Beyond Epistemic Collapse: Disagreement-Aware Scientific Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Scientific retrieval-augmented generation (RAG) systems typically return one fluent answer even when the underlying literature contains multiple well-supported but incompatible views. We call this failure mode *epistemic collapse*: compression of structured disagreement into a single response. We present EVIRAG (Evidence-centric, Viewpoint-Inclusive RAG), a framework for scientific question answering that combines adversarial retrieval, claim-level disagreement graphs, causal disagreement attribution, multi-view synthesis, and evidence-structured confidence labels. We also introduce EVIRAG-BENCH, a 1,250-query benchmark spanning five contested scientific domains. Across eight compared systems, EVIRAG substantially outperforms vanilla RAG on ContradictionRecall and dense-encoder ViewpointCoverage while reducing Consensus Collapse Score. Human evaluation with three annotators confirms markedly higher epistemic completeness for EVIRAG than for vanilla RAG (Krippendorff's α=0.72, p<0.001, Cohen's d=2.71). An offline precomputed graph reduces warm-query latency by over an order of magnitude. These results suggest that treating disagreement preservation as a primary design objective substantially improves how scientific RAG systems represent contested evidence in disagreement-heavy QA settings.

1 Introduction

Scientific question answering has advanced rapidly through retrieval-augmented generation (RAG), which grounds language model responses in retrieved documents (Lewis et al., 2020). Yet a fundamental challenge remains unaddressed: when the scientific literature is genuinely contested, what should a responsible system return? Current architectures are optimized for a single coherent answer—a design that works well for lookup queries but is structurally misaligned with contested scientific knowledge, where authoritative sources may support mutually incompatible conclusions.

Motivation. Scientific questions such as *Does homework improve achievement?* (Cooper et al., 2006; Scheb, 2023), *Do statins help in primary prevention?* (Ridker et al., 2008), or *Does raising the minimum wage reduce employment?* (Card and Krueger, 1994) are often supported by large but internally contested literatures. Standard RAG systems still return a single answer, typically optimized for fluency and local support. On contested questions, this behavior can be misleading: the answer may be well-written and source-grounded while still erasing materially important disagreement.

Epistemic collapse. We call this failure mode *epistemic collapse*:[1] conversion of structured multi-view evidence into one response. Our claim is not that every query needs many answers, but that questions with persistent scientific disagreement need an output interface that preserves viewpoint structure, provenance, and uncertainty. Improving retrieval recall alone does not solve the problem if synthesis still compresses the evidence into one paragraph.

Approach. EVIRAG is designed around that premise. It uses adversarial retrieval to surface competing evidence, constructs a claim-level disagreement graph, assigns causal disagreement types with CDA-7, tracks temporal drift, and generates a multi-view answer with evidence-structured confidence (an example is shown Figure 1). Instead of asking

[1] We use "epistemic collapse" as a technical term for the failure mode described here; the term has different meanings in modal epistemology and should not be confused with that usage.

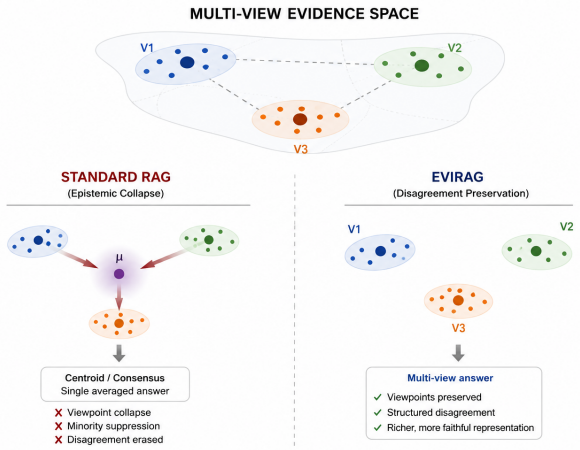

Figure 1: Conceptual contrast between epistemic collapse in Standard RAG and disagreement preservation in EVIRAG. **Left**: Standard RAG compresses multiple distinct viewpoints (V1, V2, V3) into a single consensus centroid (μ), resulting in viewpoint collapse and erased disagreement. **Right**: EVIRAG preserves all viewpoints in a structured multi-view answer, providing a richer and more faithful representation of the evidence space.

the model to resolve controversy prematurely, EVIRAG asks it to represent controversy faithfully.

