FACTWEAVER: ATOMIC FACT GRAPH-GUIDED NON-FACTOID QUESTION ANSWERING OVER LONG DOCUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Non-factoid questions seek explanations, comparisons, or assessments beyond a single fact. Answering them over long documents requires identifying dispersed facts relevant to a broad analytical question, determining their relationships, and integrating them into a coherent analysis. We introduce FactWeaver, an atomic fact graph-guided framework for non-factoid question answering over long documents. FactWeaver organizes question-relevant atomic facts under analytical perspectives and uses the retrieved context to establish candidate sequence, support, and conflict relations. The resulting atomic fact graph guides multiple agents in constructing candidate reasoning paths and auditing them for defects. Revision requirements are recorded in a shared state and guide answer generation. To evaluate the effectiveness of the framework, we construct LongNFQA, a benchmark comprising 624 non-factoid questions over 52 long documents totaling 6,048 pages. Human-reviewed evidence supports retrieval evaluation, while a human-derived multidimensional rubric and an automated judge validated against held-out human judgments support answer-quality assessment. On LongNFQA, FactWeaver improves Perspective Satisfaction@20 by 0.1384 and the Avg. answer-quality score by 0.5203 points over the strongest baseline. Ablation studies support the contributions of both atomic fact graph construction and graph-guided synthesis.

1 INTRODUCTION

Non-factoid questions seek explanations, comparisons, or assessments rather than a single entity or a short answer span (Soricut & Brill, 2004; Bolotova et al., 2022). They express broad information needs that require the integration of multiple aspects of a topic (Hashemi et al., 2020; Rosset et al., 2025). In long documents, relevant facts are often dispersed in narrative sections, tables, and charts, with varying scopes and levels of detail (Chia et al., 2025; Dong et al., 2025). Non-factoid question answering over long documents is therefore particularly challenging, as it requires identifying the facts needed to address the user's analytical goal, determining their relationships, and explaining which conclusions they jointly support.

Prior work has explored a range of approaches to non-factoid and long-document question answering. For non-factoid questions, Typed-RAG (Lee et al., 2025) guides the retrieval by question type, while LANCER (Ju et al., 2026) uses synthetic questions to improve the coverage of the answer-relevant information. For long documents, RAPTOR (Sarthi et al., 2024) builds hierarchical summaries, DOS-RAG (Laitenberger et al., 2025) preserves document order, HippoRAG 2 (Gutiérrez et al., 2025) performs graph retrieval, and MDocAgent (Han et al., 2025) coordinates document agents to locate and organize dispersed evidence. Although these approaches improve evidence access and aggregation, they leave unresolved how dispersed evidence should be integrated into comprehensive and coherent non-factoid answers over long documents.

Bridging this gap requires addressing three challenges. First, broad analytical questions leave their factual requirements implicit, whereas relevant facts are scattered throughout the document. Answering them requires both identifying the necessary facts and organizing substantial evidence into a coherent analysis that accounts for conflicts and evidential limitations. Second, suitable benchmarks rarely combine long documents, non-factoid questions, and source-level evidence needed to

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.

evaluate evidence acquisition. Third, the same question may admit multiple valid analyses, making open-ended answer quality difficult to assess consistently.

We propose FactWeaver, an atomic fact graph-guided framework for non-factoid question answering over long documents. FactWeaver first selects question-relevant analytical perspectives from the document and extracts atomic facts within each perspective. For each fact, it constructs a factoid query and uses the query to retrieve and rerank supplementary evidence from the same document. This evidence provides broader context for interpreting the facts jointly and informs the induction of candidate sequence, support, and conflict relations among the atomic-fact nodes. Using the resulting atomic fact graph, multiple agents construct candidate reasoning paths that organize selected facts and specify how they support the proposed conclusions. The paths are then audited for missing perspectives, unsupported claims, and unresolved conflicts. Audit findings are consolidated into revision requirements recorded in a shared state, which guide changes to fact selection, claim scope, and ordering. Finally, a revised path is selected to guide answer generation and verification.

To support controlled evaluation in this setting, we construct LongNFQA, a benchmark comprising 624 non-factoid questions over 52 long documents, with human-reviewed source-level evidence. We also derive a multidimensional answer-quality rubric from human assessments and validate its frozen automated judge against held-out human judgments (Hashemi et al., 2024; Liu et al., 2026). Compared with the strongest evaluated baseline, FactWeaver increases Perspective Satisfaction@20 from 0.5458 to 0.6842 and the mean answer-quality score from 4.0697 to 4.5900. Ablation studies demonstrate the contributions of atomic-fact nodes, graph relations, and multi-agent synthesis, with retrieved evidence held fixed where applicable.

Our contributions are threefold.

- We propose FactWeaver, a framework for non-factoid QA over long documents that constructs an atomic fact graph to guide multi-agent synthesis. The graph connects question-relevant atomic facts through candidate relations, enabling agents to construct, audit, and revise reasoning paths before answer generation.
- We introduce LongNFQA, a benchmark of 624 non-factoid questions over 52 long documents with human-reviewed source-level evidence for retrieval evaluation.
- We establish a human-derived multidimensional rubric and validate its frozen automated judge against held-out human judgments for answer-quality evaluation.

2 Related Work

Non-factoid QA addresses explanations and complex information needs. Early work studies passage-level answering and ranking (Soricut & Brill, 2004; Surdeanu et al., 2011), while taxonomies categorize intents (Bolotova et al., 2022). ANTIQUE (Hashemi et al., 2020), ELI5 (Fan et al., 2019), NLQuAD (Soleimani et al., 2021), and Researchy (Rosset et al., 2025) cover community retrieval, long-form explanation, long-answer extraction, and multi-perspective QA.

Concrete facts connect broad questions to retrievable evidence. Typed-RAG (Lee et al., 2025) uses question types; LANCER (Ju et al., 2026) uses synthetic nugget questions for reranking. Dense X Retrieval (Chen et al., 2024) indexes propositions, and FActScore (Min et al., 2023) assesses atomic claims. These methods aid evidence selection or factual assessment, but leave the relations among selected facts for analytical synthesis implicit. FactWeaver infers candidate relations from the retrieved context.

Long-document QA requires connecting evidence dispersed across passages (Lewis et al., 2020; Izacard & Grave, 2021). At the document level, RAPTOR (Sarthi et al., 2024) builds hierarchical summaries, while DOS-RAG (Laitenberger et al., 2025) preserves the original order of the retrieved passages. At a finer granularity, GraphRAG (Edge et al., 2024) organizes entity communities, HippoRAG 2 (Gutiérrez et al., 2025) links phrases and passages, and PropRAG (Wang & Han, 2025) searches the paths of related propositions. These approaches organize evidence for retrieval and aggregation, but their structures are not designed to distinguish sequence, support, and conflict as explicit relations among atomic facts.

Other work guides how evidence is used after or during retrieval. Self-RAG (Asai et al., 2024) critiques retrieval and generation; CoRAG (Wang et al., 2025) and ReAct (Yao et al., 2023) interleave

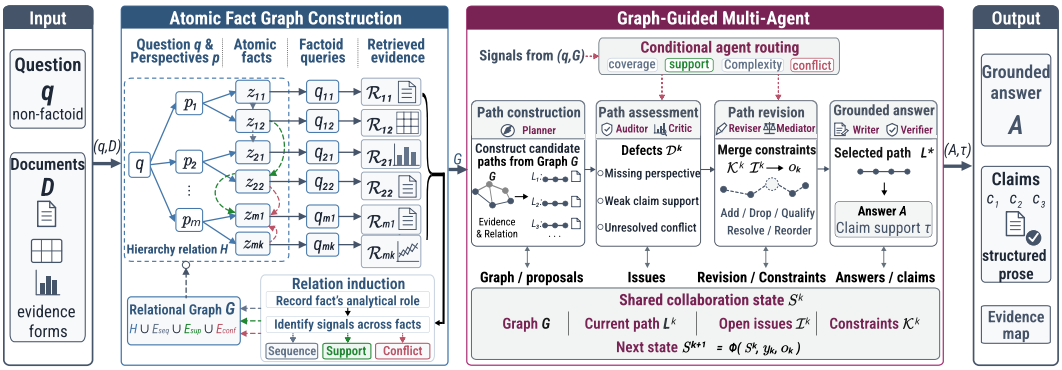

Figure 1: Overview of FactWeaver. Graph construction identifies perspectives and atomic facts, retrieves context, and induces candidate relations. Multi-agent synthesis constructs, assesses, and revises reasoning paths using a shared state recording constraints and unresolved issues before writing. Dashed arrows indicate conditional review and mediation; vertical arrows indicate access to shared reasoning information and its updates.

reasoning with retrieval or tools. MA-RAG (Nguyen et al., 2025) and MDocAgent (Han et al., 2025) coordinate specialized agents, the latter combining textual and visual evidence; ALCE (Gao et al., 2023) evaluates the support of citations in generated answers. These methods improve the use, coordination, or evaluation of evidence, but leave its fact-level relations implicit. FactWeaver addresses this gap by representing atomic facts and candidate sequence, support, and conflict relations in a shared graph. Agents audit and revise reasoning paths in that graph under common constraints before generating an answer.

3 Method

FactWeaver connects dispersed facts across documents to a coherent analytical answer. Graph construction identifies question-relevant facts and their candidate relations (Section 3.2); synthesis organizes and checks the reasoning based on these facts (Section 3.3). Figure 1 summarizes this process.

3.1 Problem Formulation and Overview

Given a non-factoid question q and a collection of long documents D, FactWeaver produces an answer A. A claim c is a factual or analytical assertion proposed for the answer whose validity must be assessed against the evidence in the document. PDF parsing produces page-aligned text, table, and chart records, normalized into a source-node catalog $\mathcal{U}_D$. Each source node retains its content, evidence form, and document location. The claim to-evidence map τ links each answer claim to its supporting source nodes. Graph nodes represent atomic facts supported by source records. Defining FactWeaver as $\mathcal{F}$, the task is

$$(A, \tau) = \mathcal{F}(q, D), \qquad \tau(c) \subseteq \mathcal{U}_D. \tag{1}$$

Graph construction selects analytical perspectives $\mathcal{P}_q$: complementary aspects that the answer should address. It extracts atomic facts z_{ij}, indexed by perspective i and fact j, and constructs factoid queries q_{ij} to retrieve evidence $\mathcal{R}_{ij} \subseteq \mathcal{U}_D$. This evidence informs relations in the atomic fact graph G. Synthesis develops candidate reasoning paths $\mathcal{L}$ that organize facts and proposed claims, and selects a revised path L^* for writing. The overall flow is

$$(q, D) \to \mathcal{P}_q \to \{z_{ij}\} \to \{q_{ij}\} \to \{\mathcal{R}_{ij}\} \to G \to \mathcal{L} \to L^* \to (A, \tau). \tag{2}$$

3.2 Atomic Fact Graph Construction

The graph must retain both the breadth of the analytical intent and the atomic facts needed to address it. We organize atomic facts into document-supported perspectives, attach retrieved evidence, and

use that evidence to propose connections. This separates the factual content of the graph from the candidate interpretations subsequently examined during synthesis.

Analytical scope. Perspective selection keeps fact acquisition focused on the question while covering its analytical scope. To identify candidate perspectives, a language model processes page-aligned chunks following major section boundaries, keeping each input within the model's context window, and induces local analytical concepts, each described by a name, definition, and source-page span. Document-level consolidation merges concepts with the same analytical role and aggregates their provenance, while retaining complementary concepts separately. The resulting inventory $\mathcal{P}_D$ supplies candidates supported by documents. For a candidate p, the prompted selector $g(q, p)$ returns 1 when p is relevant to q and adds a distinct analytical requirement and 0 otherwise. It selects

$$\mathcal{P}_q = \{p \in \mathcal{P}_D : g(q, p) = 1\}. \tag{3}$$

Atomic facts. Within each selected perspective p_i, text propositions, table rows, and normalized chart records provide independently verifiable atomic facts z_{ij}. We adopt the atomic-claim representation used by FActScore (Min et al., 2023), here to construct a graph rather than to score generated text. Each fact retains the seed provenance that identifies its document source. Its perspective specifies why it matters: the same investment figure, for example, may inform both margin pressure and capacity expansion. The objective of the answer, the perspectives and their atomic facts form a hierarchy H, with the perspectives as intermediate nodes and the atomic facts as leaves.

Retrieving evidence for relations. The local source of a fact establishes its provenance, but may not provide sufficient context to relate it to other facts. Establishing these relations may require identifying the relevant scope of the comparison or retrieving corroborating evidence. Let t denote the type of original question, following the taxonomy of evidence-based, comparison, experience, reason, instruction, and debate (Bolotova et al., 2022). The query constructor ψ_t uses this type to choose a factual template and expected answer form. For each atomic fact, it constructs a perspective-conditioned factoid query:

$$q_{ij} = \psi_t(p_i, z_{ij}). \tag{4}$$

Here, z_{ij} provides the specific factual target and p_i provides its analytical role. Templates include *what*, *when*, *who*, and quantitative questions. Broad *why/how* intent remains in q and the perspectives, while the queries seek the factual material needed to assess that intent. The candidates retrieved are reranked against q, p_i, and q_{ij}, resulting in $\mathcal{R}_{ij}$. Each atomic fact remains a leaf of H, with its retrieved source records attached as supporting evidence.

