote-RAG (Dai et al., 2025).
- **Released-policy RL references**: Search-R1 (Jin et al., 2025), R1-Searcher (Song et al., 2025), and ReARTeR (Sun et al., 2025).

All controlled systems use Qwen2.5-7B-Instruct. ClaimWeaver, EviNote-RAG, Naive RAG, Self-RAG-style, IRCoT-style, and Search-o1 use E5-base-v2 (Wang et al., 2022) over a shared passage corpus derived from the English Wikipedia snapshot `20231101.en` in `wikimedia/wikipedia`.[1] We retrieve the top-3 passages per query. For Self-RAG-style and IRCoT-style, we follow the released prompting strategies; Search-o1 receives the same 3-turn budget as ClaimWeaver. We implement EviNote-RAG as a configuration-matched free-form-note control. It uses the same base model, E5 index and top-k retrieval interface, NQ/HotpotQA training examples, GRPO schedule, maximum retrieval turns, context/output budgets, decoding, and

[1] https://huggingface.co/datasets/wikimedia/wikipedia

three training seeds as CLAIMWEAVER; its memory follows the original Supportive-Evidence Note representation and its training objective follows the original EviNote-RAG reward design. CLAIMWEAVER uses the answer, claim-grounding, and format objective described in the main paper. This comparison aligns training configuration and serving limits across the two complete methods; realized computation and reward objectives differ. Search-R1, R1-Searcher, and ReARTeR are evaluated from their released checkpoints and retain their checkpoint-native policy backbones and trajectory formats. Where their released interfaces support an external retriever, we use the same E5 corpus and evaluation pipeline. These three rows retain differences in architecture and pretraining and serve as cross-system reference comparisons. All systems use greedy decoding at inference time. CLAIMWEAVER and matched EviNote-RAG are trained with the same three independent random seeds, and their reported results are averaged over those seeds.

Graph-inspired local controls. On all seven datasets, each local control receives the same top-50 E5 candidates. EntityGraph-local links passages sharing extracted entities; ClusterTree-local clusters passage embeddings into a local summary tree; and TriplePPR-local links extracted triples and applies query-seeded personalized PageRank. Selected evidence is truncated to the common context budget and passed to the same Qwen2.5-7B-Instruct answer prompt with greedy decoding. The three controls instantiate query-local operators inspired by GraphRAG, RAPTOR, and HippoRAG, respectively.

Implementation details. In our implementation, we retrieve the top-3 passages per query. The NLI model for $R_{\text{grounding}}$ is DeBERTa-v3-large (He et al., 2023) fine-tuned on MNLI. R_{format} combines regex checks for expected structural markers with deterministic validation of references and relation structure. We set $\lambda_1 = 1.0$, $\lambda_2 = 0.3$, $\lambda_3 = 0.1$, and sample $G = 8$ rollouts per query during GRPO training. Training is conducted on 8 NVIDIA A100 GPUs using the verl framework (Sheng et al., 2024) for 800 steps with a learning rate of 1×10^{-6}, KL penalty coefficient of 0.001, and maximum sequence length of 4096 tokens.

Metrics. We report Exact Match (EM) and token-level F1 as answer quality metrics, following prior work (Jin et al., 2025; Dai et al., 2025). Both use the same answer normalization and alias handling for every system; EM requires an exact normalized alias match, while token F1 is the harmonic mean of normalized token precision and recall, maximized over gold aliases. In addition, we conduct human evaluations of intermediate-output quality, including claim faithfulness and relational-state annotation accuracy.

B TRAJECTORY FORMAT AND OUTPUT SCHEMAS

This section documents the system prompt and output schemas that define CLAIMWEAVER's trajectory structure. During inference, the model receives the system prompt. The stage-specific blocks below specify output schemas encouraged by R_{format} during training.

B.1 SYSTEM PROMPT (AGENTIC LOOP)

The system prompt defines the overall iterative retrieve-organize-reason behavior:

```
You are a question-answering agent with access to a search
tool.  To answer the user's question, you will iteratively
search for evidence and organize it before producing a final
answer.

At each step, you may either:
(1) Issue a search query to gather more evidence, OR
(2) Produce a final answer if sufficient evidence has been
collected.

When you retrieve documents, you MUST process them through
the following three-stage evidence organization pipeline
before deciding your next action:

Stage 1:  DECOMPOSE - Extract atomic claims from passages
Stage 2:  ORGANIZE - Filter, merge, chain, and retain claims
```

```
Stage 3: NOTE - Annotate each claim with its role
After organization, review your accumulated notes. If
critical information is missing, formulate a targeted
follow-up search query. Otherwise, synthesize the notes
into a final answer.
Format your output using the structural tags below.
```

Figure 2 in the main text provides a worked example.

B.2 Stage 1: Claim Decomposition

```
<decompose>
Given the query and retrieved passages below, decompose each
passage into atomic claims.
Requirements for each claim:
- ATOMIC: Express exactly one factual proposition
- SELF-CONTAINED: Disambiguate all pronouns, abbreviations,
and implicit references so the claim is interpretable
without surrounding context
- SOURCE-GROUNDED: Cite the source document
- VERIFIABLE: Must be independently checkable as true/false
Do NOT include opinions, rhetorical statements, or
meta-information about the document itself.
Query: {query}
Passages:
[Doc1] {passage_1}
[Doc2] {passage_2}
[Doc3] {passage_3}
Output format:
[c1] <claim text> [Doc1]
[c2] <claim text> [Doc1]
[c3] <claim text> [Doc2]
...
</decompose>
```

Example. Given the passage *"Toni Morrison was awarded the Nobel Prize in Literature in 1993, becoming the first African-American woman to receive it."* from Doc3, with the query *"What university did the 1993 Nobel Literature laureate teach at?"*, the model produces:

> [c1] Toni Morrison was awarded the Nobel Prize in Literature in 1993. [Doc3]
> [c2] Toni Morrison was the first African-American woman to receive the Nobel Prize in Literature. [Doc3]

The reference "it" is expanded to "the Nobel Prize in Literature," and the sentence is split into two independently interpretable claims.