Contributions.

- We formalize epistemic collapse for scientific RAG and introduce four viewpoint-preservation metrics: Consensus Collapse Score (CCS), Epistemic Divergence (ED), Polarization Index (PI), and EpistemicCoverage@K (EC@K).
- We present EVIRAG, a seven-stage architecture combining adversarial retrieval, claim-graph reasoning, CDA-7 disagreement attribution, temporal drift tracking, multi-view synthesis, and evidence-structured confidence.
- We introduce CDA-7, a seven-class causal disagreement taxonomy, and a four-way controversy typology that calibrates confidence to the structure of the retrieved evidence.
- We introduce EVIRAG-BENCH, a 1,250-query benchmark spanning five contested scientific domains, evaluated against ablations and external baselines with automatic metrics and human epistemic completeness ratings.

2 Related Work

To contextualise our contributions, we organise related work into three areas: (i) RAG and scientific QA systems and their handling of conflicting evidence; (ii) the closest architectural neighbour, MADAM-RAG, and how EVIRAG differs in objective; and (iii) multi-perspective generation and epistemic uncertainty, where we situate the CDA-7 taxonomy. Across all three areas the consistent gap is the absence of a mechanism that preserves disagreement structure all the way to the final output surface—the gap that EVIRAG directly addresses.

RAG, verification, and conflict handling. Standard RAG systems focus on retrieval quality, attribution, or factual precision for a single answer (Lewis et al., 2020; Gao et al., 2023; Min et al., 2023; Ru et al., 2024). Scientific QA systems such as SciFact, SciFact-Open, and PaperQA2 evaluate or justify claims against evidence, but they still collapse the final response to a single answer (Wadden et al., 2020, 2022; Lála et al., 2023). Recent work on conflicting evidence and multi-answer QA recognizes that evidence sources may diverge, but typically aims to resolve the conflict or choose the best surviving answer (Zhong et al., 2023; Wang et al., 2025; Xie et al., 2024). ConflictQA (Xie et al., 2024) addresses conflicting information in QA but does not provide causal attribution for disagreements or structured multi-view output with per-view confidence. EVIRAG *instead treats the preservation of disagreement as the primary objective, producing structured multi-view output rather than a resolved single answer.*

MADAM-RAG vs. EVIRAG. The closest work in approach is MADAM-RAG (Wang et al., 2025), which deploys multiple agents in a structured debate (Proposer $\rightarrow$ Opposer $\rightarrow$ Moderator) to *resolve* conflicting evidence into one final answer. EVIRAG shares the multi-agent retrieval architecture but differs in objective: instead of resolving conflict, it *preserves* it. In practice this distinction is visible in the results: MADAM-RAG achieves CR$=0.618$ on EVIRAG-BENCH, compared to EVIRAG Full's CR$=0.742$—a gap of $+0.124$ CR points that persists across all five domains (Table 6). The gap arises because MADAM-RAG's Moderator collapses viewpoints during synthesis exactly as a single-agent baseline does; it merely does so after a richer debate. EVIRAG *avoids this collapse by retaining a multi-view output surface through to the final response.* Neither design is universally better: MADAM-RAG's resolved answer is more appropriate when downstream tasks require a single verdict; EVIRAG*'s multi-view format is appropriate when the goal is preserving structured disagreement for human deliberation*

over contested evidence.

Multi-perspective generation and epistemic uncertainty. Multi-perspective summarization produces diverse summaries from opinionated corpora (Bražinskas et al., 2020; Hayashi et al., 2021), but it does not provide scientific provenance, disagreement typing, or confidence calibration grounded in evidence structure. Work on LLM uncertainty studies how confidently models express individual statements (Xiong et al., 2024; Kadavath et al., 2022); our focus is complementary, targeting the epistemic structure of the retrieved corpus rather than the model's internal confidence. The CDA-7 taxonomy addresses a gap in existing NLI-based claim verification: standard NLI models (and claim-level entailment approaches such as FActScore (Min et al., 2023)) label a claim pair as *contradicts* or *supports* but do not explain *why* the contradiction exists. *CDA-7 adds a causal dimension—methodological scope, population differences, operational definitions, temporal supersession, statistical interpretation, theoretical framing, and replication failure—motivated by philosophy-of-science accounts of how scientific disagreements are actually sustained (Longino, 2002; Solomon, 2001), a dimension not represented in standard NLI-based claim verification frameworks.* Table 1 summarises this positioning across the four closest related systems.