Inducing candidate connections. The relations encode candidate ways in which atomic facts may contribute to a joint analysis. Their induction combines two sources. Within each perspective, the retrieved evidence informs a local plan that assigns each atomic fact an analytical role and proposes an initial ordering. Across perspectives, semantic analysis links atomic facts that share a question-relevant context and detects potential numerical, causal, temporal, or scope tensions. Together with the hierarchy inherited from H, these signals define candidate edges in G: adjacent facts in a local plan receive sequence edges, cross-perspective facts that share relevant context receive support edges, and facts implicated in the same tension receive conflict edges. All induced relations remain provisional until path assessment. Consequently, G contains four types of relationships:

- **Hierarchy (H):** connects the objective answer, analytical perspectives, and atomic facts, encoding the scope and perspective membership of each fact.
- **Sequence (E_{seq}):** records the initial within-perspective order in which atomic facts can be introduced and combined during synthesis.
- **Support (E_{sup}):** links atomic facts between perspectives whose shared domain context suggests that they can jointly substantiate an analytical claim.
- **Conflict (E_{conf}):** connects atomic facts involved in the same numerical, causal, temporal, or scope tension, marking evidence that requires reconciliation or qualification.

3.3 Graph-Guided Multi-Agent Synthesis

Synthesis turns the graph into an analytical answer by deciding which claims its atomic facts jointly support and how those claims should be organized. Agents operate on explicit reasoning paths and

shared audit findings rather than independently drafting answers. Path construction, assessment, and revision precede answer realization; the graph remains fixed while the reasoning paths change.

Constructing candidate reasoning paths. The Planner proposes alternative paths $\mathcal{L}$ from the atomic facts and relations in G, informed by perspective-local briefs. Each path L contains ordered atomic-fact nodes $\mathbf{v}_L$, proposed claims $\mathcal{C}_L$, and conflict or qualification notes $\mathcal{N}_L$, giving $L = (\mathbf{v}_L, \mathcal{C}_L, \mathcal{N}_L)$. A path specifies how selected facts and relations lead to proposed answer claims; subsequent assessment may revise both the ordering of those facts and the scope of the claims.

Assessing coverage and support. The Auditor assesses each candidate path against the question, graph G, and the selected perspectives. The resulting defect record $\mathcal{D}^k$ distinguishes three principal deficiencies: missing perspective coverage, claims that are insufficiently supported by the selected atomic facts, and unresolved conflicts among those facts. Each entry identifies the affected part of the path and the reason that a revision is required (Appendix C, Table 9). The complexity of the question and the coverage, support and conflict signals at the graph-level determine whether an Extra Auditor, Domain Critic, or Mediator participates. These roles contribute assessments and constraints to the same reasoning process; the Auditor and the Reviser participate for every question.

Revising through shared constraints. Revision converts the assessment into changes to the reasoning path. Feedback is consolidated into accepted requirements and unresolved issues, with mediation when activated. In step k, let L^k be the current path, $\mathcal{K}^k$ the accepted constraints, and $\mathcal{I}^k$ the unresolved problems. A structured role output y_k informs the requested change o_k. The update rule Φ combines constraint updates, LLM-based revision, and structural checks. With G fixed, the shared state S^k evolves as

$$S^k = (G, \mathcal{K}^k, \mathcal{I}^k, L^k), \qquad S^{k+1} = \Phi(S^k, y_k, o_k), \tag{5}$$

Constraints specify what the path must change, such as covering an omitted perspective. Fully supported claims are retained, while identified defects are addressed through the following operations:

$$o_k = \begin{cases} \mathrm{AddNode}, & \text{a required perspective is missing,} \\ \mathrm{DropClaim}, & \text{a claim is unsupported,} \\ \mathrm{QualifyClaim}, & \text{a claim is partially supported,} \\ \mathrm{ResolveConflict}, & \text{an exposed conflict remains,} \\ \mathrm{ReorderPath}, & \text{the evidence order is inconsistent.} \end{cases} \tag{6}$$

The Reviser applies the operations triggered by $\mathcal{D}^k$ jointly under $\mathcal{K}^k$, coordinating changes to evidence selection, claim scope, conflict resolution, and ordering. These changes are interdependent: adding evidence to restore perspective coverage, for example, may require an existing claim to be qualified or repositioned within the path. All changes in evidence are restricted to atomic facts in G and re-validated against the graph before selecting L^*.

Realizing the grounded answer. Path recommendations and audit feedback jointly guide the selection of the final path L^*. The Writer expresses its claims as A, retaining the accepted constraints and outstanding issues. The Verifier checks the draft against the selected path and constraints, making targeted corrections where needed. The source records attached to the atomic facts supply the claim-to-evidence map τ. Thus, the final answer inherits an assessed reasoning path and its evidence, rather than relying on an unconstrained interpretation of a flat collection. Appendix C specifies execution, routing, and failure handling.

4 LongNFQA Benchmark Construction

Evaluating whether a system acquires the facts needed for analysis and uses them effectively requires both source-level annotations and assessment of open-ended answers. The benchmarks address complementary but incomplete conditions: FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021) and MP-DocVQA (Tito et al., 2023) emphasize localized factoid or numerical answers, while ANTIQUE (Hashemi et al., 2020), ELI5 (Fan et al., 2019), NLQuAD (Soleimani et al., 2021) and SQuALITY (Wang et al., 2022) lack the alignment of evidence based in document or the evaluation of dispersed-evidence required here. LongNFQA supports separate retrieval and answer-quality evaluation over long reports; Appendix B details these boundaries.

4.1 Dataset and Question Construction

LongNFQA contains 624 questions on 52 PDF documents, 6,048 pages, with 104 questions per question type (Table 1). Stratified purposive sampling covers five report families (Table 1), varying conditions without estimating population frequencies. PDF parsing produces text, table, and chart nodes with document, page, and section provenance; tables retain titles, units, and footnotes.

The construction of questions targets document-supported analysis that requires complementary records. GPT-5.5 proposes questions related to analyst needs, complementary records, and document structure; the released NF-CATS classifier (Bolotova et al., 2022) initializes six categories of labels, followed by human review. All evaluated methods receive the reviewed question-type label as part of their initial input. Reviewers retain only scoped, multi-record questions that are neither factoid, ambiguous, single-span, nor external-knowledge dependent. An automated evaluator assesses the sufficiency of the document, and human reviewers make the final inclusion decisions (Joren et al., 2025). Appendix B details preprocessing and annotation statistics.

Table 1: LongNFQA: document-family composition (left) and corpus-wide statistics (right).

Document family	PDFs	Questions	Property	Value
Financial/business	22	264	Pages per PDF (min./mean/max.)	51 / 116.3 / 201
Scientific/technical	8	96	Canonical source nodes	21,646
Medical/public-health	8	96	Text / table / chart nodes	20,371 / 1,191 / 84
Policy/public-sector	8	96	Mean gold nodes per question	8.88
Regulatory/safety	6	72	Gold node assignments (text / table)	4,449 / 1,095

4.2 Gold Evidence Annotation

Gold evidence is annotated at the source-node level. Nodes used to construct each question initialize its candidate evidence set; GPT-5.5 then proposes additional candidates from the full source-node inventory. Annotators inspect all candidates in their source context and retain only those that directly support the answer, including necessary qualifications and boundary conditions. Each retained node receives a provenance-aware selection rationale, and the resulting question-level set supplies the reference labels for node-level retrieval metrics.

Annotation is independent of system outputs. Annotators do not observe method-generated intermediate representations, retrieval outputs, or graph relations when making node-level decisions. Each selected node is reviewed individually, and uncertain cases are adjudicated. Final labels and subsequent corrections retain their source provenance and version history and are stored separately from method inputs and answer-quality judgments.

5 Experimental Setup

Our experiments test whether graph construction recovers evidence relevant to the analytical intent and whether graph-guided synthesis turns it into well-supported answers.

5.1 Evaluation Protocol

Source-level metrics assess evidence recovery independently of the system's graph; the answer rubric evaluates the resulting analysis.

Retrieval quality. Precision@20, Recall@20, and F1@20 assess node relevance, while Perspective Satisfaction@20 measures coverage across distinct parts of the requested analysis. After fixing node-level gold labels, we group retained evidence by answer requirement into evaluation perspectives, derived from the question and reviewed evidence independently of the method-selected $\mathcal{P}_q$. Applicable retrievers use a frozen source-node catalogue; the first 20 nodes are scored without post-hoc reranking, matching the minimum selected-node count across evaluated methods. For question q, let $\mathcal{P}_q^{\text{gold}}$ denote the evaluation perspectives, E_p^{req} the required evidence assigned to perspective

p, and $\mathcal{R}_q^{20}$ the first 20 retrieved nodes. Then

$$\mathrm{PS@20}(q) = \frac{1}{|\mathcal{P}_q^{\mathrm{gold}}|} \sum_{p \in \mathcal{P}_q^{\mathrm{gold}}} \mathbf{1}\left[E_p^{\mathrm{req}} \subseteq \mathcal{R}_q^{20}\right]. \tag{7}$$

An evaluation perspective is satisfied when its required evidence is retrieved. Precision, Recall, and Perspective Satisfaction are macro-averaged over questions; F1 is the harmonic mean of aggregate Precision and Recall. Thus, node-level metrics depend only on the independently reviewed gold set, whereas Perspective Satisfaction evaluates the subsequent grouping of those labels.

Answer quality. An evidence-grounded rubric assesses answers beyond surface-form similarity (Xu et al., 2023; Yang et al., 2024). Questions form document-disjoint development (84), held-out human-validation (84), and automatic-evaluation (456) sets. Two reviewers independently assess anonymized development answers against reviewed evidence; a third adjudicates disagreements. Following multidimensional rubric evaluation (Vu et al., 2024; Hashemi et al., 2024; Liu et al., 2026), their rationales and recurring defects are consolidated into dimension definitions, failure descriptions, and score levels 1–5 (Kim et al., 2024). The rubric covers Factual and Numerical Correctness (FNC), Evidence Grounding (EG), Overclaim Control (OC), Completeness (CMP), Reasoning Soundness (RS), and Focus and Clarity (FC); Avg. is their unweighted mean.

Held-out validation assesses judge agreement with human rankings (Saad-Falcon et al., 2024). With the rubric and prompt frozen after development, the judge achieves mean within-question Spearman $\rho = 0.890$ on 84 held-out questions (1,260 answers), indicating strong agreement with human answer rankings. Held-out labels do not revise the protocol. Appendix A gives the full protocol.

5.2 Baselines, Ablations, and Controls

Nine baselines span retrieval and synthesis strategies under a shared generator: no-document generation (Direct LLM), single-pass retrieval (Naive RAG), hierarchy or order preservation (RAPTOR (Sarthi et al., 2024) and DOS-RAG (Laitenberger et al., 2025)), graph memory (HippoRAG 2 (Gutiérrez et al., 2025)), adaptive retrieval (Self-RAG (Asai et al., 2024), CoRAG (Wang et al., 2025), and Search-R1 (Jin et al., 2025)), and type-aware decomposition (Typed-RAG (Lee et al., 2025)). Prompted adaptive protocols replace specialized training; Appendix D specifies all adapters.

Eleven ablations target FactWeaver's three stages: *evidence acquisition* (atomic-fact acquisition, perspective selection, and atomic-fact queries; w/o AFA/, w/o PS/, w/o AFQ), *relational organization* (graph representation, relational edges, and retrieved relation context; w/o GR/, w/o RE/, w/o RC), and *graph-guided synthesis* (collaboration, path construction, conditional roles, path assessment and revision, and grounded realization; w/o CS/, w/o PC/, w/o Cond/, w/o PAR/, w/o GAR). Appendix D defines their retained components and execution conditions.

Controls align source nodes, corpus scope, evidence context, generator, decoding, and citation instructions where applicable (Appendix D). Direct LLM and external-search Search-R1 are excluded from retrieval evaluation; every method enters answer-quality evaluation.

6 Results

Results examine the two modules underlying FactWeaver: constructing a graph from dispersed information relevant to the analytical intent, and using it to develop a supported answer. We report evidence retrieval, answer quality, ablations, and token consumption over 624 LongNFQA questions. All generative components use Qwen3.5-Flash, with method-specific retrievers retained.

6.1 Evidence Retrieval

We compare the first 20 nodes, matching the minimum output count across evaluated retrievers. FactWeaver exceeds Typed-RAG by 0.1384 in PS@20 and 0.0749 in F1@20, with higher precision and recall at the same evaluation depth. The recovered evidence thus better matches both individual gold nodes and the reviewed analytical requirements. Appendix E provides subgroup results.

Table 2: The Retrieval results at $k = 20$ (↑). Best values in each column are bold.