B.3 Stage 2: Claim Organization

```
<organize>
Given the query, the newly decomposed claims, and any
existing notes from previous rounds, organize all evidence
for answering.
If existing notes are provided, treat them as additional
claims when discovering relations. A new claim may OVERLAP
with an existing note (merge to strengthen provenance) or
BRIDGE with an existing note (extend a reasoning chain
across rounds).
Step A - Relevance Triage:
Identify claims that are NON-CONTRIBUTORY to answering the
query. A claim is non-contributory only if it has NEITHER:
```

```
(a) direct utility (contains information that directly
answers the query), NOR
(b) indirect utility (serves as an intermediate step in a
reasoning chain toward the answer).
Be conservative: when uncertain, RETAIN the claim.

Step B - Relation Discovery (among retained claims only):
Examine retained claim pairs and output every predicted
OVERLAPPING or BRIDGING relation. Unlisted pairs are
assigned NONE.

- OVERLAPPING: Both claims provide the same or redundant
information for answering the query.
- BRIDGING: One claim introduces an entity or fact needed to
interpret the other, forming a directed logical dependency
for multi-hop reasoning. Specify the direction: ci -> cj
means ci provides a premise for cj.
- NONE: The pair has neither relation. After all pairs are
processed, a retained claim with no OVERLAPPING or BRIDGING
edge is marked ISOLATED.

Step C - Structural Resolution:
Apply resolutions in this order:
- EXCLUDE: Remove non-contributory claims
- MERGE: Form overlap-connected components, fuse each
component into one representative claim, and combine
provenance
- CHAIN: Project bridge edges onto the merged claims and
link them with directed edges, specifying the shared entity
or logical link; reject self-edges and directed cycles
- RETAIN: Keep isolated claims unchanged

Query: {query}
New claims: {claims from Stage 1}
Existing notes (from previous rounds, if any): {accumulated
notes or "None"}

Output format:
[Excluded] c2, c5 (reason: <brief justification>)
[Merge] c1 + c4 -> C1: "<merged claim text>" [Doc1, Doc3]
[Chain] C1 -> c3 (via <bridge entity>)
[Isolated] c6
</organize>
```

B.4 Stage 3: Per-Claim Note Generation

```
<note>
For each retained/merged/chained claim from the organization
stage, generate a usage note. Each note must specify:

(1) ROLE: One of the following:
  - Entry point: Starting claim of a reasoning chain
  - Bridge: Intermediate step connecting other claims
  - Answer-bearing: Directly contains the answer
  - Supporting: Provides confirming evidence
  - Independent: Standalone relevant fact

(2) SOURCE COUNT: Number of distinct source documents
supporting this claim; multiple passages from the same
document count once

(3) CONNECTIONS: What this claim follows from or leads to
(if part of a chain)

(4) USAGE: How the downstream generator should use this
claim when formulating the answer

Query: {query}
Organized claims: {output from Stage 2}

Output format:
```

```
C1: "<claim text>" [Doc1, Doc3]
  Note: Role=Entry point. Sources=2. Leads to c3 via
<entity>. Use: identifies the target entity for the next
reasoning step.
c3: "<claim text>" [Doc2]
  Note: Role=Answer-bearing. Sources=1. Follows from C1.
Use: provides the final answer directly.
c6: "<claim text>" [Doc1]
  Note: Role=Independent. Sources=1. Use: optional
supporting context outside the primary reasoning path.
</note>
```

B.5 Search Decision and Answer Generation

After generating notes, the model decides whether to search again or answer:

```
<decision>
Review the accumulated notes. Determine whether sufficient
evidence exists to answer the query.
If a reasoning chain is INCOMPLETE (e.g., a bridging claim
has no follow-up, or the answer-bearing claim is missing),
formulate a targeted search query to fill the gap:
<search> {new query targeting the missing information}
</search>
If the evidence is SUFFICIENT, produce the final answer:
<answer> {concise answer} </answer>
</decision>
```

B.6 Two-Round Trajectory Illustration

The following schematic trajectory illustrates the memory update for the query *"Where was the director of Parasite born?"*:

```
<search> director of Parasite film </search>
[Retrieved: Doc1, Doc2, Doc3]
<decompose>
[c1] Bong Joon-ho directed the 2019 film Parasite. [Doc1]
[c2] Parasite won the Palme d'Or at Cannes 2019. [Doc1]
[c3] Bong Joon-ho directed the film Parasite. [Doc2]
[c4] Parasite grossed over $250 million worldwide. [Doc3]
</decompose>
<organize>
[Excluded] c2 (award info, not about birthplace), c4 (box
office, irrelevant)
[Merge] c1 + c3 -> C1: "Bong Joon-ho directed the 2019 film
Parasite." [Doc1, Doc2]
[Isolated] C1
</organize>
<note>
C1: "Bong Joon-ho directed the 2019 film Parasite." [Doc1,
Doc2]
  Note: Role=Entry point. Sources=2. Use: identifies the
director; need birthplace next.
</note>
<search> Bong Joon-ho birthplace born </search>
[Retrieved: Doc4, Doc5, Doc6]
<decompose>
[c5] Bong Joon-ho was born in Daegu, South Korea. [Doc4]
[c6] Bong Joon-ho was born in Daegu, South Korea. [Doc5]
[c7] Bong Joon-ho studied sociology at Yonsei University.
[Doc5]
```

Hyperparameter	Value
Base model	Qwen2.5-7B-Instruct
Optimizer	AdamW
Learning rate	1×10^{-6}
LR schedule	Cosine with warmup
Warmup steps	30
Training steps	800
Batch size	128 queries
Rollouts per query (G)	8
Max sequence length	4096 tokens
Max retrieval turns	3
KL penalty coefficient	0.001
λ_1 (R_{answer})	1.0
λ_2 ($R_{\text{grounding}}$)	0.3
λ_3 (R_{format})	0.1
Temperature (sampling)	0.7
Top-p (sampling)	0.95
Retriever	E5-base-v2
Passages per query	3
NLI model ($R_{\text{grounding}}$)	DeBERTa-v3-large-MNLI
Hardware	8× NVIDIA A100 80GB
Framework	verl
Training time	∼30 hours