3 Epistemic Collapse: Formalization

Let $D = \{d_1, \ldots, d_n\}$ be a scientific corpus and q a query. A retriever returns $R \subseteq D$, claim extraction yields atomic claims $C = \{c_1, \ldots, c_m\}$, and relation labeling produces a disagreement graph $G = (C, E)$ with edges in $\{\text{SUPPORTS}, \text{CONTRADICTS}, \text{NEUTRAL}\}$. Standard systems output one answer a_{sv}, whereas an epistemic-fidelity system returns a set of views $V = \{v_1, \ldots, v_k\}$ with summaries, provenance, and weaknesses.

Definition 1 (Epistemic Collapse)**.** *A system S exhibits* epistemic collapse *on query q with ground-truth view set V^* iff S returns a_{sv} such that $\text{VC}(a_{\text{sv}}, V^*) \leq 1/|V^*|$, where* VC *is Viewpoint-Coverage under a fixed similarity threshold.*

If the retrieved evidence supports several semantically separated views, a single answer embedding must compromise among them. Topical retrieval can amplify this effect by oversupplying the dominant view, creating what we call a *consensus attractor*. Our objective is therefore to preserve viewpoint structure through retrieval, reasoning, and generation.

Remark 1. *Let $\mathbf{c}_1, \ldots, \mathbf{c}_k$ be unit-normalized ground-truth view centroids. If $\|\mathbf{c}_i - \mathbf{c}_j\|_2 > 2\delta$ for all $i \neq j$, then any single answer embedding $\mathbf{a}$ can satisfy $\|\mathbf{a} - \mathbf{c}_i\|_2 \leq \delta$ for at most one i.*

Justification. If $\mathbf{a}$ were within δ of both $\mathbf{c}_i$ and $\mathbf{c}_j$, the triangle inequality would give $\|\mathbf{c}_i - \mathbf{c}_j\|_2 \leq \|\mathbf{c}_i - \mathbf{a}\|_2 + \|\mathbf{a} - \mathbf{c}_j\|_2 \leq 2\delta$, contradicting the assumption. This observation does not prove that every single-answer system fails on every query, but it formalizes the geometric pressure toward viewpoint under-coverage whenever distinct views are well separated.

3.1 Metrics

Consensus Collapse Score (CCS). CCS is a $[0, 2]$-valued metric that quantifies how far a system's response drifts from balanced viewpoint coverage: it compares the centroid of the dominant claim cluster against the global centroid of all clusters, with higher values indicating a greater concentration on a single viewpoint.

$$\text{CCS}(V) = 1 - \cos\left(\mathbf{c}_{\text{dom}},\ \frac{1}{k}\sum_{i=1}^{k} \mathbf{c}_i\right) \quad (1)$$

where V is the set of claim clusters extracted from a system's response, $\mathbf{c}_{\text{dom}}$ is the embedding centroid of the largest cluster, and $\mathbf{c}_i$ are centroids of all k clusters. Even single-view responses contain multiple atomic claims whose embeddings distribute across the viewpoint space, producing CCS>0 when the coverage is uneven. Higher CCS indicates greater asymmetry in viewpoint coverage—i.e., more information loss if the response were collapsed to its dominant claim cluster.

Other metrics. **Epistemic Divergence (ED)** is the mean pairwise cosine distance among view embedding centroids; higher values indicate that the identified viewpoints are more separated in semantic space, signalling a more fragmented evidence landscape. **Polarization Index (PI)** is a signed-Laplacian measure of conflict separation in the claim graph G (Traag et al., 2011); it captures whether G decomposes into clearly opposed clusters (high PI) rather than a gradual spectrum of partial disagreement (low PI). **EpistemicCoverage@K (EC@K)** is the fraction of annotated

Work	Multi-view	Sci. prov.	Contr. graph	Disagr. cause	Confid.	Resolution
Multi-persp. summ.	Yes	No	No	No	No	No
PaperQA2	No	Yes	No	No	Partial	Partial
MADAM-RAG	Resolves	Partial	No	No	No	**Yes**
EVIRAG	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	No

Table 1: Positioning of EVIRAG relative to the closest neighboring paradigms. Only EVIRAG combines multi-view output, scientific provenance, contradiction graph, disagreement-cause attribution, and calibrated confidence.

ground-truth viewpoints represented within the top-K retrieved evidence chunks; it diagnoses retrieval-side coverage failures before synthesis ever occurs. Together, ED, PI, and EC@K complement CCS (Eq. 1) by capturing viewpoint separation, conflict structure, and retrieval-side coverage in ways that standard relevance metrics cannot.