Method	Precision@20	Recall@20	F1@20	PS@20
Naive RAG	.1759	.4193	.2478	.2472
DOS-RAG	.1756	.4286	.2491	.2560
RAPTOR	.1965	.4801	.2789	.3140
Self-RAG	.2159	.5293	.3067	.3443
HippoRAG 2	.2368	.5809	.3364	.4050
CoRAG	.2582	.6356	.3672	.4661
Typed-RAG	.2759	.6690	.3907	.5458
FactWeaver	**.3358**	**.7590**	**.4656**	**.6842**

6.2 Answer Quality

Table 3 shows that FactWeaver leads all baselines on Avg. and all six criteria. The Avg. gain of FactWeaver over Typed-RAG is 0.5203. Improvements are largest in RS (+0.6971), OC (+0.6058), and FNC (+0.5833), compared with +0.2179 in CMP. The benefit therefore extends beyond covering more content to how evidence is interpreted and integrated, consistent with the objective of graph-guided synthesis. The 95% paired corpus-unit resampling interval for the Avg. gain is [0.4640, 0.5710], conditional on finalized scores and corpus composition (Appendix E.4). Gains occur across all five document families and six question types; component contributions are examined below.

Table 3: Main answer-quality results (1–5; ↑). Best and second-best values are bold and underlined, respectively. Bar lengths are min–max normalized within each metric across displayed methods.

Method	FNC ↑	EG ↑	OC ↑	CMP ↑	RS ↑	FC ↑	Avg. ↑
Direct LLM	2.5385	2.2708	2.7917	2.3173	2.1619	2.7692	2.4749
Naive RAG	2.4920	2.8446	2.7276	2.6603	2.6202	3.0080	2.7254
DOS-RAG	2.9247	3.2644	3.1731	3.1330	3.0481	3.4151	3.1597
RAPTOR	2.6955	3.0657	2.9423	2.9167	2.8574	3.2821	2.9599
Self-RAG	3.3301	3.6971	3.5192	3.4808	3.4006	3.8734	3.5502
HippoRAG 2	3.3830	3.7484	3.5994	3.5833	3.4856	3.8910	3.6151
CoRAG	3.7516	4.1042	3.8910	3.8990	3.8237	4.2901	3.9599
Search-R1	3.2788	3.1442	3.6090	3.1218	3.0817	3.4439	3.2799
Typed-RAG	<u>3.8798</u>	<u>4.1522</u>	<u>4.0401</u>	<u>4.0673</u>	<u>3.9295</u>	<u>4.3494</u>	<u>4.0697</u>
FactWeaver	**4.4631**	**4.7324**	**4.6458**	**4.2853**	**4.6266**	**4.7869**	**4.5900**

6.3 Component Analysis

Table 4 organizes the ablations by the three stages of FactWeaver. Each block contains a broad stage substitution and, where applicable, more localized controls.

Evidence acquisition. The joint removal of atomic-fact acquisition lowers Avg. by 1.2652. Its nested controls further show lower Avg. after removing perspective selection (0.2999) or replacing atomic-fact queries (0.4201). These answer-quality comparisons support the document-anchored acquisition stage under the evaluated configuration.

Relational organization. With evidence fixed, removing only relational edges lowers Avg. by 0.5601, including a 0.7179 decrease in RS; restricting relation induction to seed context lowers Avg. by 0.3405. The broader paragraph-based graph-representation intervention produces a larger 0.8502 decrease, consistent with the joint contribution of atomic-fact structure and explicit relations.

Graph-guided synthesis. Removing path assessment and revision lowers Avg. by 0.6301, whereas removing the entire path-construction dependency chain lowers it by 0.7097. Replacing grounded answer realization with ordinary path-to-prose generation lowers Avg. by 0.3603. At the broader protocol level, one-pass synthesis lowers Avg. by 1.1402, while removing conditional review and mediation lowers it by 0.1800. Together, these comparisons support candidate-path construction, defect-driven revision, and grounded realization as complementary stages.

Table 4: Ablations method results (1–5; ↑). Best and second-best values are bold and underlined, respectively. Bar lengths are min–max normalized within each metric across displayed methods.

Method	FNC ↑	EG ↑	OC ↑	CMP ↑	RS ↑	FC ↑	Avg. ↑
Evidence acquisition							
w/o AFA	3.3301	3.3381	3.5096	3.0449	3.3365	3.3894	3.3248
w/o PS	<u>4.1891</u>	4.3974	4.3301	4.1074	4.2949	4.4215	4.2901
w/o AFQ	3.9936	4.3221	4.2131	3.9311	4.1474	4.4119	4.1699
Relational organization							
w/o GR	3.5080	3.9231	3.6987	3.6010	3.7532	3.9551	3.7399
w/o RE	3.9647	4.2067	3.9487	3.9519	3.9087	4.1987	4.0299
w/o RC	4.1474	4.3606	4.2564	4.1026	4.2051	4.4247	4.2495
Graph-guided synthesis							
w/o CS	3.1202	3.7901	3.3221	3.2676	3.4519	3.7468	3.4498
w/o PC	3.7436	4.0032	3.8942	3.7196	3.8686	4.0529	3.8803
w/o COND	4.1747	<u>4.6619</u>	<u>4.3990</u>	<u>4.1362</u>	<u>4.4487</u>	<u>4.6394</u>	<u>4.4100</u>
w/o PAR	3.8446	4.1122	3.8638	3.9071	3.8429	4.1891	3.9599
w/o GAR	4.0849	4.3446	4.1571	4.0897	4.3253	4.3766	4.2297
FactWeaver	**4.4631**	**4.7324**	**4.6458**	**4.2853**	**4.6266**	**4.7869**	**4.5900**

6.4 TOKEN CONSUMPTION

FactWeaver achieves an Avg. 0.6301 points higher than CoRAG while using 4.2% fewer tokens; relative to Typed-RAG, its 0.5203-point gain requires 14.6% more tokens. Within FactWeaver, removing path assessment and revision reduces token use by 14.9% but lowers Avg. by 0.6301, whereas removing conditional roles saves 8.0% with a 0.1800-point decrease. The larger change under the former ablation indicates that path assessment and revision materially affect answer quality and token consumption, while conditional routing produces a smaller quality–cost adjustment under this configuration. Appendix E reports exact totals for these quality–token trade-offs.

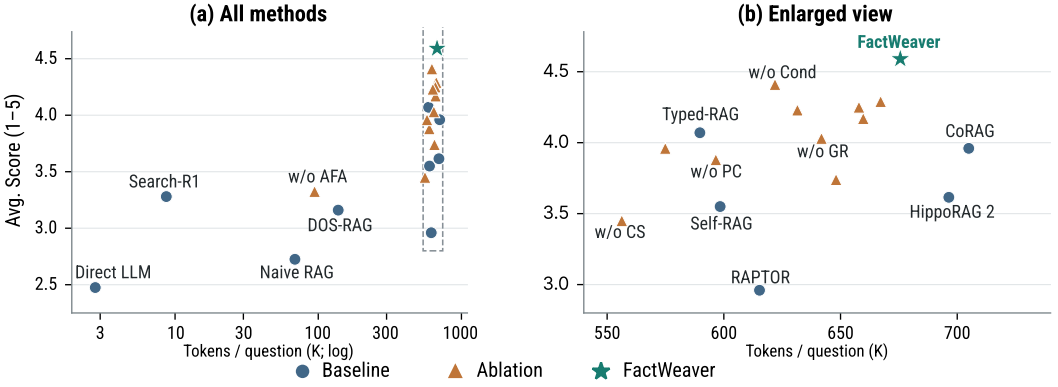

Figure 2: Answer quality and mean online token consumption. Each point averages the same 624 questions; (a) uses a logarithmic token axis, and (b) enlarges the dashed region on a linear axis.

7 CONCLUSION

FactWeaver addresses long-document non-factoid QA by constructing an atomic fact graph for multi-agent synthesis. Analytical perspectives organize dispersed facts, retrieved evidence informs candidate relations, and shared-state path revision organizes and checks reasoning before writing. LongNFQA separately assesses evidence acquisition and answer quality using reviewed evidence and a human-derived rubric. Results support complementary roles of atomic-fact graph construction and graph-guided synthesis: evidence recovery is necessary, while organization and use affect the resulting analysis. These gains increase token consumption. Purposive sampling, model-assisted evaluation, and dependence on parsing, retrieval, and semantic judgment limit generalizability.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization and language editing, and with literature discovery and screening. The authors reviewed AI-assisted text and checked citations against the original sources. Model-assisted dataset construction, annotation, and evaluation are described in Sections 4–5. The authors reviewed the AI-assisted work and take responsibility for the final text, scientific claims, code, data, and analyses.

ETHICS STATEMENT

LongNFQA is constructed from publicly accessible professional reports. These documents remain governed by their publishers' copyright and license terms; public availability is not treated as permission for unrestricted redistribution. Document provenance and source links can be distributed independently of the source files, and any redistribution of document content is conditioned on the applicable terms. We do not intentionally collect private personal data or infer sensitive attributes, although public reports may contain names or institutional records already in the public domain. Third-party model APIs process excerpts from these public documents, which introduces provider data-retention and service-jurisdiction considerations.

Question construction, evidence proposal, and rubric-based evaluation are model-assisted. Item-level human review reduces, but does not eliminate, model, annotator, domain, and purposive-sampling biases. Generated answers can remain incomplete or unsupported and must not be used as financial, medical, legal, regulatory, or policy advice. LongNFQA evaluates research systems within a bounded corpus; it does not certify decision-support reliability or population-level representativeness.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix C specify the method, execution rules, and failure handling. The supplementary attachment provides the executable FactWeaver implementation, frozen configurations, all eleven reported ablations, offline smoke tests, five anonymized execution traces, complete per-question retrieval and answer-quality results, and token accounting summaries. It also includes the 624-question release with human-reviewed gold evidence and the frozen rubric-development and held-out judge-validation artifacts.

The anonymous repository additionally provides processed JSON representations of all 52 documents, document-family manifests, and a source catalogue for the omitted public PDFs, together with the versioned code and evaluation artifacts. The released configuration records the model, decoding, decomposition, retrieval, graph, routing, and shared evaluation settings; validation scripts check artifact schemas, record counts, score aggregation, judge agreement, and anonymity constraints. Sections 4–5 and Appendices A–E document the construction, controls, metrics, and analyses corresponding to these artifacts. Raw source PDFs are not redistributed; their official access URLs and version qualifications are listed in the released source catalogue. API-served models remain subject to provider-side updates.

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/forum?id=hSyW5go0v8.

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1196–1207, 2022. doi: 10.1145/3477495.3531926.

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*, pp. 15159–15177, 2024. doi: 10.18653/v1/2024.emnlp-main.845.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, 2021. doi: 10.18653/v1/2021.emnlp-main.300.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Maojia Song, Chaoqun Liu, Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-LongDoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9233–9250, 2025. doi: 10.18653/v1/2025.emnlp-main.469.

Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. MMDocIR: Benchmarking multimodal retrieval for long documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30971–31005, 2025. doi: 10.18653/v1/2025.emnlp-main.1576.

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.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567, 2019. doi: 10.18653/v1/P19-1346.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398.

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*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. MDocAgent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*, 2025. URL https://arxiv.org/abs/2503.13964.

Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. ANTIQUE: A non-factoid question answering benchmark. In *Advances in Information Retrieval*, pp. 166–173, 2020. doi: 10.1007/978-3-030-45442-5_21.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, 2024. doi: 10.18653/v1/2024.acl-long.745.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, 2021. doi: 10.18653/v1/2021.eacl-main.74.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O. Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Rwhi91ideu.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Snp3iEj7NJ.

Jia-Huei Ju, François G. Landry, Eugene Yang, Suzan Verberne, and Andrew Yates. LANCER: LLM reranking for nugget coverage. In *Advances in Information Retrieval*, pp. 188–203, 2026. doi: 10.1007/978-3-032-21300-6_12.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8euJaTveKw.

Alex Laitenberger, Christopher D. Manning, and Nelson F. Liu. Stronger baselines for retrieval-augmented generation with long-context language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32559–32569, 2025. doi: 10.18653/v1/2025.emnlp-main.1656.

DongGeon Lee, Ahjeong Park, Hyeri Lee, Hyeonseo Nam, and Yunho Maeng. Typed-RAG: Type-aware decomposition of non-factoid questions for retrieval-augmented generation. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pp. 129–152, 2025. doi: 10.18653/v1/2025.xllm-1.14.

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*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. OpenRubrics: Towards scalable synthetic rubric generation for reward modeling and LLM alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17417–17437, 2026. doi: 10.18653/v1/2026.acl-long.791.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang 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*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741.

Thang Nguyen, Peter Chin, and Yu-Wing Tai. MA-RAG: Multi-agent retrieval-augmented generation via collaborative chain-of-thought reasoning. *arXiv preprint arXiv:2505.20096*, 2025. URL https://arxiv.org/abs/2505.20096.

Corbin Rosset, Ho-Lam Chung, Guanghui Qin, Ethan C. Chau, Zhuo Feng, Ahmed Awadallah, Jennifer Neville, and Nikhil Rao. Researchy questions: A dataset of multi-perspective, decompositional questions for deep research. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3712–3722, 2025. doi: 10.1145/3726302.3730275.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, 2024. doi: 10.18653/v1/2024.naacl-long.20.

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/forum?id=GN921JHCRw.

Amir Soleimani, Christof Monz, and Marcel Worring. NLQuAD: A non-factoid long question answering data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1245–1255, 2021. doi: 10.18653/v1/2021.eacl-main.106.