Table 5: Training hyperparameters for CLAIMWEAVER.

```
</decompose>
<organize>
[Excluded] c7 (education, not about birthplace)
[Merge] c5 + c6 -> C2: "Bong Joon-ho was born in Daegu,
South Korea." [Doc4, Doc5]
[Chain] C1 -> C2 (via Bong Joon-ho)
</organize>
<note>
C1: "Bong Joon-ho directed the 2019 film Parasite." [Doc1,
Doc2]
 Note: Role=Entry point. Sources=2. Leads to C2. Use:
establishes director identity.
C2: "Bong Joon-ho was born in Daegu, South Korea." [Doc4,
Doc5]
 Note: Role=Answer-bearing. Sources=2. Follows from C1.
Use: provides birthplace directly.
</note>
<answer> Daegu, South Korea </answer>
```

C TRAINING DETAILS

Table 5 lists the training hyperparameters. The policy receives reward on complete trajectories; retrieved passage tokens are environment observations and are excluded from policy-gradient updates. For query q, the relative advantage of sampled trajectory τ_i is

$$\hat{A}_i = \frac{R(\tau_i) - \text{mean}_{j=1}^{G} R(\tau_j)}{\text{std}_{j=1}^{G} R(\tau_j) + \epsilon}, \tag{5}$$

where $G = 8$ and ϵ stabilizes groups with zero or near-zero reward variance. These advantages weight policy-generated tokens, with a KL penalty against the frozen reference policy.

Training data. We train on the combined training sets of NQ and HotpotQA, totaling approximately 150K queries. During each training step, we sample a batch of 128 queries and generate

$G = 8$ rollout trajectories per query. Each trajectory includes the full retrieve-organize-reason loop with up to 3 retrieval turns.

Reward implementation. R_{answer} is computed via exact match between the predicted and gold answers. $R_{\text{grounding}}$ averages DeBERTa-v3-large-MNLI entailment probabilities over the decomposed claims in a nonempty claim trajectory. R_{format} checks via regex whether the trajectory contains the expected structural markers for both the global agentic actions and the internal CLAIMWEAVER operations. Specifically, it verifies the presence of search query tags, claim decomposition tags, relational state predictions, per-claim notes, and a final answer block. $R_{\text{format}} = 1$ if all required markers are present and all structural validation checks pass, and 0 otherwise.

Inference. At inference time, we use greedy decoding with temperature 0. The retrieval setup remains identical to training: E5 retriever with top-3 passages and up to 3 retrieval turns.

Iterative evidence accumulation. Across multiple retrieval turns, CLAIMWEAVER maintains a cumulative note buffer. After each retrieval round, newly decomposed claims are organized jointly with the existing notes from previous rounds. This allows cross-round relation discovery: a claim retrieved in round 2 may be identified as bridging with a claim from round 1. The model decides to issue a follow-up search when its notes indicate an incomplete reasoning chain (e.g., an entry-point claim with no answer-bearing follow-up). The maximum of 3 retrieval turns bounds context growth; because each turn can return multiple passages, retrieval turns and reasoning hops need not coincide.

Format reward details. R_{format} uses regex checks for the following structural markers: (1) at least one `<search>` tag; (2) a `<decompose>` block with numbered claims and source citations; (3) an `<organize>` block with at least one operation (Excluded/Merge/Chain/Isolated); (4) a `<note>` block with Role annotations; and (5) an `<answer>` tag. A deterministic structural validator separately rejects malformed references, merge self-links, and directed chain cycles. The reward is binary: $R_{\text{format}} = 1$ if all checks pass, and 0 otherwise. This component supervises trajectory syntax; the relation audit below evaluates relation semantics.

D PER-DATASET COMPONENT ABLATIONS

Table 6 reports complementary component-level ablations across datasets. Table 2 in the main paper summarizes both these components and the fine-grained filtering, merging, and chaining ablations.

Setting	HotpotQA		2Wiki		MuSiQue		Bamboogle	
	EM	F1	EM	F1	EM	F1	EM	F1
CLAIMWEAVER (Full)	54.0	59.4	54.0	58.0	28.2	37.0	46.4	56.0
− Notes	48.8	54.6	48.4	52.8	24.2	32.2	41.6	50.0
− Organization	45.0	51.2	43.2	48.6	20.4	26.8	37.3	45.6
− Org − Notes	42.4	48.8	39.4	45.2	17.6	22.4	33.9	38.6
− $R_{\text{grounding}}$	51.2	57.0	52.0	55.8	26.4	34.6	44.3	53.4
Prompt-only (no RL)	24.8	28.6	23.4	27.2	8.2	14.4	22.4	29.0

Setting	NQ		TriviaQA		PopQA		Avg.	
	EM	F1	EM	F1	EM	F1	EM	F1
CLAIMWEAVER (Full)	52.4	56.8	70.8	77.8	47.6	49.6	50.5	56.4
− Notes	50.4	54.8	68.4	74.6	47.2	49.2	47.0	52.6
− Organization	48.8	55.0	66.2	72.4	47.0	49.0	44.0	49.8
− Org − Notes	47.4	54.4	64.4	69.8	46.2	48.4	41.6	46.8
− $R_{\text{grounding}}$	51.4	55.6	69.4	76.0	45.8	47.2	48.6	54.2
Prompt-only (no RL)	30.6	35.4	38.2	41.8	28.8	32.2	25.2	29.8

Table 6: Per-dataset component ablations. Avg. is the seven-dataset macro-average. Averages and component-level differences are calculated before rounding.