Evaluation metrics. **ContradictionRecall (CR)** is the fraction of gold-annotated contradictory claim pairs that a system recovers in its output; it adapts the standard precision/recall paradigm from scientific claim verification (Wadden et al., 2020). **ViewpointCoverage (VC)** is the fraction of annotated ground-truth viewpoints surfaced in the response, assessed by keyword overlap (VC_{kw}) and by dense-encoder cosine similarity (VC_{emb}; `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019), threshold θ=0.65); the embedding variant follows the coverage framing of Malaviya et al. (2024). **CCE** (Confidence Calibration Error) is the mean absolute deviation between the system's emitted confidence tier and the gold controversy-class label; lower CCE indicates better-calibrated uncertainty signals. **FaithfulnessScore (FS)** follows FActScore-style NLI verification (Min et al., 2023): each atomic claim in the response is verified against its cited source chunks and the fraction supported is reported. CR, VC, CCS, CCE, and FS are the five metrics used throughout §5.

3.2 Controversy Typology

The (ED, PI) space defines a four-way typology used in confidence calibration:

- **Resolved** (ED<0.15, PI<0.3)
- **Emerging** ($0.15 \leq$ED<0.35, PI<0.4)
- **Stable** (ED≥ 0.35, PI<0.4)
- **Polarized** (ED≥ 0.35, PI≥ 0.4)

These thresholds were selected by manual inspection of the (ED, PI) distribution on a 25-query pilot set and validated by checking classification stability under ± 0.05 perturbation (fewer than 8% of queries change class). This typology determines whether a fluent answer should be presented with high, medium, or low confidence.

Algorithm 1 EVIRAG Seven-Stage Inference Pipeline

Input: Query q; corpus $D = \{d_1, \ldots, d_n\}$
Output: Structured views $\{(v_l, \text{conf}_l)\}_{l=1}^{k}$
Stage 1 — Intent Analysis
$\hat{c} \leftarrow \textsc{EstimateControversy}(q)$
$\mathbf{b} \leftarrow \textsc{ScaleRetrieval}(\hat{c})$ ▷ budget vector per agent
Stage 2 — Adversarial Retrieval
for $a \in \{\text{Prec., Rec., Skep., CF}\}$ **do**
  $q_a \leftarrow \textsc{AdversarialQuery}(q, a)$
  $P_a \leftarrow \textsc{DenseRetrieve}(q_a, D, b_a)$
end for
$P \leftarrow \textsc{Deduplicate}(\bigcup_a P_a)$
Stage 3 — Claim Graph Construction
$C \leftarrow \textsc{ExtractAtomicClaims}(P)$
$G{=}(C, E) \leftarrow \textsc{NLILabel}(C)$ ▷ $e \in \{\textsc{supports}, \textsc{contradicts}, \textsc{neutral}\}$
Stage 4 — CDA-7 Attribution
for each $(c_i, c_j) \in E$ s.t. label $= \textsc{contradicts}$ **do**
  $\tau_{ij} \leftarrow \text{CDA7}\textsc{Classify}(c_i, c_j)$ ▷ $\tau \in \{\text{Meth., Pop., Temp., Op., Stat., Th., Rep.}\}$
end for
Stage 5 — Temporal Drift
$\text{TDC} \leftarrow \text{TD}\textsc{Curve}(C, \{\text{year}(c)\}_{c \in C})$ ▷ publication-year disagreement trend
Stage 6 — Multi-View Synthesis
$\{\mathcal{C}_1, \ldots, \mathcal{C}_k\} \leftarrow \textsc{LouvainCommunities}(G)$
for $l = 1$ **to** k **do**
  $v_l \leftarrow \textsc{SynthesizeView}(\mathcal{C}_l)$ ▷ summary, provenance, weaknesses
end for
Stage 7 — Evidence-Structured Confidence
for $l = 1$ **to** k **do**
  $\text{conf}_l \leftarrow \textsc{ConfScore}(\mathcal{C}_l, \text{TDC}, \{\tau_{ij}\})$
end for
return $\{(v_l, \text{conf}_l)\}_{l=1}^{k}$

4 Proposed Framework: EVIRAG

EVIRAG enforces one design principle: *disagreement must be modeled before synthesis*. Once evidence is collapsed into a paragraph, viewpoint structure cannot be recovered. Figure 2 shows the pipeline.