Radu Soricut and Eric Brill. Automatic question answering: Beyond the factoid. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 57–64. Association for Computational Linguistics, 2004. URL https://aclanthology.org/N04-1008/.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37(2):351–383, 2011. doi: 10.1162/COLI_a_00051.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical Multimodal Transformers for Multipage DocVQA. *Pattern Recognition*, 144:109834, 2023. doi: 10.1016/j.patcog.2023.109834.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. Foundational autoraters: Taming large language models for better automatic evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17086–17105, 2024. doi: 10.18653/v1/2024.emnlp-main.949.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. SQuALITY: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1139–1156, 2022. doi: 10.18653/v1/2022.emnlp-main.75.

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. doi: 10.18653/v1/2025.emnlp-main.317.

Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. Chain-of-retrieval augmented generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 66673–66700, 2025. doi: 10.52202/085713-2003.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3225–3245, 2023. doi: 10.18653/v1/2023.acl-long.181.

Sihui Yang, Keping Bi, Wanqing Cui, Jiafeng Guo, and Xueqi Cheng. LINKAGE: Listwise ranking among varied-quality references for non-factoid QA evaluation via LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6985–7000, 2024. doi: 10.18653/v1/2024.findings-emnlp.410.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3277–3287, 2021. doi: 10.18653/v1/2021.acl-long.254.

APPENDIX CONTENTS

Appendix	Contents	Page
A	Answer-Quality Rubric and Judge Validation	14
B	LongNFQA Construction and Evidence Annotation	17
C	Method and Implementation Details	19
D	Baselines, Ablations, and Experimental Controls	22
E	Additional Results and Analyses	24
F	Qualitative Examples and Error Analysis	34

A ANSWER-QUALITY RUBRIC AND JUDGE VALIDATION

This appendix specifies the answer-quality protocol summarized in Section 5.1. The complete LongNFQA evaluation comprises 13,104 answers from 21 methods on 624 questions. Judge development and held-out validation were conducted on the original 15-system set (9,360 answers); the six supplemental mechanism ablations were evaluated by automatic-evaluation after the protocol was frozen. Documents are assigned as indivisible groups to a rubric-development split (84 questions from seven documents), a held-out human-validation split (84 questions from seven documents), and an automatic-evaluation split (456 questions from 38 documents). The three splits share neither questions nor documents, and each covers all five document families and six question types.

A.1 HUMAN REVIEW, BLINDING, AND AGGREGATION

For every question in the development and human-validation splits, two reviewers independently scored all 15 original-system answers. Method names were replaced by deterministic blind identifiers and answer order was randomized. Reviewers received the question, its human-reviewed evidence, and the required perspectives, but no reference answer. A third reviewer adjudicated disagreements. Development rationales were used to specify the rubric below; the rubric, judge instructions, and aggregation rule were fixed before the held-out split was evaluated. Human-validation annotations were used only to measure agreement and did not revise the judge.

Each answer receives an integer score from 1 to 5 on six dimensions. Avg. is the unweighted arithmetic mean of these six scores; it is not a separate holistic judgment. The two human-reviewed splits use adjudicated scores in the reported aggregate, while the remaining 456 questions use the frozen judge.

A.2 COMPLETE SCORING RUBRIC

Factual and Numerical Correctness (FNC). Whether factual statements, quantities, dates, units, comparisons, and attributions agree with the reviewed evidence.

1. Multiple central facts, values, units, dates, comparisons, or attributions contradict the evidence.
2. At least one major factual or numerical error materially changes the answer; correct details do not repair it.
3. The central direction is plausible, but one material detail is wrong, imprecise, or insufficiently checked.
4. Substantive facts and numbers agree with the evidence, with only a minor imprecision.
5. All checkable facts, numbers, units, dates, comparisons, and attributions are accurate and appropriately precise.

Evidence Grounding (EG). Whether substantive claims are supported by, and traceable to, the reviewed document evidence.

1. Central claims are unsupported by or contrary to the evidence.
2. Some evidence is used, but major conclusions lack support or cannot be traced to the reviewed evidence.

3. Most central claims have support, but grounding is incomplete, indirect, or uneven across important perspectives.
4. Nearly all substantive claims are well supported, with at most one minor traceability gap.
5. Every substantive claim is directly supported or transparently derived from the reviewed evidence.

Overclaim Control (OC). Whether the answer respects evidential boundaries, qualifies uncertainty, and avoids unsupported extrapolation.

1. The answer repeatedly presents unsupported conclusions, forecasts, causes, or external assumptions as established facts.
2. A major conclusion crosses the evidence boundary or omits a necessary uncertainty qualification.
3. The answer is generally cautious but contains one meaningful overstatement or ambiguous qualification.
4. Evidence boundaries and uncertainty are handled well, with only a minor wording risk.
5. The answer consistently distinguishes disclosed facts, warranted inference, uncertainty, and unavailable information.

Completeness (CMP). Whether the answer covers the required analytical perspectives and their important supporting details.

1. Most required perspectives are absent or the answer does not address the analytical task.
2. The answer covers a minority of required perspectives or omits at least one central part of the requested synthesis.
3. Central perspectives are partly covered, but an important perspective or supporting relation remains missing or shallow.
4. All central perspectives are covered, with only a minor supporting detail or qualification omitted.
5. All required perspectives and their important relations, qualifications, and supporting details are addressed.

Reasoning Soundness (RS). Whether comparisons, causal links, calculations, and cross-evidence synthesis are logically warranted.

1. Comparisons, causal links, calculations, or synthesis are incoherent or contradicted by the evidence.
2. A major reasoning step is invalid, unexplained, or connects evidence that does not support the conclusion.
3. The overall reasoning is plausible, but one material link, calculation, or synthesis step is weak or implicit.
4. Reasoning is coherent and evidence-aware, with only a minor gap in explanation or integration.
5. Comparisons, calculations, causal qualifications, and cross-evidence synthesis are explicit, coherent, and warranted.

Focus and Clarity (FC). Whether the answer is relevant, well organized, readable, and proportionate without harmful repetition or digression.

1. The response is largely irrelevant, disorganized, unreadable, or dominated by repetition and digression.
2. The requested conclusion is difficult to recover because of substantial verbosity, poor organization, or irrelevant content.
3. The answer is understandable but noticeably repetitive, diffuse, unevenly organized, or insufficiently direct.

4. The answer is clear, relevant, and well organized, with only minor verbosity or presentation issues.
5. The answer is direct, proportionate, coherent, easy to navigate, and free of harmful repetition or digression.

A.3 Frozen Judge and Held-Out Validation

The judge uses Qwen3.5-Flash with temperature zero. It evaluates each anonymous answer independently against the question, reviewed evidence, evaluation perspectives obtained after source-level annotation, and the dimension-specific scoring descriptions. For each dimension it returns an integer score and an answer-specific justification; auxiliary fields record missing perspectives, unsupported claims, factual or numerical errors, boundary violations, and a concise error summary. The prompt explicitly prohibits inferring method identity, rewarding unsupported external knowledge, or treating answer length and formatting as evidence of quality.

Table 5 compares frozen-judge scores with adjudicated human scores on 1,260 held-out answers. Dimension-level statistics pool aligned answer instances. Mean rank correlation is computed over the 15 answers within each question and then averaged over the 84 questions.

Table 5: Agreement between the frozen answer-quality judge and adjudicated human scores on the held-out validation split.

Dimension	Spearman ρ	MAE	Exact agreement
FNC	0.942	0.088	0.912
EG	0.905	0.121	0.879
OC	0.924	0.106	0.894
CMP	0.918	0.116	0.884
RS	0.903	0.126	0.874
FC	0.715	0.241	0.759
Mean within-question ranking	0.890	–	–

Agreement is strongest for Factual and Numerical Correctness and Overclaim Control and lower for the more subjective Focus and Clarity criterion. The predefined aggregate retains all six dimensions. As auxiliary bias checks, the correlations of method-centered scores with presentation position and answer length are 0.048 and 0.065, respectively. The evaluation artifacts preserve split assignments and hashes, blind-ID maps, both reviewers' records, adjudication records, the versioned rubric and prompt, raw dimension scores, and aggregate outputs.

A.4 Judge Inputs, Outputs, and Frozen Specification

Each request groups the 15 anonymous candidate answers for one question with its reviewed evidence and post-annotation evaluation perspectives. Independent scoring means that each candidate receives its own six criterion scores; candidates are presented in the same question-level request. The saved input contains the question ID and text, question type, document ID and family, evaluation split, corpus scope, and a `gold_evidence` object containing the frozen source-level evidence and its answer-requirement groups $\mathcal{P}_q^{\text{gold}}$. Each candidate record contains `answer_id`, `final_answer`, `answer_characters`, and `selected_node_count`. The blind-ID mapping is kept separately; neither method names nor human scores appear in the candidate records.

The requested output assigns every supplied answer ID a mapping from each dimension ID to an integer `score` and a `reason`, together with missing perspectives, unsupported claims, factual or numerical errors, error codes, and a short justification. Per-question records are flattened into dimension-level rows keyed by question ID, split, answer ID, and dimension. The blind mapping restores method identities only for aggregation. The unweighted mean of the six dimension scores gives each answer's Avg. score; method-level scores average the answer-level values over questions. Human-adjudicated scores replace judge scores on the two reviewed splits, as specified above. The judge and generator share a model family, so the held-out human comparison provides the independent reference for assessing evaluation agreement.

Frozen judge specification. Inputs: the question and its scope, human-reviewed source-level evidence with evaluation perspectives, and 15 anonymized candidate answers. **Scoring instruction:** assess every answer independently on FNC, EG, OC, CMP, RS, and FC using the fixed 1–5 descriptions; do not infer method identity, introduce external knowledge, or treat answer length or formatting as evidence. **Outputs:** for each answer and dimension, return an integer score and justification, together with missing perspectives, unsupported claims, factual or numerical errors, error codes, and a concise summary. **Aggregation:** compute Avg. as the unweighted mean of the six dimension scores and restore blind method identities only after scoring. The complete machine-readable frozen prompt is preserved with the evaluation artifacts; task-specific records instantiate the variable inputs above.

B LongNFQA Construction and Evidence Annotation

This appendix expands the construction protocol in Section 4. LongNFQA is designed for controlled evaluation of long-document non-factoid retrieval and answer synthesis; its purposive sampling supports coverage across document conditions rather than an estimate of their population frequency.

B.1 Positioning against Existing Benchmarks

Table 6 makes explicit the evaluation boundaries summarized in Section 4. The table characterizes each resource relative to the present setting; it does not rank their suitability for their original tasks.

Table 6: Benchmark positioning for the LongNFQA setting. Checkmarks denote whether a resource is designed for the corresponding long-document or non-factoid QA condition.

Benchmark	Reference	Evaluation boundary in the present setting	Long document	Non-factoid QA
FinQA	(Chen et al., 2021)	Targets numerical or factoid answers rather than open-ended analytical synthesis with reviewed source-node evidence.	–	–
TAT-QA	(Zhu et al., 2021)	Evaluates short table–text answers, not acquisition and synthesis for broad non-factoid questions.	–	–
MP-DocVQA	(Tito et al., 2023)	Focuses on page-localized answer recovery rather than dispersed-evidence synthesis.	✓	–
M-LongDoc	(Chia et al., 2025)	Provides open-ended QA over super-long documents, but each question is grounded in a single evidence page rather than dispersed, multi-perspective evidence.	✓	✓
MMDocIR	(Dong et al., 2025)	Evaluates page- and layout-level retrieval over long multi-modal documents, without evaluating open-ended answer synthesis.	✓	–
ANTIQUE	(Hashemi et al., 2020)	Does not provide a bounded source document or source-level evidence annotations.	–	✓
ELI5	(Fan et al., 2019)	Does not align answer claims with reviewed source-level document evidence.	–	✓
NLQuAD	(Soleimani et al., 2021)	Uses contiguous subsection answers and therefore does not test dispersed-evidence acquisition.	–	✓
SQuALITY	(Wang et al., 2022)	Frames questions as narrative summarization and provides no retrieval annotations.	✓	–
Researchy	(Rosset et al., 2025)	Provides non-factoid, multi-perspective web queries and clicked URLs, but no bounded long-document corpus or reviewed source-node evidence.	–	✓
LongNFQA	*This work*	Provides reviewed source-node evidence and open-ended answer-quality assessment, enabling the two evaluations to be reported separately.	✓	✓

B.2 Corpus Composition and Question Sampling

LongNFQA contains 52 English PDF reports organized into 32 sampling and split units: 20 temporally or functionally linked document pairs and 12 singletons. Questions are answered from one scoped document; document pairing keeps related reports together during sampling and split construction. The reports contain 6,048 pages in total and range from 51 to 201 pages, with a mean of 116.3 pages.

Table 7: LongNFQA composition by document family.

Document family	PDFs	Questions
Financial/business	22	264
Policy/public-sector	8	96
Medical/public-health	8	96
Scientific/technical	8	96
Regulatory/safety	6	72
Total	52	624

PDF parsing retains page, section, document order, and source identifiers for text blocks, complete tables, and chart records. The parser produces 87,447 raw elements (84,498 text, 2,653 table, and 296 chart elements). Removing empty or very short text elements and serializing complete tables and charts yields the canonical evaluation catalogue of 21,646 nodes: 20,371 text, 1,191 table, and 84 chart nodes. Gold annotations and retrieval outputs are resolved to this catalogue.