E VARIANCE ACROSS TRAINING SEEDS

Table 7 reports the standard deviation of CLAIMWEAVER's EM scores across three training seeds for each dataset. The observed standard deviations range from 0.2 to 0.8 EM. On Bamboogle, the three runs obtain 45.6, 46.4, and 47.2 EM, yielding a sample standard deviation of 0.8.

Dataset	Mean EM	Std
HotpotQA	54.0	0.4
2WikiMultiHopQA	54.0	0.6
MuSiQue	28.2	0.5
Bamboogle	46.4	0.8
NQ	52.4	0.3
TriviaQA	70.8	0.2
PopQA	47.6	0.4

Table 7: Standard deviation of CLAIMWEAVER EM across three training seeds.

F PERFORMANCE BY QUESTION TYPE

We analyze performance by reasoning depth, distinguishing questions that require two hops from those requiring three or more hops. Table 8 reports gains over EviNote-RAG for these groups, alongside direct single-hop questions and a high-distractor subset. High-distractor questions are an exploratory, orthogonal subset. Under the fixed top-3 retrieval output, we assign this label when at least two passages contain no gold supporting fact. We construct the subset only for datasets with supporting-fact annotations; membership may overlap the hop-count bins.

Question Type	Δ EM	Δ F1
Single-hop direct	+0.0	+0.8
Multi-hop 2-hop	+3.8	+4.6
Multi-hop (3+ hops)	+5.2	+6.4
High distractor	+2.8	+3.4

Table 8: Performance improvement of CLAIMWEAVER over EviNote-RAG, broken down by question type. Single-hop values are macro-averaged over NQ, TriviaQA, and PopQA.

The EM improvement is +3.8 on two-hop questions and +5.2 on questions requiring three or more hops. This trend is consistent with the intended role of bridge chains in connecting longer evidence paths. The high-distractor subset gains +2.8 EM, aligning with the role of relevance triage in limiting distracting claims.

G EFFICIENCY ANALYSIS

Table 9 compares the computational cost and accuracy of each method on HotpotQA. CLAIMWEAVER uses approximately 3,200 total trajectory tokens per query (including search queries, retrieved passages, intermediate structures, and generated output), resulting in higher latency than simpler baselines. The accuracy gains over EviNote-RAG (+5.0 EM) and Naive RAG (+21.0 EM) therefore come with a material deployment cost. The later diagnostic brings trajectory-token counts closer by extending the baseline retrieval horizon without retraining. Compute and latency remain descriptive measurements under the recorded serving configuration.

Method	Traj. Tokens	Latency	EM	F1
Naive RAG	500	1.2s	33.0	37.1
Search-R1	2,800	6.5s	42.0	46.4
EviNote-RAG	1,800	4.2s	49.0	55.7
CLAIMWEAVER	3,200	7.8s	54.0	59.4

Table 9: Descriptive efficiency comparison on HotpotQA under one serving configuration.

Per-stage token breakdown. On HotpotQA, CLAIMWEAVER's average 3,200 trajectory tokens per query are distributed as follows: first-round search queries and retrieved passages account for $\sim$900 tokens (28%), claim decomposition for $\sim$600 tokens (19%), organization (triage + relation + resolution) for $\sim$500 tokens (16%), notes for $\sim$400 tokens (12%), and the answer decision for $\sim$100 tokens (3%). The remaining $\sim$700 tokens (22%) arise from second- and third-round retrieval in multi-hop questions. The NLI model for $R_{\text{grounding}}$ is used only during training to compute rewards; inference generates the full trajectory without NLI calls. The mean number of retrieval turns is 1.8 on single-hop datasets and 2.1 on multi-hop datasets; only 12% of multi-hop queries use all 3 turns.

H REMAINING ERROR ATTRIBUTION

To identify bottlenecks for future improvement, we manually inspect 100 incorrect predictions from CLAIMWEAVER on HotpotQA and trace each to the earliest pipeline stage where the error originates. Retrieval is assigned when a required supporting fact is absent from the retrieved passages; decomposition when present evidence is omitted or distorted in the extracted claims; organization when the required claims are available but filtered, merged, or chained incorrectly; and answer generation when sufficient correctly organized notes are available but the final answer is wrong. Recording the earliest applicable category makes the four labels mutually exclusive. Table 10 reports the results.

Source Stage	%
Retrieval (missing from retrieved passages)	41
Claim organization	28
– Incorrect filtering (useful claim excluded)	12
– Missed bridging (chain not constructed)	11
– Wrong merge (non-equivalent claims fused)	5
Answer generation (wrong answer from correct notes)	22
Claim decomposition (omitted or distorted)	9

Table 10: Attribution of remaining errors by pipeline stage on HotpotQA (100 manually inspected incorrect predictions).

Retrieval accounts for 41% of errors and is the largest observed bottleneck. Within the organization stage, incorrect filtering (12%) and missed bridging (11%) are the dominant failure modes, while wrong merges are relatively rare (5%). The filtering errors typically involve claims with indirect utility (e.g., intermediate steps in a reasoning chain) being mistakenly excluded, while missed bridging errors arise when the model fails to identify the shared entity connecting two claims across documents. Claim decomposition accounts for 9% of the inspected answer failures. This stage-attribution rate has a different denominator from the claim-faithfulness audit and includes both omitted and distorted evidence. These results suggest that pairing CLAIMWEAVER with stronger retrievers and improving conservative bridging detection are the most promising directions for future gains.

I MODEL SIZE SCALING

To evaluate generalizability across model scales, we train and evaluate CLAIMWEAVER with Qwen2.5-3B-Instruct in addition to the default 7B model. The matched EviNote-RAG control uses

the same scale-specific training and serving configuration at both model sizes. Table 11 reports the results.

Base Model	Method	Avg EM	Avg F1
Qwen2.5-3B	Naive RAG	24.6	27.4
	EviNote-RAG	38.4	43.6
	CLAIMWEAVER	41.2	46.6
Qwen2.5-7B	Naive RAG	31.0	33.2
	EviNote-RAG	47.9	53.9
	CLAIMWEAVER	50.5	56.4

Table 11: Scaling analysis. CLAIMWEAVER consistently outperforms baselines at both 3B and 7B scales. The 7B results correspond to those reported in the main paper.