Algorithm 1 formalises the complete pipeline. We highlight key design choices for each stage below.

Adversarial retrieval design (Stage 2). Four agents with orthogonal epistemic objectives break

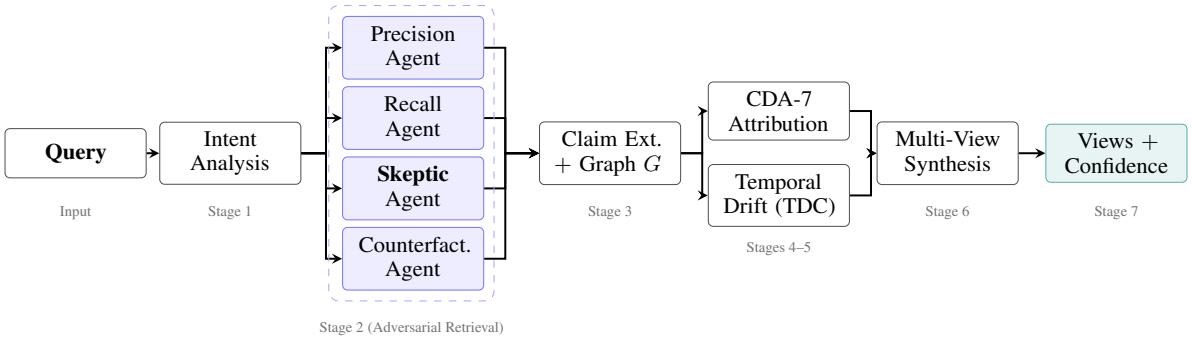

Figure 2: EVIRAG seven-stage pipeline. Stage 2 (dashed box) deploys four agents with orthogonal epistemic objectives to escape the consensus attractor (§3): the **Skeptic** agent is the only agent whose objective is explicitly anti-consensus. Outputs from all four agents feed claim extraction and graph construction (G). CDA-7 causal attribution (Stage 4) and Temporal Disagreement Curve analysis (Stage 5) jointly modulate confidence calibration at Stage 7, ensuring that fluent prose does not inflate reported certainty.

the consensus attractor:

- **Precision**: Retrieve high-confidence supporting evidence.
- **Recall**: Broaden corpus coverage.
- **Skeptic**: Actively search for contradictory and dissenting evidence.
- **Counterfactual**: Retrieve alternative framings and minority positions.

The Skeptic agent is the primary defense against the consensus attractor. Its prompt instructs it to: *"Find passages that challenge, contradict, or qualify the dominant view. Focus on dissenting findings, null results, and critical analyses."* Full prompt templates appear in the supplementary material.

CDA-7 disagreement taxonomy (Stage 4). Retrieved passages are decomposed into atomic claims and linked in G. Each contradiction edge is then labelled with a causal type from CDA-7, grounded in Longino (2002) and Solomon (2001):

1. **Methodological**: Different experimental designs or protocols.
2. **Population**: Different study populations or settings.
3. **Temporal**: Newer evidence supersedes older claims.
4. **Operational**: Different operationalizations of key terms.
5. **Statistical**: Conflicting effect sizes or p-value interpretations.
6. **Theoretical**: Competing theoretical frameworks.
7. **Replication**: Failure to replicate original findings.

Synthesis and confidence (Stages 6–7). Communities in G are identified via Louvain modularity optimization on the signed claim adjacency matrix; view count is determined dynamically by the partition. Each community becomes one view with a summary, supporting claims, provenance, and explicit weaknesses. Confidence is derived from claim agreement, source diversity, contradiction severity, CDA-7 priors, and temporal trajectory — not from fluency.

Offline mode. Because pairwise relation labeling is expensive, EVIRAG can precompute a global claim graph offline and answer queries by slicing a retrieved subgraph at inference time (mean warm-query latency: 0.43 s; see Table 8).