Question construction follows three complementary routes. *Need-first* construction starts from an analyst's information need and identifies the document evidence needed to answer it. *Evidence-first* construction begins with dispersed, complementary records and formulates a synthesis question that requires their joint interpretation. *Structure-first* construction targets relations across sections, periods, or evidence forms. GPT-5.5 proposes candidate questions and metadata; reviewers then verify the scope, evidence dependence, and non-factoid character of every retained item. The NF-CATS classifier provides an initial taxonomy label, which is accepted or corrected by review. Candidates are excluded when they are factoid, ambiguous, answerable from a single local span, or dependent on external knowledge.

The final collection is balanced across the six question types: evidence-based, comparison, experience, reason, instruction, and debate each contain 104 questions. A deterministic, document-level allocation produces the 84/84/456 development, human-validation, and automatic-evaluation splits. Documents are indivisible, and post-allocation checks confirm zero question or document overlap between splits.

B.3 Source-Level Gold Evidence and Coverage Groups

Each question has one final question-level gold evidence set. Question-construction provenance supplies the initial candidate nodes, but these nodes are not treated as complete gold labels because relevant support may occur elsewhere in the document. GPT-5.5 proposes additional candidate nodes, after which annotators inspect the source content and provenance and make item-level keep or remove decisions. They retain nodes that directly support an answer claim or a necessary qualification and reject material that is only topically related.

After these node-level decisions are fixed, the retained evidence is organized by distinct answer requirements for coverage analysis. Each group contains a short analytical question, its answer-relevant nodes, and one required evidence set E_p^{req} whose recovery marks that requirement as satisfied. Section 5.1 calls these groups evaluation perspectives. They are constructed from the question and reviewed evidence, without access to method outputs. Node-level metrics evaluate the full frozen gold set, while Perspective Satisfaction evaluates recovery across these post-annotation groups.

Every record stores the question and document identifiers, document family, question type, corpus scope, selected canonical node IDs, coverage-group assignments, required evidence, and a provenance-aware selection rationale. Versioned review decisions retain changes to node selection and subsequent group membership. The final 624 records contain 2,018 coverage groups (3.234 per question) and 5,544 answer-relevant node assignments (8.885 per question), of which 4,449 are text and 1,095 are table nodes. The question-level gold sets contain between 7 and 11 nodes. Of the 624 questions, 476 have three required groups, 147 have four, and one has two.

B.4 Leakage and Integrity Checks

Gold evidence is produced over the canonical catalogue independently of the method being scored. Gold node IDs and coverage-group labels are not provided to retrieval or generation systems. Evaluation checks enforce unique question and node identifiers, corpus-scope validity, nonempty group assignments, required-evidence membership in the corresponding gold set, and complete selection rationales. Question generation, evidence annotation, method execution, and answer judging retain separate versioned artifacts and content hashes.

C Method and Implementation Details

This appendix details atomic fact graph construction and graph-guided synthesis in Section 3, including their execution and validation rules. All final FactWeaver runs use Qwen3.5-Flash for semantic mapping, query construction, reranking, and collaborative generation; the stored run manifests record the model used by each stage.

C.1 Fact Identification and Evidence Retrieval

PDFs are converted to page-aligned JSON records containing text blocks, complete tables, chart records, bounding boxes, and cross-page links. Major section boundaries define document chunks for perspective induction, while each retained record preserves its original page and source ID. Local analytical concepts are induced per chunk and consolidated only when their definitions express the same analytical role; provenance spans from all merged concepts are retained.

At query time, the stored question type and selected document-supported perspectives condition atomic-fact extraction. Chart records are normalized from extracted chart data into structured records analogous to serialized table rows, after which they share the same atomic-fact extraction and retrieval pipeline. Narrative propositions, table rows, and normalized chart records are converted into factoid retrieval queries while retaining their perspective and seed provenance. Retrieval uses `all-mpnet-base-v2` embeddings over the scoped canonical nodes. Candidates are reranked using the original question, the associated perspective, and the factoid query; the evaluation wrapper exposes the final ordered canonical node IDs and their full node records. The shared experimental protocol admits at most 25 nodes and scores the first 20.

C.2 Atomic Fact Graph Construction

Retrieved evidence informs semantic relation induction; deterministic ID validation and edge construction then produce the graph. Table 8 specifies this boundary.

Table 8: Construction of the four graph relations.

Relation	Construction rule	Execution type
Hierarchy H	Connect the answer objective to selected perspectives and each perspective to atomic-fact nodes admitted by its perspective assignment.	Deterministic from stored provenance.
Sequence E_{seq}	Connect consecutive valid atomic-fact nodes in the perspective-local evidence plan.	Inherited local plan order; deterministic adjacency and ID checks.
Support E_{sup}	Assign question-grounded domain labels and propose cross-perspective links only for matching non-generic labels.	Semantic labels; deterministic pair validation.
Conflict E_{conf}	Identify potentially incompatible numerical, causal, temporal, or scope assertions within a perspective; connect units implicated in the same conflict.	Semantic conflict detection; deterministic ID and pair validation.

Local causal anchoring provides candidate claims and explanations, entity alignment assigns question-grounded domains, narrative analysis characterizes discourse relations, and conflict auditing records potential tensions. These annotations are retained with each graph leaf's atomic fact and retrieved source nodes. Local evidence plans rank atomic facts using evidence form, explanatory content, and question alignment; sequence edges inherit this initial order, not a separately inferred

causal or temporal ordering. Consequently, a path-level claim can be traced through the graph leaf to the page-aligned source node, even though relation labels themselves remain proposals subject to later audit.

C.3 Collaborative Synthesis Protocol

Table 9 summarizes the structural and semantic checks used in path assessment.

Table 9: Path auditing separates structural checks from semantic judgment.

Audit target	Assessment
Structure and coverage	Deterministic checks flag empty paths, inadmissible units, and missing perspectives.
Claim support	Semantic review distinguishes fully supported, partially supported, and unsupported claims.
Conflict relations	Deterministic checks locate exposed conflict edges; semantic review interprets their implications.
Evidence order	Semantic review assesses coherence with reference to sequence relations.

The implementation realizes the state transition in Equation 5 through structured role outputs and a routing controller. The fixed chain first proposes candidate reasoning paths, audits their evidence use, applies accepted revisions, realizes the selected path as prose, and verifies the resulting answer. Conditional review is activated from the graph and question signals: detected conflicts or question complexity may invoke mediation or domain criticism, while compound high-risk signals may invoke an additional audit. Conditional roles may propose constraints or diagnoses, but they do not write the answer directly.

The controller validates every referenced graph ID, recomputes structural coverage and relation exposure, and commits accepted constraints and open issues to the shared state. Semantic outputs formulate claims, interpret conflicts, and propose revisions. The Reviser applies the permitted operations in Equation 6; revised paths undergo the same structural checks before selection. The Writer receives the selected path, accepted constraints, and provenance-bearing evidence. The Verifier checks the draft against these objects and either returns it or supplies a corrected answer.

Table 10: Deterministic and model-mediated responsibilities.

Mechanism	Operations
Deterministic	Node-ID and corpus-scope validation; path non-emptiness; perspective coverage; graph-edge exposure; state serialization; provenance resolution; output-schema checks.
Model-mediated	Perspective mapping; atomic fact and query formulation; semantic reranking; local causal and narrative analysis, domain alignment, and conflict detection; path and claim construction; support qualification; conflict interpretation; revision, answer realization, and semantic verification.

C.4 Routing, Selection, and Termination

Routing rules. Routing is computed once from the question and the graph before path proposal. Graph coverage is the fraction of perspectives in the synthesis brief represented by at least one atomic-fact node; it uses no gold annotations. The implementation's low-coverage threshold is 0.75, and its additional-audit conflict threshold is two edges. These are configurable defaults, not parameters estimated from the evaluation scores. The local routing tag is assigned by an ordered keyword check for decision/debate, comparison, causal, evidence-support, and guidance intent, with a general fallback. The first four tags count as complex. This operational tag is distinct from the six-category question taxonomy used for factoid query construction. Mediation is activated by any conflict, low coverage, or complex intent; an Extra Auditor is activated by at least two conflicts or the conjunction of low coverage and complex intent; the Domain Critic is activated by complex intent or the absence of cross-perspective support edges.

Candidate and state updates. The Planner is instructed to propose multiple paths, without a fixed candidate count. Path records contain an ID, an ordered list of atomic-fact node IDs, claims, and qualification notes; the additional named fields in the JSON record retain the thesis, assumptions, links, and revision rationale. Missing or duplicate path IDs are normalized before auditing. The controller records invalid references, missing perspectives, and support or conflict edges whose two endpoints occur in the path. Its structural status is `needs_revision` for an empty path or an invalid reference; `usable` alone does not certify semantic support or complete coverage.

Auditor feedback supplies path-specific diagnoses and global constraints. Additional critiques are consolidated before mediation. Accepted constraints are merged with the existing constraints and deduplicated by their serialized content; mediation records the remaining open issues and the revision plan. The Reviser receives the candidates, graph, audit findings, mediation output, and shared state in a single request, allowing multiple defects to be addressed together. Its revised paths are structurally audited again. The primitive names in Equation 6 describe the permitted semantic changes; the model returns revised structured paths rather than invoking separate executable functions for each primitive.

Path selection. Selection first considers revised paths not marked rejected; if this leaves no candidates, the implementation retains the returned set for best-effort selection. The provisional path ID is taken from the Reviser, otherwise the Mediator, otherwise the Planner. It is accepted when it identifies an eligible candidate with no invalid node references and no rejecting audit. Otherwise, candidates are ranked by the implementation's feedback heuristic, with ties resolved by candidate order.

For reproducibility, this heuristic sums the model-reported coverage, coherence, and faithfulness fields, adds the stored structural-audit score, and adds 0.6 times the audit's evidence-support field and 0.4 times its conflict-resolution field. It subtracts 5 for a rejecting audit, 1 for a revision request, and 1.5 per invalid node. The stored structural score itself contains the same three model fields, plus 0.5 per covered perspective and 0.25 per exposed support edge, minus 0.75 per invalid node and 0.15 per missing perspective, rounded to three decimals. Thus the three model fields enter the fallback ranking twice. Missing or nonnumeric scores become zero. The semantic fields are prompted model judgments without calibrated scales; the coefficients are fixed implementation heuristics, not learned weights or formal measures of coverage, coherence, or entailment. They are distinct from Perspective Satisfaction and the answer-quality rubric.

Finite execution and failure handling. Each enabled role is called once in a finite synthesis pass: Planner, Auditor, optional Extra Auditor and Domain Critic, optional Mediator, Reviser, Writer, and Verifier. There is one revision pass and one final verification pass, with five to eight role calls; termination does not depend on all defects being resolved. When mediation is disabled, the controller consolidates feedback directly. Candidate and revised path lists fall back to a graph-ranked path when empty. That path selects one highest-ranked unit per represented perspective, using the graph's stored selection scores; these combine the local evidence-plan score, 0.6 times entity-alignment confidence, and 0.15 times the number of candidate causes. Without perspective labels, the fallback selects up to five highest-ranked units.

Each role requests a JSON object. Parsing first attempts the complete response, then its outermost brace-delimited object. A failed call or parse returns the role's fallback record with an error field: revision preserves the input paths, and verification preserves the Writer's answer. The final output uses a nonempty verifier correction, otherwise the Writer's answer, and finally an evidence-based fallback assembled from the local briefs if both are empty. Fallback outputs preserve traceability but do not establish that semantic verification succeeded.

Each run preserves the question and model configuration, selected perspectives and atomic facts, retrieved and reranked nodes, graph nodes and edges, routing decisions, candidate and revised paths, shared state, final answer, and stage-level token and timing records. Error fields and structural-audit findings distinguish normal outputs from fallback or imperfect paths. Private credentials and service endpoints are excluded from the reproducibility materials.

D BASELINES, ABLATIONS, AND EXPERIMENTAL CONTROLS

D.1 BASELINE IMPLEMENTATIONS

The baselines cover complementary strategies for evidence acquisition and answer synthesis: document-order context organization, hierarchical retrieval, graph propagation, adaptive retrieval, retrieval chains, and type-aware query decomposition. A shared source-node interface and answer generator support evaluation on LongNFQA while the implementations retain the distinct operations specified in Table 11. The table identifies the protocol adaptations used in the experiments; method names in the results refer to these evaluated implementations.

Table 11: Baseline implementations and evaluation coverage. † denotes an adaptation to the evaluation protocol.

Method	Evaluated implementation	Retrieval	Answer
Direct LLM	No-document generation control.	–	✓
Naive RAG	Single dense retrieve–rerank–generate pipeline over canonical source nodes.	✓	✓
DOS-RAG†	Retrieves from the shared candidate catalogue and restores the selected evidence to document order before generation.	✓	✓
RAPTOR†	Resource-bounded collapsed summary tree with document-order clustering, two extractive levels, TF–IDF retrieval, and backtracking to leaf source nodes.	✓	✓
Self-RAG†	Same-generator adaptation with retrieval-decision, relevance, support, and one additional-retrieval check.	✓	✓
HippoRAG 2†	Local personalized PageRank over OCR links and shared phrase nodes, without OpenIE; graph results map to canonical source nodes.	✓	✓
CoRAG†	Same-generator retrieval-chain adaptation that alternates query proposal, retrieval, and intermediate evidence assessment.	✓	✓
Search-R1†	Iterative external-search protocol without the original RL training; it emits no local source-node IDs.	–	✓
Typed-RAG†	Type-aware question-decomposition adapter under the shared retriever, reranker, and answer generator.	✓	✓

Type-aware decomposition. The Typed-RAG implementation retains type-conditioned subquery construction, per-subquery retrieval, and evidence aggregation. A unified structured prompt proposes up to four question-only subqueries by default based on the given question ; this prompt and the shared answer generator replace the original type-specific processing branches. Neither target-document content nor gold evidence or coverage-group labels enter the decomposition prompt. If no subquery is returned, retrieval uses the original question.