CLAIMWEAVER outperforms both baselines at both scales, improving over EviNote-RAG by 2.8 EM at 3B and 2.6 EM at 7B, yielding similar margins across the two model capacities.

J TRAINING DYNAMICS

We observe distinct learning phases during GRPO training (Figure 3). The grounding reward ($R_{\text{grounding}}$) stabilizes within the first 50 steps. The format reward (R_{format}) reaches near-perfect compliance by step 80. The answer reward (R_{answer}) improves more gradually as the complete retrieval and organization policy is optimized.

Validation EM on HotpotQA shows steady improvement throughout training, while NQ saturates earlier (around step 150). The different trajectories align with the larger final gains observed on multi-hop QA.

K QUALITATIVE CASE STUDIES

We provide two qualitative case studies comparing CLAIMWEAVER with passage-concatenation and free-form-note baselines. The first case study illustrates multi-hop reasoning with a same-nationality distractor and shows how filtering and bridging can separate the evidence paths. The second case study focuses on a single-hop question with redundant evidence, where overlapping merge and isolated claim retention are most relevant.

K.1 MAIN CASE STUDY: MULTI-HOP REASONING WITH MISLEADING DISTRACTORS

For the query *"In what city was the director of the film that won the 2020 Academy Award for Best Picture born?"* (Figure 4), retrieval returns passages about *Parasite*, Bong Joon-ho, and Park Chan-wook. The correct reasoning path identifies *Parasite* as the winner, links it to Bong Joon-ho, and then extracts his birthplace. Retrieved passages also mention Bong's relocation and education in Seoul and Park Chan-wook, another South Korean director born in Seoul.

- **Naive RAG** overweights repeated mentions of Seoul and answers "Seoul" despite a passage stating that Bong was born in Daegu.
- **EviNote-RAG** identifies the director but drops the Daegu birthplace fact during free-form compression and also answers Seoul.
- **CLAIMWEAVER** separates birthplace evidence from biographical distractors, excludes claims about the wrong predicate or director, and chains the Best Picture, director, and birthplace claims to answer "Daegu, South Korea."

Question: In what city was the director of the film that won the 2020 Academy Award for Best Picture born?

Naive RAG

<information>
Doc 1 (Title: "Parasite (2019 film)"): Parasite is a 2019 South Korean black comedy thriller film directed by Bong Joon-ho. The film won the Palme d'Or at the 2019 Cannes Film Festival and the Academy Award for Best Picture at the 92nd Academy Awards in February 2020, becoming the first non-English-language film to win Best Picture ...
Doc 2 (Title: "Bong Joon-ho"): Bong Joon-ho (born September 14, 1969) is a South Korean filmmaker, born in **Daegu, South Korea**. While Bong was in elementary school, the family relocated to Seoul ... He enrolled in Yonsei University in Seoul ... and is best known for Memories of Murder (2003), The Host (2006), Snowpiercer (2013), and Parasite (2019) ...
Doc 3 (Title: "Park Chan-wook"): Park Chan-wook (born 23 August 1963) is a South Korean filmmaker and former film critic, born in Seoul, South Korea. He is widely regarded as a leading figure in 21st-century world cinema, known for Joint Security Area (2000), Oldboy (2003), The Handmaiden (2016), and Decision to Leave (2022) ...
</information>
<think>
The question asks for the birth city of the director of the film that won Best Picture in 2020. The retrieved passages say that Parasite won the Academy Award for Best Picture, and that Parasite was directed by Bong Joon-ho. The passages about Korean directors mention both Daegu and Seoul, but Seoul appears repeatedly in the retrieved context and is strongly associated with the director's background and Korean film context. Therefore, the director was likely born in Seoul.
</think>
<answer>
Seoul
</answer>

Frequency bias → wrong answer

EviNote-RAG

<information>
[same three docs as Naive RAG]
</information>
<summary>
The 2020 Academy Award for Best Picture was won by *Parasite*, directed by Bong Joon-ho. *Bong Joon-ho is a South Korean filmmaker primarily associated with Seoul*, where he was raised after his family relocated during his childhood, where he later attended Yonsei University, and where he established his filmmaking career. Doc 2 also briefly notes Daegu in connection with Bong's early life, but this detail appears peripheral to his lifelong identification with Seoul.
</summary>
<answer>
Seoul
</answer>

Wrong predicate marked as key

ClaimWeaver

<information>
[same three docs as Naive RAG]
</information>
<decompose>
c1: Parasite won the 2020 Academy Award for Best Picture. [Doc 1]
c2: Parasite was directed by Bong Joon-ho. [Doc 1]
c3: Bong Joon-ho was born in Daegu, South Korea. [Doc 2]
c4: Bong Joon-ho's family relocated to Seoul during his elementary school years. [Doc 2]
c5: Bong Joon-ho attended Yonsei University in Seoul. [Doc 2]
c6: Park Chan-wook is a South Korean filmmaker. [Doc 3]
c7: Park Chan-wook was born in Seoul, South Korea. [Doc 3]
</decompose>

Claim decomposition

<organize>
[Triage]
c1, c2, c3 — on the target chain ✓
c4, c5 — about Bong's relocation/education, predicate ≠ born_in ✗
c6, c7 — about Park Chan-wook, ≠ director of target film ✗
[Exclude] c4, c5 — predicate mismatch ("relocated_to" / "attended" are not "born_in").
[Exclude] c6, c7 — bridging entity "Park Chan-wook" fails the target relation (not director of Parasite).
[Chain] c1 → c2 → c3 via entities "Parasite" → "Bong Joon-ho".
</organize>

Filter non-contributory claims

Construct bridging chain

<note>
c1 [Entry point]: Identifies the 2020 Best Picture film.
c2 [Bridge]: Identifies the director of that film.
c3 [Answer-bearing]: Provides the birth city of the target director.
</note>

Per-claim note generation

<answer>
Daegu, South Korea
</answer>

Correct answer

Figure 4: Case study on multi-hop reasoning with misleading distractors. CLAIMWEAVER filters distractor claims and constructs the correct reasoning chain; the baselines follow spurious associations in the retrieved evidence.