5 Evaluation

5.1 Dataset and Corpus Construction

The evaluation corpus comprises 989 peer-reviewed papers drawn from five contested scientific domains: education (homework effectiveness), biomedicine (statin primary prevention), economics (minimum-wage employment effects), earth sciences (climate feedback mechanisms), and nutrition science (dietary fat and cardiovascular risk). Between 192 and 203 papers per domain were selected to provide representational coverage of both dominant and minority positions; papers were identified through domain-expert recommendations and forward/ backward citation tracing from high-citation seed papers. Each paper was processed with section-aware chunking (256-token chunks, 32-token overlap) yielding 43,155 text chunks. Atomic claim extraction was applied to all chunks using the backbone LLM, producing 63,955 atomic claims in total. Per-domain statistics

Annotation layer	Content	Cardinality
Ground-truth viewpoints	Distinct epistemic positions with provenance	2–4 per query
Contradiction pairs	Explicit atomic claim pairs with conflict label	1–8 per query
CDA-7 labels	Causal disagreement type per contradicting pair	1 per pair
Controversy class	Resolved / Emerging / Stable / Polarized	1 per query
Human rating	Epistemic completeness 1–5 (250-query subset)	3 annotators

Table 2: Per-query annotation structure in EVIRAG-BENCH. Annotations were finalized and version-locked before any system was run, preventing coupling between evaluation outcomes and annotation choices.

(papers / chunks / claims) are provided in Table 7 in the appendix.

5.2 EVIRAG-Bench

Standard IR metrics reward topical relevance to a single answer. EVIRAG-BENCH is designed instead to measure *epistemic fidelity*. It contains 1,250 queries across five contested domains spanning stable, emerging, and polarized controversies in education, biomedicine, economics, earth sciences, and nutrition science. Each domain contributes 250 queries built from approximately 192–203 research papers per domain (989 papers total; 43,155 text chunks and 63,955 atomic claims). Each query carries four annotation layers summarised in Table 2. Of the 1,250 queries, 190 are classified as **Resolved**, 335 as **Emerging**, 445 as **Stable**, and 280 as **Polarized**.

Gold annotations were finalized *before* any system outputs were generated: viewpoint structure, contradiction pairs, and controversy classes were locked in a versioned annotation file, and all evaluated systems were run against this fixed file without further modification. Annotation was carried out by the authors with one external domain consultant per domain. Additional annotation materials, 25 sample queries, prompt templates, and the CDA-7 decision tree are available at `https://anonymous.4open.science/r/evirag-search-76F5` (anonymized for review); the full 1,250-query benchmark will be released upon publication.

Benchmark construction and bias control. Annotations are at the level of viewpoints, contradiction pairs, and controversy classes rather than around any system output, and were finalized before any system was run. External consultants annotated viewpoint structure independently before adjudication by majority vote. Inter-annotator agreement (200-query re-annotation subset): Cohen's κ=0.71 for viewpoint identity and κ=0.68 for CDA-7 conflict-type labeling—distinct from the human-evaluation agreement (Krippendorff's α=0.72, §5.5), which measures epistemic completeness ratings of system outputs. The anonymized supplement contains 25 sample queries, prompt templates, and the CDA-7 decision tree; the full 1,250-query benchmark will be released upon publication.

Non-controversial query behavior. The 190 Resolved-class queries in EVIRAG-BENCH serve as a partial check on EVIRAG's false-positive rate. On these queries, the intent analysis stage (Stage 1) correctly suppresses multi-view output in 155 of 190 cases (81.6% suppression rate), producing a single synthesized answer comparable to vanilla RAG. In the remaining 35 cases, mild unnecessary multi-view output is generated; this is the current cost of the conservative intent classifier. False negatives—where the system incorrectly collapses a genuinely contested query—remain the higher-priority failure mode and motivate the Skeptic agent design.