Comparison scope. The shared generator executes each baseline's retrieval and synthesis protocol, controlling model identity while allowing method-specific evidence selection and intermediate reasoning. Self-RAG, CoRAG, and Search-R1 use prompted decision protocols with this generator rather than their original specialized training; RAPTOR and HippoRAG 2 use the indexing configurations specified in the table. The results therefore compare these strategies under the stated implementations and controls, without assuming equivalence to their original trained systems. Search-R1 additionally provides an external-search comparison for answer quality, with a different evidence scope from the local-document retrievers.

D.2 ABLATION DEFINITIONS

The eleven ablations follow the evidence-acquisition, relational-organization, and graph-guided-synthesis stages used in the main text. Table 12 details the first two stages, and Table 13 details graph-guided synthesis. The descriptions identify the retained components and interpretation boundaries of each intervention.

Table 12: Evidence-acquisition and relational-organization ablations. Components not named in an intervention are retained.

Variant	Intervention
Evidence acquisition	
w/o Atomic-Fact Acquisition	Use the original question to retrieve raw document nodes for graph construction, without perspective selection, atomic-fact queries, or reranking; retain downstream relation construction and synthesis.
w/o Perspective Selection	Bypass question-relevant perspective selection and acquire facts from the document inventory under the matched fact, query, and evidence budgets; retain perspective identities and all downstream stages.
w/o Atomic-Fact Queries	Retain the selected perspectives and atomic facts, but replace fact-conditioned queries with perspective-conditioned queries under the matched acquisition budget; rerun all downstream stages.
Relational organization	
w/o Graph Representation	Retain the same upstream evidence and agent schedule, but supply independent perspective paragraphs without hierarchy, sequence, support, or conflict edges.
w/o Relational Edges	Retain the same atomic facts, evidence, perspective membership, hierarchy, and agent schedule, but remove sequence, support, and conflict edges before synthesis.
w/o Retrieved Context for Relations	Induce relations from each fact's seed provenance rather than its supplementary retrieved context; retain the same final evidence for downstream synthesis.

Table 13: Graph-guided-synthesis ablations. Components not named in an intervention are retained.

Variant	Intervention
w/o Collaborative Synthesis	Retain retrieved evidence and the atomic fact graph, but replace the graph-guided multi-agent synthesis protocol with one-pass answer synthesis.
w/o Path Construction	Remove explicit candidate-path construction and the fixed assessment and revision stages that depend on it; retain conditional roles, Writer, and Verifier with only the structural interface needed to execute them.
w/o Conditional Roles	Disable Extra Auditor, Domain Critic, and Mediator while retaining Planner, Auditor, Reviser, Writer, and Verifier.
w/o Path Assessment and Revision	Retain Planner candidate paths, but omit auditing, mediation, and revision; select an initial valid path by a fixed rule before the unchanged Writer and Verifier.
w/o Grounded Answer Realization	Retain the final revised path and its evidence, but replace constraint-aware Writer–Verifier realization with a single path-to-prose generation call.

w/o Atomic-Fact Acquisition jointly changes decomposition, query construction, and reranking and therefore does not isolate atomic-fact granularity. Likewise, w/o Graph Representation changes the representation from atomic-fact graph to perspective paragraphs, whereas w/o Relational Edges preserves the nodes and hierarchy and isolates the explicit relational edges. w/o Path Construction removes a dependency chain, while w/o Path Assessment and Revision retains the initial paths and isolates their subsequent assessment and revision. The resulting stage substitutions and localized controls are not additive; these scope differences govern the comparisons drawn in Section 6.3.

D.3 Shared Controls and Cost Accounting

Table 14 records the configuration values that materially determine decomposition breadth, evidence acquisition, graph sparsity, routing, and matched evaluation. We report these decision-relevant settings rather than an exhaustive list of serialization, retry, and logging options. Method-specific acquisition budgets apply inside FactWeaver; the shared context controls apply to the cross-method evaluation harness.

Table 14: Key method hyperparameters and shared evaluation controls.

Scope	Setting	Value
Document induction	Number of document Perspectives	5–12
Question mapping	Selected Perspectives per question	3–5
Atomic-fact filtering	Narrative / table preselection per Perspective	20 / 7
Atomic-fact reranking	Retained narrative / table facts per Perspective	10 / 5
Supplemental retrieval	Candidates / retained nodes per atomic-fact query	10 / 3
Evidence graph	Selected atomic-fact nodes per Perspective	5
Dynamic routing	Low-coverage / Extra-Auditor conflict thresholds	0.75 / 2
Decoding	Temperature / final-answer output limit	0 / 2,000 tokens
Shared evaluation	Candidate / final-node limits	60 / 25
Shared evaluation	Context target / retrieval evaluation depth	8,000 tokens / 20 nodes

Where the method permits, comparisons use the same canonical source nodes, question-specific corpus scope, Qwen3.5-Flash answer generator, temperature zero, 2,000-token final-answer limit, and citation instructions. Shared evidence selection uses a 60-candidate limit, at most 25 final context

nodes, and an 8,000-token target. The selector can exceed this target to retain at least 20 nodes when sufficient candidates exist; it is not a strict bound on total context or online inference. Retrieval is evaluated at the first 20 positions. These values serve distinct protocol roles: 60 bounds the candidate pool before final selection, 25 caps the selected context, and 20 fixes the evidence count used for cross-method retrieval comparison. They are fixed configuration choices; the reported experiments do not establish their optimality or sensitivity to alternative limits. We retain and report each method's defining operations, including RAPTOR's tree, HippoRAG 2's graph, CoRAG's retrieval chain, and Search-R1's external search, rather than forcing them into an identical internal pipeline. Intermediate calls and their cumulative token use therefore vary across methods.

Token accounting includes prompt and completion tokens from every online LLM call. For stage-specific ablations, the reported total combines the retained upstream stages with the newly executed modified stage, so all values represent end-to-end token consumption. Direct LLM and Search-R1 do not emit canonical local source-node IDs and are excluded from retrieval evaluation; all methods and all eleven ablations enter answer-quality evaluation.

E Additional Results and Analyses

The main text reports the principal aggregate comparisons. This appendix provides the complete retrieval and answer-quality summaries used in those comparisons, followed by document-family, question-type, and evaluation-split breakdowns and a conditional resampling analysis of the finalized answer scores. All retrieval results use the same depth $k = 20$, and every full-set answer-quality row aggregates the same 624 LongNFQA questions. The subgroup analyses are descriptive; no claim of statistical significance is made.

E.1 Complete Retrieval Results

Table 15 restates the aggregate retrieval results with methods as rows so that all four metrics and evaluated retrieval systems are directly comparable. FactWeaver obtains the highest precision, recall, F1, and Perspective Satisfaction. w/o Atomic-Fact Acquisition remains close to Naive RAG, while the remaining structure-aware and iterative baselines generally improve recall and perspective-level sufficiency. F1 is the harmonic mean of the aggregate Precision@20 and Recall@20 values.

Table 15: Complete aggregate retrieval results at $k = 20$. Best and second-best values are bold and underlined, respectively.

Method	Precision@20	Recall@20	F1@20	PS@20
Naive RAG	.1759	.4193	.2478	.2472
w/o AFA	.1787	.4366	.2536	.2529
DOS-RAG	.1756	.4286	.2491	.2560
RAPTOR	.1965	.4801	.2789	.3140
Self-RAG	.2159	.5293	.3067	.3443
HippoRAG 2	.2368	.5809	.3364	.4050
CoRAG	.2582	.6356	.3672	.4661
Typed-RAG	<u>.2759</u>	<u>.6690</u>	<u>.3907</u>	<u>.5458</u>
FactWeaver	**.3358**	**.7590**	**.4656**	**.6842**

Document-family breakdown. Table 16 reports every final subgroup value for the five document families. FactWeaver is highest on all three metrics in each family. Relative to Typed-RAG, its Perspective Satisfaction gain ranges from 0.1216 in regulatory/safety documents to 0.1545 in scientific/technical assessments.

Table 16: Retrieval results by document family at $k = 20$. P, R, and PS denote Precision, Recall, and Perspective Satisfaction; best and second-best values are bold and underlined, respectively.

Family (N)	Metric	Naive RAG	w/o AFA	DOS-RAG	RAPTOR	Self-RAG	HippoRAG 2	CoRAG	Typed-RAG	**FactWeaver**
Financial/business (264)	P@20	.1845	.1862	.1839	.2049	.2242	.2428	.2636	.2824	**.3436**
	R@20	.4275	.4416	.4365	.4879	.5370	.5847	.6444	.6769	**.7655**
	PS@20	.2582	.2617	.2648	.3245	.3532	.4119	.4751	.5562	**.6922**
Policy/public-sector (96)	P@20	.1792	.1802	.1776	.1974	.2182	.2391	.2573	.2771	**.3375**
	R@20	.4232	.4368	.4301	.4786	.5314	.5828	.6360	.6678	**.7593**
	PS@20	.2587	.2622	.2622	.3142	.3550	.4123	.4635	.5503	**.6875**
Medical/public-health (96)	P@20	.1672	.1698	.1677	.1891	.2083	.2318	.2557	.2698	**.3276**
	R@20	.4151	.4318	.4255	.4790	.5274	.5836	.6355	.6663	**.7560**
	PS@20	.2318	.2396	.2431	.3064	.3290	.3993	.4549	.5313	**.6736**
Scientific/technical (96)	P@20	.1641	.1698	.1651	.1844	.2047	.2281	.2516	.2656	**.3271**
	R@20	.4056	.4300	.4175	.4659	.5169	.5728	.6240	.6548	**.7510**
	PS@20	.2352	.2387	.2465	.2960	.3342	.3906	.4583	.5278	**.6823**
Regulatory/safety (72)	P@20	.1674	.1729	.1667	.1903	.2069	.2299	.2514	.2722	**.3278**
	R@20	.4078	.4331	.4165	.4741	.5171	.5719	.6181	.6642	**.7495**
	PS@20	.2280	.2454	.2454	.3090	.3310	.3970	.4618	.5451	**.6667**

Question-type breakdown. Table 17 gives the corresponding results for all six non-factoid question types. The gains are not confined to one analytical intent: FactWeaver leads every method on precision, recall, and Perspective Satisfaction in each type. Its Perspective Satisfaction improvement over Typed-RAG ranges from 0.1306 on debate questions to 0.1482 on experience questions.

Table 17: Retrieval results by question type at $k = 20$. P, R, and PS denote Precision, Recall, and Perspective Satisfaction; best and second-best values are bold and underlined, respectively.

Question type (N)	Metric	Naive RAG	w/o AFA	DOS-RAG	RAPTOR	Self-RAG	HippoRAG 2	CoRAG	Typed-RAG	**FactWeaver**
Evidence-based (104)	P@20	.1841	.1841	.1832	.2048	.2221	.2418	.2601	.2793	**.3447**
	R@20	.4268	.4379	.4346	.4881	.5337	.5859	.6404	.6741	**.7682**
	PS@20	.2596	.2596	.2620	.3237	.3494	.4159	.4800	.5569	**.6939**
Comparison (104)	P@20	.1808	.1841	.1798	.2000	.2207	.2433	.2582	.2774	**.3399**
	R@20	.4268	.4458	.4360	.4852	.5373	.5927	.6355	.6709	**.7645**
	PS@20	.2516	.2660	.2636	.3269	.3518	.4279	.4655	.5497	**.6851**
Experience (104)	P@20	.1726	.1736	.1721	.1933	.2125	.2327	.2587	.2745	**.3327**
	R@20	.4228	.4348	.4305	.4838	.5313	.5823	.6381	.6730	**.7621**
	PS@20	.2428	.2436	.2460	.3045	.3357	.3902	.4551	.5321	**.6803**
Reason (104)	P@20	.1769	.1822	.1769	.1976	.2178	.2375	.2582	.2764	**.3341**
	R@20	.4182	.4420	.4286	.4789	.5302	.5796	.6352	.6680	**.7519**
	PS@20	.2548	.2580	.2668	.3157	.3574	.4095	.4776	.5545	**.6875**
Instruction (104)	P@20	.1635	.1673	.1639	.1832	.2048	.2264	.2543	.2683	**.3250**
	R@20	.4051	.4253	.4155	.4642	.5167	.5644	.6253	.6568	**.7485**
	PS@20	.2348	.2380	.2444	.2949	.3333	.3854	.4535	.5288	**.6747**
Debate (104)	P@20	.1774	.1808	.1774	.2000	.2173	.2389	.2596	.2793	**.3385**
	R@20	.4161	.4338	.4262	.4805	.5262	.5805	.6390	.6713	**.7589**
	PS@20	.2396	.2524	.2532	.3181	.3381	.4014	.4647	.5529	**.6835**

E.2 Complete Answer-Quality and Token-Consumption Results

Table 18 reports every aggregate answer-quality score and mean end-to-end tokens per question for all 21 systems on the same 624 questions. The six criteria use a 1–5 scale, and Avg. is their unweighted mean. FactWeaver is highest on every quality criterion. The ablations follow the three-stage processing order used in the main text; broad substitutions and targeted controls are shown together within their corresponding stage.