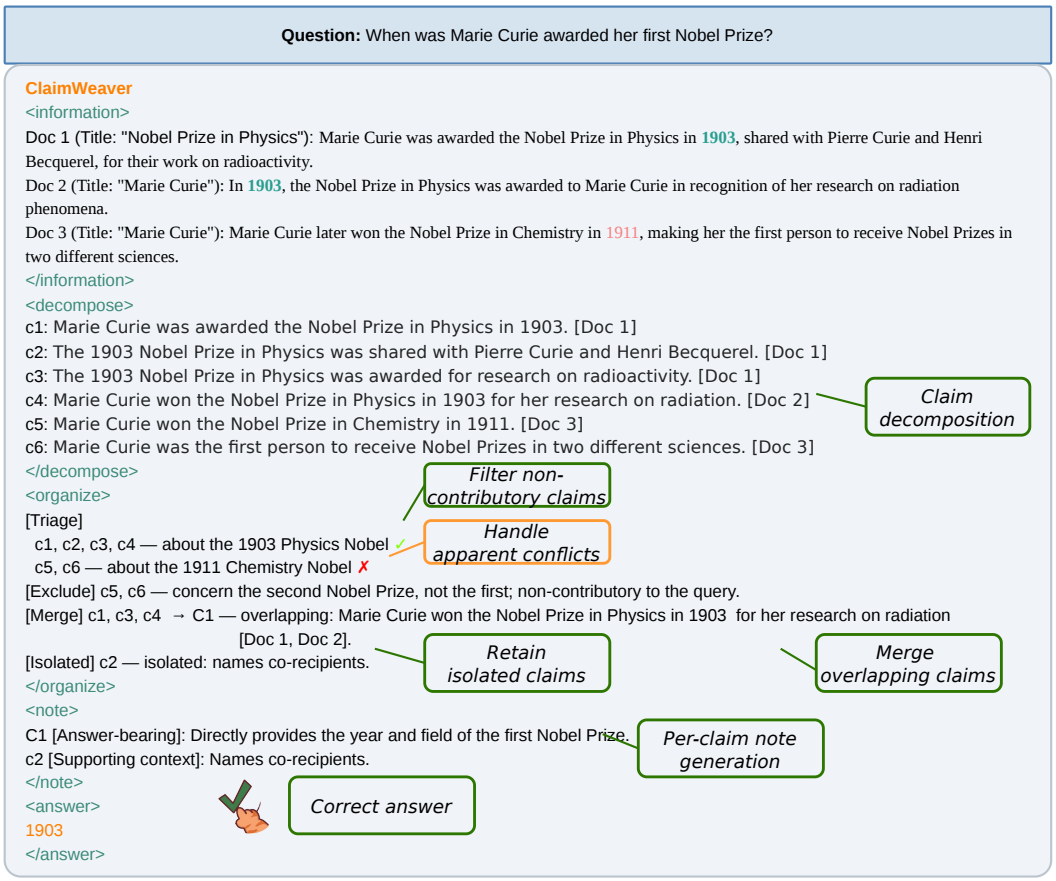


Figure 5: Case study on single-hop evidence organization. CLAIMWEAVER merges overlapping claims to reduce redundancy and excludes unrelated facts, while retaining isolated supporting context.

K.2 SUPPLEMENTARY CASE STUDY: OVERLAPPING MERGE AND ISOLATED RETENTION

Figure 5 considers the single-hop query *"When was Marie Curie awarded her first Nobel Prize?"* CLAIMWEAVER merges three claims supporting 1903, excludes the later 1911 Chemistry prize from the selected answer evidence, and retains an isolated contextual claim. This case illustrates redundancy reduction and conservative retention outside multi-hop bridging.

L STATISTICAL SIGNIFICANCE TESTS

Table 12 reports paired bootstrap tests (10K resamples) over evaluation examples. For CLAIMWEAVER, each example contributes its mean score across the three training seeds; this is paired with the corresponding mean score from the same three matched EviNote-RAG runs before resampling.

Dataset	Δ EM	Δ F1	p-value	Unadj. Sig.
HotpotQA	+5.0	+3.7	<0.001	**
2WikiMultiHopQA	+4.6	+4.4	<0.001	**
MuSiQue	+4.2	+3.4	<0.001	**
Bamboogle	+4.0	+3.2	0.003	**
NQ	0.0	+0.5	–	–
TriviaQA	+0.5	+1.3	0.04	*
PopQA	−0.4	+0.5	0.42	n.s.

Table 12: Per-dataset paired bootstrap significance tests (CLAIMWEAVER vs. EviNote-RAG, 10K resamples). **: $p < 0.01$; *: $p < 0.05$; n.s.: unadjusted $p \geq 0.05$. Tests use EM as the primary metric; NQ has zero mean EM difference.

All four multi-hop comparisons satisfy the seven-test Bonferroni threshold of $0.05/7 \approx 0.007$. The single-hop differences are small; TriviaQA's unadjusted $p = 0.04$ falls above the adjusted threshold.

M RELATIONAL STATE EVALUATION METHODOLOGY

Table 13 reports the quality of the intermediate organization judgments. The evaluation protocol and sampling procedure are described below.

Pair Relation	N_{pred}	Prec.	Rec.	H-Mean
Overlapping	50	92.0	84.6	88.1
Bridging	50	82.0	72.8	77.1

Table 13: Precision, recall, and descriptive cross-sample H-Mean of predicted pair relations. N_{pred} is the per-type precision sample size; recall uses a separate pool of 320 annotated candidate pairs. Relevance triage receives a separate unary evaluation. Isolated is the residual state for a retained claim with no pair relation.