Table 18: Complete answer-quality (1–5) and token-consumption results. FNC, EG, OC, CMP, RS, and FC denote Factual and Numerical Correctness, Evidence Grounding, Overclaim Control, Completeness, Reasoning Soundness, and Focus and Clarity, respectively. Avg. is their unweighted mean; Tok. is mean end-to-end online token consumption per question in thousands. Best and second-best values in each column are bold and underlined, respectively.

Method	FNC ↑	EG ↑	OC ↑	CMP ↑	RS ↑	FC ↑	Avg. ↑	Tok. (K) ↓
Baselines								
Direct LLM	2.5385	2.2708	2.7917	2.3173	2.1619	2.7692	2.4749	**2.772**
Naive RAG	2.4920	2.8446	2.7276	2.6603	2.6202	3.0080	2.7254	68.731
DOS-RAG	2.9247	3.2644	3.1731	3.1330	3.0481	3.4151	3.1597	137.846
RAPTOR	2.6955	3.0657	2.9423	2.9167	2.8574	3.2821	2.9599	615.285
Self-RAG	3.3301	3.6971	3.5192	3.4808	3.4006	3.8734	3.5502	598.417
HippoRAG 2	3.3830	3.7484	3.5994	3.5833	3.4856	3.8910	3.6151	696.402
CoRAG	3.7516	4.1042	3.8910	3.8990	3.8237	4.2901	3.9599	704.936
Search-R1	3.2788	3.1442	3.6090	3.1218	3.0817	3.4439	3.2799	<u>8.688</u>
Typed-RAG	3.8798	4.1522	4.0401	4.0673	3.9295	4.3494	4.0697	589.736
Proposed method								
FactWeaver	**4.4631**	**4.7324**	**4.6458**	**4.2853**	**4.6266**	**4.7869**	**4.5900**	675.609
Evidence acquisition								
w/o AFA	3.3301	3.3381	3.5096	3.0449	3.3365	3.3894	3.3248	94.272
w/o PS	<u>4.1891</u>	4.3974	4.3301	4.1074	4.2949	4.4215	4.2901	667.184
w/o AFQ	3.9936	4.3221	4.2131	3.9311	4.1474	4.4119	4.1699	659.743
Relational organization								
w/o GR	3.5080	3.9231	3.6987	3.6010	3.7532	3.9551	3.7399	648.083
w/o RE	3.9647	4.2067	3.9487	3.9519	3.9087	4.1987	4.0299	641.837
w/o RC	4.1474	4.3606	4.2564	4.1026	4.2051	4.4247	4.2495	657.894
Graph-guided synthesis								
w/o CS	3.1202	3.7901	3.3221	3.2676	3.4519	3.7468	3.4498	556.282
w/o PC	3.7436	4.0032	3.8942	3.7196	3.8686	4.0529	3.8803	596.577
w/o COND	4.1747	<u>4.6619</u>	<u>4.3990</u>	<u>4.1362</u>	<u>4.4487</u>	<u>4.6394</u>	<u>4.4100</u>	621.923
w/o PAR	3.8446	4.1122	3.8638	3.9071	3.8429	4.1891	3.9599	574.918
w/o GAR	4.0849	4.3446	4.1571	4.0897	4.3253	4.3766	4.2297	631.527

Complete token accounting. Table 19 lists the token totals for all evaluated methods separately from the answer-quality criteria. Totals sum API-reported input and output tokens across online LLM calls.

Table 19: Mean end-to-end online token consumption per question for evaluated baselines, FactWeaver, and its ablation variants (thousands). Ablation abbreviations follow Section 5.2.

Baseline or full method	Tokens (K)	Ablation	Tokens (K)
Direct LLM	2.772	w/o AFA	94.272
Naive RAG	68.731	w/o PS	667.184
DOS-RAG	137.846	w/o AFQ	659.743
RAPTOR	615.285	w/o GR	648.083
Self-RAG	598.417	w/o RE	641.837
HippoRAG 2	696.402	w/o RC	657.894
CoRAG	704.936	w/o CS	556.282
Search-R1	8.688	w/o PC	596.577
Typed-RAG	589.736	w/o COND	621.923
FactWeaver	675.609	w/o PAR	574.918
		w/o GAR	631.527

Table 20: FactWeaver online token use by document family, question type, and source-PDF length (624 questions). Values are per-question means in thousands.

Stratum	PDFs	Questions	Input (K)	Output (K)	Total (K)
Document family					
Financial/business	22	264	430.036	280.541	710.576
Policy/public-sector	8	96	475.674	275.596	751.270
Medical/public-health	8	96	389.726	278.634	668.360
Scientific/technical	8	96	389.312	271.234	660.547
Regulatory/safety	6	72	278.140	198.121	476.261
Question type					
Evidence-based	52	104	395.128	257.431	652.560
Comparison	52	104	423.740	283.205	706.945
Experience	52	104	381.890	265.088	646.978
Reason	52	104	391.701	248.784	640.485
Instruction	52	104	434.392	288.090	722.482
Debate	52	104	415.532	268.672	684.204
Source-PDF length (pages)					
1–100	23	276	374.408	251.844	626.252
101–150	19	228	404.248	276.858	681.106
> 150	10	120	487.524	291.161	778.685
All questions	**52**	**624**	**407.064**	**268.545**	**675.609**

Table 21: FactWeaver online token use by method phase over all 624 questions. Values are per-question means in thousands.

Method phase	Input (K)	Output (K)	Total (K)	Share (%)
Perspective and atomic-fact acquisition	161.197	171.159	332.356	49.2
Evidence retrieval	0.000	0.000	0.000	0.0
Evidence reranking	11.263	22.741	34.004	5.0
Local analysis and graph-guided synthesis	234.604	74.645	309.249	45.8
All questions, end-to-end	**407.064**	**268.545**	**675.609**	**100.0**

The quality–token pattern is not monotonic across methods: Search-R1 and Direct LLM use substantially fewer tokens but produce lower Avg. scores, whereas CoRAG is both more costly and lower-scoring than FactWeaver. Within FactWeaver, w/o Conditional Roles reduces token use by 8.0% with a 0.1800-point Avg. decrease, while w/o Collaborative Synthesis reduces token use by 17.7% with a 1.1402-point decrease. Removing path assessment and revision reduces token use by 14.9% with a 0.6301-point decrease. These comparisons characterize the observed trade-off rather than equating additional inference with answer quality.

Figure 2 in Section 6.4 visualizes these operating points. Figure 3 resolves the quality differences by criterion; its stage-wise rows describe evaluated removals, not additive component effects.

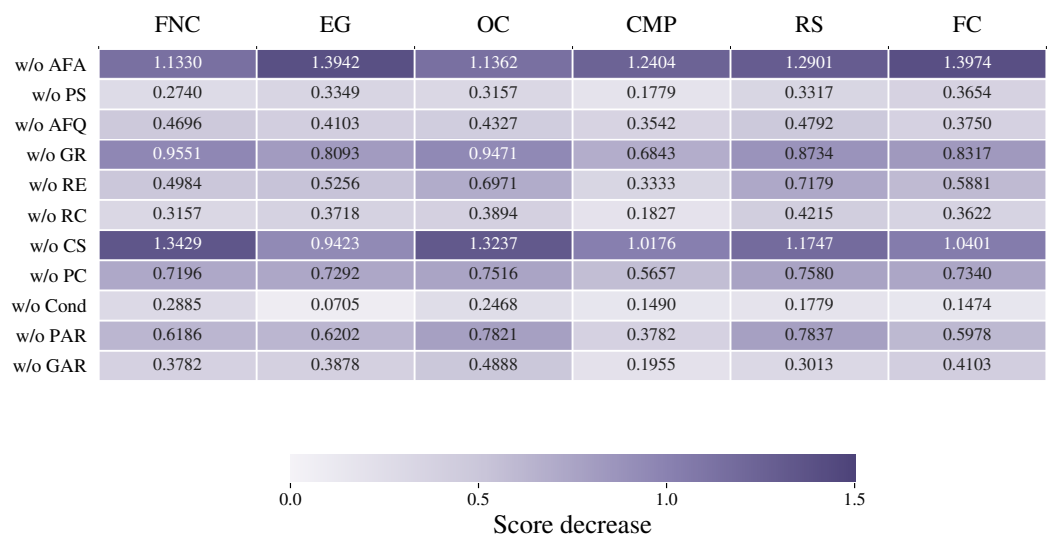

Figure 3: Criterion-level score decreases under 11 ablations. Each cell is the FactWeaver mean minus the ablated mean over the same 624 questions; darker cells indicate larger decreases. Ablation labels follow Section 5.2.

E.3 Answer-Quality Breakdown

Tables 22 and 23 report Avg. scores for all nine baselines, FactWeaver, and the eleven ablations by document family and question type. Each score averages the six criterion scores over the same subset of questions for every method. Weighting the subgroup means by their question counts recovers the aggregate results in Table 18.

Figure 4 summarizes both breakdowns on a shared score scale, showing the cross-category ordering alongside the complete numerical comparisons below.

	Financial/ business (264)	Policy/ public-sector (96)	Medical/ public-health (96)	Scientific/ technical (96)	Regulatory/ safety (72)
Direct LLM	2.4766	2.4896	2.5608	2.4427	2.3773
Naive RAG	2.7374	2.7517	2.7222	2.6944	2.6921
DOS-RAG	3.1616	3.1615	3.1684	3.1267	3.1829
RAPTOR	2.9684	2.9358	2.9740	2.9288	2.9838
Self-RAG	3.5587	3.5365	3.5503	3.4983	3.6065
HippoRAG 2	3.6162	3.6146	3.6076	3.5747	3.6759
CoRAG	3.9735	3.9410	3.9757	3.8924	4.0046
Search-R1	3.2708	3.2951	3.3403	3.2622	3.2361
Typed-RAG	4.0713	4.0694	4.0608	4.0243	4.1366
FactWeaver	**4.5593**	**4.5920**	**4.5590**	**4.6753**	**4.6273**
w/o AFA	3.2917	3.3420	3.3229	3.3733	3.3611
w/o PS	4.2828	4.2917	4.2830	4.3108	4.2963
w/o AFQ	4.1610	4.1736	4.1615	4.1927	4.1782
w/o GR	3.7045	3.7049	3.7396	3.8212	3.8079
w/o RE	4.0208	4.0330	4.0208	4.0538	4.0394
w/o RC	4.2405	4.2517	4.2413	4.2726	4.2593
w/o CS	3.4110	3.4149	3.4497	3.5399	3.5185
w/o PC	3.8479	3.8767	3.8733	3.9497	3.9213
w/o Cond	4.3782	4.3924	4.3941	4.4965	4.4560
w/o PAR	3.9501	3.9635	3.9514	3.9861	3.9676
w/o GAR	4.2216	4.2326	4.2222	4.2517	4.2361

	Evidence-based	Comparison	Experience	Reason	Instruction	Debate
Direct LLM	2.4936	2.4920	2.4487	2.4840	2.4551	2.4760
Naive RAG	2.7292	2.7564	2.7163	2.7356	2.6939	2.7212
DOS-RAG	3.1683	3.1955	3.1474	3.1571	3.1298	3.1603
RAPTOR	2.9663	2.9824	2.9615	2.9760	2.9135	2.9599
Self-RAG	3.5609	3.5769	3.5497	3.5272	3.5304	3.5561
HippoRAG 2	3.6202	3.6587	3.6138	3.6026	3.5929	3.6026
CoRAG	3.9679	3.9936	3.9359	3.9760	3.9279	3.9583
Search-R1	3.2869	3.2997	3.2708	3.2917	3.2516	3.2788
Typed-RAG	4.0657	4.0817	4.0641	4.0817	4.0513	4.0737
FactWeaver	**4.5657**	**4.5897**	**4.5913**	**4.6026**	**4.5897**	**4.6010**
w/o AFA	3.3221	3.3638	3.3365	3.3221	3.2821	3.3221
w/o PS	4.2788	4.2885	4.2901	4.2965	4.2901	4.2965
w/o AFQ	4.1571	4.1683	4.1699	4.1763	4.1699	4.1779
w/o GR	3.7340	3.7340	3.7484	3.7452	3.7372	3.7404
w/o RE	4.0176	4.0288	4.0288	4.0369	4.0304	4.0369
w/o RC	4.2388	4.2484	4.2484	4.2564	4.2484	4.2564
w/o CS	3.4439	3.4519	3.4647	3.4519	3.4391	3.4471
w/o PC	3.8558	3.8910	3.8910	3.9054	3.8734	3.8654
w/o Cond	4.3974	4.4054	4.4295	4.4231	4.4071	4.3974
w/o PAR	3.9455	3.9599	3.9599	3.9679	3.9583	3.9679
w/o GAR	4.2196	4.2292	4.2292	4.2356	4.2292	4.2356

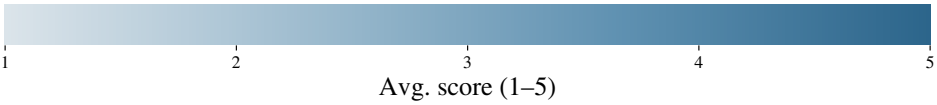

Figure 4: Avg. answer quality for all 21 methods by document family (top) and question type (bottom). Both panels use the same 1–5 scale; ablation labels follow Section 5.2.