Overlapping judgments achieve the highest precision (92.0%) in this annotated sample. Bridging judgments have lower precision (82.0%), with errors primarily involving direction reversals or missing intermediate steps. Precision and recall characterize spurious and missed relations on their respective annotated samples.

Precision evaluation. We randomly sample 50 model predictions for each pair-relation type (overlapping and bridging; 100 total) from the HotpotQA evaluation outputs. For each prediction, three annotators independently judge whether the relation is correct given the query, the two claims, and their source passages. Majority vote determines the gold label. Claim-level relevance triage uses a separate evaluation.

Recall evaluation. We draw a separate set of 80 queries from the HotpotQA evaluation split. For each query, annotators identify all claims participating in at least one non-isolated relation and annotate all pairs among these candidates, yielding 320 annotated claim pairs (average 4 per query). We then measure how many ground-truth relations were correctly detected by the model.

Independence of samples. The precision and recall samples are drawn from disjoint queries, keeping the two estimates based on separate annotated items. The reported H-Mean is the harmonic mean of these two sample-level estimates and serves as a descriptive cross-sample summary.

Sample size and interpretation. The relation study is a descriptive audit with 50 precision samples per type and 320 recall candidate pairs. These estimates characterize relation quality and recurring error patterns within the annotated sample.

N Organization Quality Under Outcome-Driven RL

The semantic correctness of organization receives no direct reward component. We examine its training trajectory, association with answer errors, and downstream sensitivity under fixed-evidence interventions.

Organization quality during training. We checkpoint the model at steps 0, 100, 200, 400, and 800, and measure relation F1 and triage precision on a fixed set of 100 HotpotQA validation examples with human-annotated relations and relevance decisions. Table 14 shows organization quality and answer accuracy increasing together under the answer, grounding, and format rewards.

Step	Relation F1	Triage Prec.	EM
0 (init)	31.2	52.4	23.0
100	58.6	71.8	37.0
200	68.4	78.2	45.0
400	74.8	83.6	50.0
800 (final)	79.2	86.4	54.0

Table 14: Organization quality and answer accuracy across training steps. Relation F1 measures overlap/bridge prediction; Triage Prec. measures the precision of relevance-triage decisions.

Association with answer correctness. We analyze 200 trajectories from the final model on HotpotQA validation (100 correct and 100 incorrect answers) and annotate organization errors. Among correct-answer trajectories, 91% have fully correct organization. Among incorrect-answer trajectories, 67% contain a wrong merge, missed bridge, or incorrect exclusion. The fixed-evidence intervention below complements this descriptive association by holding retrieved evidence constant.

Fixed-evidence interventions. We perturb the stored organization structure in 100 HotpotQA trajectories. Retrieved passages and search traces are held fixed; after each perturbation, we regenerate downstream notes and the final answer with greedy decoding. This design measures sensitivity to organization conditional on a shared retrieved evidence set:

- **Shuffle relations**: Randomly shuffle the overlap, bridge, and isolation assignments in the organization state. EM drops from 54.0 to 35.0 (-19.0).
- **Reverse chains**: Flip all bridging directions. EM drops to 39.0 (-15.0).
- **Remove exclusions**: Restore all excluded claims. EM drops to 45.0 (-9.0).
- **Disable merge**: Keep overlapping claims as separate entries. EM drops to 50.0 (-4.0).

Under fixed search traces, the performance drops establish direct downstream sensitivity to the learned organization and suggest a hierarchy: relational-state assignment and chain direction matter most, followed by relevance filtering, then deduplication.

Summary. Across the three analyses, organization quality improves during training, organization errors are associated with answer failures, and fixed-evidence perturbations reduce accuracy. The observed state quality and intervention effects connect organization to answer behavior under fixed search traces.

O Bridging Error Impact Analysis

The relational-state evaluation reported in the main paper shows lower cross-sample H-Mean for bridging than overlap predictions (77.1 versus 88.1), with errors including direction reversal and missed intermediate steps. To understand how these errors affect downstream answer quality, we analyze the 50 precision-sampled bridging predictions and cross-reference them with final answer correctness.

Error breakdown. Among the 9 incorrect bridging predictions (18% error rate out of 50 samples):

- **Direction reversal** (5 cases): The model correctly identifies that two claims are bridging-related but assigns the wrong direction (e.g., predicts $c_i \to c_j$ when the logical dependency is $c_j \to c_i$).
- **Missing intermediate** (3 cases): The model links two claims that require an intermediate step, skipping a necessary bridge entity.
- **Spurious bridge** (1 case): The model predicts a bridging relation between claims that are actually isolated.

Impact on final answers. Of the 9 bridging errors, 6 (67%) occur in trajectories with incorrect final answers. Four of the 5 direction reversals and 2 of the 3 missing-intermediate errors co-occur with wrong answers. The trajectory containing the single spurious bridge still yields a correct answer. These observational counts characterize associations between bridge-error types and answer outcomes in the annotated sample.

Implications. The error patterns complement the fixed-evidence intervention in which reversing chains causes the second-largest EM drop (-15.0). Together they motivate closer evaluation of direction assignment and missing intermediate claims.

P TOKEN-BUDGET COMPARISON

We evaluate the contribution of trajectory length with an inference-time diagnostic. EviNote-RAG and Search-R1 receive up to five retrieval turns, producing trajectory-token counts comparable to or larger than CLAIMWEAVER's three-turn trajectory.

Method	Traj. Tokens	EM	F1
Search-R1 (3 turns)	2,800	42.0	46.4
Search-R1 (5 turns)	4,200	43.4	47.8
EviNote-RAG (3 turns)	1,800	49.0	55.7
EviNote-RAG (5 turns)	3,100	50.2	56.4
CLAIMWEAVER (3 turns)	3,200	54.0	59.4

Table 15: Token-budget diagnostic on HotpotQA. Baselines receive additional retrieval turns using their original trained checkpoints.