Document families. FactWeaver has the highest Avg. score in each of the five families, and Typed-RAG is the strongest external baseline throughout. Their difference ranges from 0.4880 in financial/business documents to 0.6510 in scientific/technical reports. All eleven ablations score below FactWeaver within each family.

Table 22: Avg. answer-quality scores (1–5) by document family. Best and second-best values are bold and underlined, respectively.

Method	Financial/ business ($N = 264$)	Policy/ public-sector ($N = 96$)	Medical/ public-health ($N = 96$)	Scientific/ technical ($N = 96$)	Regulatory/ safety ($N = 72$)
Baselines					
Direct LLM	2.4766	2.4896	2.5608	2.4427	2.3773
Naive RAG	2.7374	2.7517	2.7222	2.6944	2.6921
DOS-RAG	3.1616	3.1615	3.1684	3.1267	3.1829
RAPTOR	2.9684	2.9358	2.9740	2.9288	2.9838
Self-RAG	3.5587	3.5365	3.5503	3.4983	3.6065
HippoRAG 2	3.6162	3.6146	3.6076	3.5747	3.6759
CoRAG	3.9735	3.9410	3.9757	3.8924	4.0046
Search-R1	3.2708	3.2951	3.3403	3.2622	3.2361
Typed-RAG	4.0713	4.0694	4.0608	4.0243	4.1366
Proposed method					
FactWeaver	**4.5593**	**4.5920**	**4.5590**	**4.6753**	**4.6273**
Evidence acquisition					
w/o AFA	3.2917	3.3420	3.3229	3.3733	3.3611
w/o PS	4.2828	4.2917	4.2830	4.3108	4.2963
w/o AFQ	4.1610	4.1736	4.1615	4.1927	4.1782
Relational organization					
w/o GR	3.7045	3.7049	3.7396	3.8212	3.8079
w/o RE	4.0208	4.0330	4.0208	4.0538	4.0394
w/o RC	4.2405	4.2517	4.2413	4.2726	4.2593
Graph-guided synthesis					
w/o CS	3.4110	3.4149	3.4497	3.5399	3.5185
w/o PC	3.8479	3.8767	3.8733	3.9497	3.9213
w/o COND	<u>4.3782</u>	<u>4.3924</u>	<u>4.3941</u>	<u>4.4965</u>	<u>4.4560</u>
w/o PAR	3.9501	3.9635	3.9514	3.9861	3.9676
w/o GAR	4.2216	4.2326	4.2222	4.2517	4.2361

Question types. The ordering between FactWeaver and Typed-RAG also holds across all six non-factoid question types, with differences ranging from 0.5000 on evidence-based questions to 0.5385 on instruction questions. The complete comparison includes both broad stage removals and targeted mechanism ablations; all eleven variants remain below FactWeaver for every question type.

Table 23: Avg. answer-quality scores (1–5) by non-factoid question type. Each type contains 104 questions; best and second-best values are bold and underlined, respectively.

Method	Evidence-based ($N = 104$)	Comparison ($N = 104$)	Experience ($N = 104$)	Reason ($N = 104$)	Instruction ($N = 104$)	Debate ($N = 104$)
Baselines						
Direct LLM	2.4936	2.4920	2.4487	2.4840	2.4551	2.4760
Naive RAG	2.7292	2.7564	2.7163	2.7356	2.6939	2.7212
DOS-RAG	3.1683	3.1955	3.1474	3.1571	3.1298	3.1603
RAPTOR	2.9663	2.9824	2.9615	2.9760	2.9135	2.9599
Self-RAG	3.5609	3.5769	3.5497	3.5272	3.5304	3.5561
HippoRAG 2	3.6202	3.6587	3.6138	3.6026	3.5929	3.6026
CoRAG	3.9679	3.9936	3.9359	3.9760	3.9279	3.9583
Search-R1	3.2869	3.2997	3.2708	3.2917	3.2516	3.2788
Typed-RAG	4.0657	4.0817	4.0641	4.0817	4.0513	4.0737
Proposed method						
FactWeaver	**4.5657**	**4.5897**	**4.5913**	**4.6026**	**4.5897**	**4.6010**
Evidence acquisition						
w/o AFA	3.3221	3.3638	3.3365	3.3221	3.2821	3.3221
w/o PS	4.2788	4.2885	4.2901	4.2965	4.2901	4.2965
w/o AFQ	4.1571	4.1683	4.1699	4.1763	4.1699	4.1779
Relational organization						
w/o GR	3.7340	3.7340	3.7484	3.7452	3.7372	3.7404
w/o RE	4.0176	4.0288	4.0288	4.0369	4.0304	4.0369
w/o RC	4.2388	4.2484	4.2484	4.2564	4.2484	4.2564
Graph-guided synthesis						
w/o CS	3.4439	3.4519	3.4647	3.4519	3.4391	3.4471
w/o PC	3.8558	3.8910	3.8910	3.9054	3.8734	3.8654
w/o COND	<u>4.3974</u>	<u>4.4054</u>	<u>4.4295</u>	<u>4.4231</u>	<u>4.4071</u>	<u>4.3974</u>
w/o PAR	3.9455	3.9599	3.9599	3.9679	3.9583	3.9679
w/o GAR	4.2196	4.2292	4.2292	4.2356	4.2292	4.2356

Evaluation splits. Table 24 reports all methods on the development, held-out human-validation, and automatic-evaluation subsets. The difference between FactWeaver and Typed-RAG is positive in all three subsets: 0.4921, 0.5476, and 0.5205, respectively. These within-split comparisons retain the score assignments used in the aggregate table.

Table 24: Avg. answer-quality scores (1–5) by evaluation split. Best and second-best values are bold and underlined, respectively.

Method	Development ($N = 84$)	Held-out human validation ($N = 84$)	Automatic evaluation ($N = 456$)
Baselines			
Direct LLM	2.4107	2.5317	2.4762
Naive RAG	2.8016	2.6587	2.7237
DOS-RAG	3.2262	3.0972	3.1590
RAPTOR	2.9960	2.9266	2.9594
Self-RAG	3.6151	3.4762	3.5519
HippoRAG 2	3.6488	3.5833	3.6148
CoRAG	3.8988	4.0159	3.9609
Search-R1	3.3274	3.2381	3.2789
Typed-RAG	4.1310	4.0099	4.0694
Proposed method			
FactWeaver	**4.6230**	**4.5575**	**4.5899**
Evidence acquisition			
w/o AFA	3.2877	3.3611	3.3249
w/o PS	4.3512	4.2440	4.2873
w/o AFQ	4.2242	4.1151	4.1700
Relational organization			
w/o GR	3.8016	3.6786	3.7398
w/o RE	3.9603	4.0972	4.0303
w/o RC	4.3254	4.1726	4.2496
Graph-guided synthesis			
w/o CS	3.4722	3.4246	3.4503
w/o PC	3.8095	3.9504	3.8805
w/o COND	<u>4.4583</u>	<u>4.3611</u>	<u>4.4101</u>
w/o PAR	4.0060	3.9167	3.9594
w/o GAR	4.1607	4.2976	4.2299

E.4 CONDITIONAL RESAMPLING OF FINALIZED ANSWER SCORES

We assess sensitivity to corpus composition using the finalized per-answer score table. Averaging its 624 answers per method reproduces all 147 reported method–metric means (21 methods and seven metrics) to numerical precision; each answer's Avg. score equals its unweighted six-criterion mean. The analysis holds these scores fixed and does not reconstruct their transformation from individual judge or adjudication records.

We use a paired cluster bootstrap with 20,000 replicates and random seed 20260908. The 52 documents form 32 corpus units: 20 document pairs and 12 standalone documents. Units are sampled with replacement within each document family, retaining all questions and all methods from each sampled unit. Each replicate recomputes question-weighted means and FactWeaver-minus-comparator differences. This preserves pairing between methods and dependence among questions within a document pair. The 2.5th and 97.5th percentiles define pointwise 95% resampling intervals, conditional on the finalized scores and the observed document-family composition. These are not simultaneous confidence bounds and do not capture scoring transformations, evaluator uncertainty, or variation across independent generation runs.

Table 25: Differences in finalized answer scores under paired corpus-unit resampling. Differences are FactWeaver minus the named comparator; dimension-level rows compare FactWeaver with Typed-RAG. Intervals are pointwise and conditional on the supplied score table.

Comparator / criterion	Difference	95% resampling interval
Avg. score: baseline comparisons		
Direct LLM	+2.1151	[1.8680, 2.3360]
Naive RAG	+1.8646	[1.6060, 2.0990]
DOS-RAG	+1.4303	[1.2180, 1.6150]
RAPTOR	+1.6301	[1.4030, 1.8290]
Self-RAG	+1.0398	[0.8840, 1.1680]
HippoRAG 2	+0.9749	[0.8420, 1.0870]
CoRAG	+0.6301	[0.5490, 0.6980]
Search-R1	+1.3101	[1.1280, 1.4670]
Typed-RAG	+0.5203	[0.4640, 0.5710]
Avg. score: ablation comparisons		
w/o AFA	+1.2652	[1.0900, 1.4160]
w/o GR	+0.8502	[0.7300, 0.9480]
w/o CS	+1.1402	[0.9950, 1.2640]
w/o COND	+0.1800	[0.1170, 0.2500]
w/o PC	+0.7097	[0.6280, 0.7770]
Criterion-level differences from Typed-RAG		
Factual and Numerical Correctness	+0.5833	[0.5130, 0.6610]
Evidence Grounding	+0.5801	[0.4430, 0.7010]
Overclaim Control	+0.6058	[0.5200, 0.6850]
Completeness	+0.2179	[0.1460, 0.2850]
Reasoning Soundness	+0.6971	[0.6010, 0.7900]
Focus and Clarity	+0.4375	[0.3380, 0.5220]

Table 25 characterizes the stability of the reported score differences under corpus-unit resampling. The Avg. difference from Typed-RAG is 0.5203 with interval [0.4640, 0.5710]; among the six criteria, RS has the largest mean difference and CMP the smallest. These summaries concern the fixed finalized scores rather than an independent validation of the scoring procedure or evidence of robustness to new generators, judges, or document distributions.

F Qualitative Examples and Error Analysis

F.1 A Successful Revision Trace

We summarize a held-out human-validation example from the U.S. Department of Energy FY 2023 Agency Financial Report.

Question. *How do DOE's FY 2023 balance sheet, net cost, and budgetary-resource statements compare in explaining financial stewardship?*

Graph construction begins with three selected perspectives: (i) financial reporting, internal control, and audit compliance; (ii) federal accounting standards and liability management; and (iii) operational risk management and payment integrity. Representative factoid queries include "How many consecutive years has DOE received an unmodified audit opinion?", "Which entities are included in the consolidated statements under SFFAS 47?", and "What was DOE's overall improper-payment rate for FY 2023?" These queries retrieve localized facts while the original comparative intent remains in the perspectives and synthesis state.

The atomic fact graph covers all three perspectives and contains one conflict edge between the reported low improper-payment rate and evidence that some contractors resisted oversight-related data requests. Because the question is comparative, a conflict is exposed, and no cross-perspective support edge is available, routing activates the Domain Critic and Mediator in addition to the fixed Planner, Auditor, Reviser, Writer, and Verifier roles.

Table 26: Compressed path audit and revision trace for the DOE example.

State	Recorded behavior
Candidate path	Connects accrual liabilities and delayed budgetary execution but omits the operational-risk perspective.
Audit findings	Missing perspective; an "off-balance-sheet" characterization is not warranted; low improper payments must be qualified by the data-access conflict.
Revision	Adds the payment-integrity and oversight nodes, replaces the unsupported characterization with "accrued liability excluded from capitalization," and places the accounting-basis explanation before the risk qualification.
Selected path	Covers all three perspectives, explicitly traverses the conflict edge, and is accepted as usable by the repeated structural audit.

The resulting answer distinguishes the complementary functions of the three statements: the balance sheet records the scale of recognized long-term obligations, the net-cost statement records program expenses and adjustments, and the budgetary-resource statement records the pace of legal spending authority and cash execution. It further qualifies process-level audit assurance with the operational visibility limitation. The example illustrates the intended role of paths-before-prose: the revision changes the evidence and claim structure before the Writer realizes the answer, rather than attempting to repair an unsupported sentence only after drafting.

F.2 Residual Failure Pattern

The lowest-scoring FactWeaver cases remain generally readable but reveal a common boundary failure: retrieved facts are accurate while the synthesis assigns them an unsupported causal interpretation. In a held-out FCA FY 2024 example, the answer correctly aggregates met, not-met, and not-applicable metrics, but then interprets a monitoring maturity gap as a deliberate consequence of fiscal prioritization and treats an inapplicable metric as evidence that no mission failure occurred. The reviewed evidence does not establish either causal conclusion. This case shows that graph coverage and path revision do not guarantee warranted interpretation when the semantic audit accepts a plausible but unstated mechanism. More conservative causal qualification and stronger claim-level entailment checks remain important directions for improving the framework.