With five retrieval turns, Search-R1 gains +1.4 EM and EviNote-RAG gains +1.2 EM. CLAIMWEAVER with three turns remains ahead by +10.6 and +3.8 EM, respectively. This diagnostic varies the retrieval horizon of each trained checkpoint and reports the resulting accuracy and trajectory-token count.

Q CLAIM FAITHFULNESS EVALUATION

To evaluate the faithfulness of decomposed claims, we randomly sample 200 claims from each of three settings: CLAIMWEAVER, $-R_{\text{grounding}}$, and Prompt-only (600 claims in total). Three annotators independently judge whether each claim is entailed by its cited source passage. We use majority vote as the final label. The average pairwise Cohen's κ is 0.82.

Setting	Faithful n (%)	Hallucinated n (%)
CLAIMWEAVER (Full)	188 (94.0)	12 (6.0)
$-R_{\text{grounding}}$	163 (81.5)	37 (18.5)
Prompt-only	143 (71.5)	57 (28.5)

Table 16: Claim-faithfulness evaluation on 200 manually annotated claims per setting (600 claims in total).

The full model has 188 of 200 decomposed claims (94.0%) judged faithful. Removing $R_{\text{grounding}}$ increases the hallucinated count from 12 of 200 (6.0%) to 37 of 200 (18.5%). This ablation associates the entailment-based reward with fewer unsupported claims. The Prompt-only setting has the highest hallucination rate (57 of 200; 28.5%).

Scope of the faithfulness audit. The labels target source entailment at the decomposition stage. Merged claims, generated notes, and final answers are downstream outputs requiring separate evaluation. The human labels measure claim quality independently of the frozen NLI training signal.

R Claim Count Analysis

Since $R_{\text{grounding}}$ averages entailment probability over generated claims, omitting low-entailment claims can increase this reward component. We audit this incentive by tracking claim generation statistics across training and at convergence.

Claim count during training. Table 17 reports the average number of claims generated per query at different training checkpoints, measured on the same 200 HotpotQA validation examples.

Step	Avg Decomp. Claims	R_{ground}	EM
0 (init)	4.2	0.61	25.5
100	6.8	0.78	39.0
200	7.4	0.84	45.5
400	7.8	0.89	50.0
800 (final)	7.6	0.91	53.5

Table 17: Average decomposed claims per query before relevance triage. The count increases during training and stabilizes between 7.6 and 7.8 at the last two checkpoints.

The average number of pre-triage claims increases from 4.2 at initialization to 7.6 at the final checkpoint. The late checkpoints remain close (7.8 at step 400 and 7.6 at step 800). Relevance triage occurs after the count used in $R_{\text{grounding}}$ is defined.

Final model claim statistics. At convergence, the model generates an average of 7.6 decomposed claims per query on multi-hop benchmarks and 5.2 on single-hop benchmarks. Manual inspection of 50 HotpotQA examples identifies an average of 8.1 relevant source claims per question. These averages provide descriptive coverage context. Proposition recall requires an explicit one-to-one or many-to-one matching protocol.

Coverage pressure from $\mathbf{R}_{\text{answer}}$. The answer reward has weight $\lambda_1 = 1.0$, compared with $\lambda_2 = 0.3$ for grounding, and supplies implicit pressure to retain answer-supporting claims. On multi-hop questions, dropping a required bridge can change R_{answer} from 1 to 0, while deleting one of 7–8 claims changes the weighted grounding contribution by at most approximately 0.04. This asymmetry discourages broad coverage reduction. Selective omission of difficult, low-entailment propositions remains a possible failure mode.

S Human Annotation Details

We conduct two three-annotator human evaluation tasks: claim faithfulness evaluation (600 claims) and pair-relation prediction quality (100 precision samples + 320 recall candidate pairs). Both tasks employed the same pool of three annotators. The separate manual inspections used for stage attribution, organization diagnostics, and descriptive claim-count analysis form additional diagnostic audits, each with its own sample and evaluation unit.

Recruitment and qualifications. Annotators were recruited from graduate students with backgrounds in NLP or computational linguistics. All annotators had prior experience with annotation tasks involving natural language inference and information extraction.

Compensation. Annotators were compensated at an hourly rate above the local minimum wage. Total annotation time was approximately 25 hours across all tasks.

Consent and data use. Annotators were informed of the research purpose and the intended use of their annotations for aggregate evaluation metrics prior to participating. They were informed of their right to withdraw at any time without penalty. No personally identifying information about annotators is collected or released in this work.

T Domain and Task Transfer

Transfer protocol. We evaluate the trained model in two transfer settings without further updates. For biomedical domain transfer, we use the 85 factoid questions from the four official test batches of BioASQ 12b Task B Phase B (Nentidis et al., 2024), together with the provided PubMed articles and snippets as evidence. We report Strict Accuracy (SAcc), Lenient Accuracy (LAcc), and Mean Reciprocal Rank (MRR), macro-averaged over batches. For long-form task-format transfer, we use the ASQA development split (Stelmakh et al., 2022) after screening exact and normalized question overlap against the NQ and HotpotQA training data. All systems use the same long-form answer prompt, retrieval/context budget, and maximum output length. We report ROUGE-L, Disambig-F1, and their official combined score (DR). ASQA is NQ-derived, so this evaluation measures task-format transfer.

Method	SAcc	LAcc	MRR
Naive RAG	26.8	39.6	31.9
EviNote-RAG	31.8	46.6	37.4
ClaimWeaver	**34.6**	**49.8**	**40.1**

Table 18: Zero-shot transfer on BioASQ 12b Phase B factoid QA ($N = 85$), using the official PubMed evidence. Scores are macro-averaged over four test batches.

Method	R-L	D-F1	DR	Len.
Naive RAG	27.4	31.6	29.4	68.0
EviNote-RAG	30.2	36.4	33.2	71.4
ClaimWeaver	**31.6**	**38.2**	**34.7**	70.8

Table 19: Zero-shot task-format transfer on overlap-screened ASQA development questions. R-L: ROUGE-L; D-F1: Disambig-F1; Len.: average output length.