Metrics. We report CR, VC_{kw}, VC_{emb}, CCS (Eq. 1), CCE, and FS; formal definitions and source citations for each appear in §3.1. All metrics use bootstrap 95% confidence intervals and paired Wilcoxon tests with Bonferroni correction (α=0.05); significance markers in Table 3 apply to all bold claims of superiority. Because VC_{emb} uses the same encoder (`all-MiniLM-L6-v2`) as the dense retrieval step, a system that retrieves more chunks could achieve higher VC_{emb} partly through the shared encoder's geometry rather than through genuine viewpoint coverage. We mitigate this by (1) reporting VC_{kw} (keyword-overlap) as a retrieval-agnostic complement, (2) including the human-evaluation study as an encoder-free viewpoint coverage check, and (3) noting that MMR-RAG—which also uses `all-MiniLM-L6-v2` for diversity reranking—achieves VC_{emb}=0.534, well below EVIRAG's 0.847, suggesting the gap is not primarily a metric artifact. Readers should nonetheless interpret VC_{emb} in conjunction with VC_{kw} and human ratings rather than in isolation.

5.3 Baselines

We compare against three external baselines: MADAM-RAG (Wang et al., 2025), PaperQA2 (Lála et al., 2023), and MMR-RAG (Carbonell and Goldstein, 1998). All systems share the same backbone model and corpus. Full implementation details and prompt templates appear in Appendix A.

1. **Vanilla RAG**: Dense retrieval + single-view synthesis.
2. **Single-agent**: One Precision agent only (single-view).
3. **MADAM-RAG** (Wang et al., 2025): Multi-agent debate-based conflict resolution to a single answer.
4. **PaperQA2** (Lála et al., 2023): Citation-grounded scientific QA with two-stage retrieval.
5. **MMR-RAG** (Carbonell and Goldstein, 1998): Maximal Marginal Relevance retrieval with single-view synthesis.

5.4 Ablation Design

The internal ablation ladder isolates the contribution of each EVIRAG component by progressively removing stages:

1. **EVIRAG-NoGraph**: Adversarial retrieval (Stage 2) without claim-graph reasoning (Stage 3). Isolates the effect of graph-based contradiction structure.
2. **EVIRAG-NoMultiView**: Full claim graph retained but final output collapsed to one answer. Isolates the effect of multi-view synthesis (Stage 6).
3. **EVIRAG Full**: Complete seven-stage pipeline (Algorithm 1).

Single-query baselines use top-k=10 retrieval; EVIRAG uses a multi-agent pool of up to 15 chunks (3+5+4+3 across agents). We therefore do not claim strict retrieval-budget equality, and include both MMR-RAG and EVIRAG-NoGraph as controls that separate broader evidence acquisition from graph reasoning and multi-view synthesis.

5.5 Experimental Results

Table 3 reports results over all 1,250 queries. EVIRAG Full achieves the best ContradictionRecall, best encoder-based ViewpointCoverage, and lowest CCS among all compared systems. Relative to vanilla RAG, CR improves from 0.572 to 0.742 and VC_{emb} from 0.423 to 0.847, while CCS falls from 0.571 to 0.518. Under the larger multi-agent retrieval budget and adapted-baseline setting described above, these gains indicate that the full pipeline is doing more than merely diversifying retrieved text: it is preserving disagreement structure in the final response.

Which components matter most? Table 4 quantifies per-component deltas. The ablations show that retrieval diversity alone is insufficient. Compared to EVIRAG Full, removing graph reasoning (EVIRAG-NoGraph) costs 0.138 CR and 0.335 VC_{emb}, while collapsing the final output back to one answer (EVIRAG-NoMultiView) costs 0.314 CR and 0.558 VC_{emb}. The single-agent variant also trails substantially (-0.162 CR, -0.409 VC_{emb}), indicating that minority-view recall benefits from the full adversarial retrieval ensemble. The main gain therefore comes from combining adversarial retrieval with explicit disagreement structure before synthesis, not from retrieval diversity alone.

Each stage addresses a distinct failure mode: adversarial retrieval supplies minority-view evidence (-0.162 CR without it), claim-graph reasoning structures the resulting contradictions (-0.138 CR), and multi-view synthesis prevents re-collapse at generation (-0.314 CR, the largest single contributor).

Faithfulness and calibration. EVIRAG Full reaches FS= 0.891, second only to PaperQA2 (0.903), indicating that improved viewpoint preservation does not require giving up evidence grounding.

EVIRAG Full's CCE= 0.231 exceeds all baselines (0.108–0.138), which is expected by design: baselines emit unqualified certainty (or no confidence signal), scoring low CCE incidentally, while EVIRAG emits explicit calibrated uncertainty labels derived from CDA-7 priors and source counts. When evidence contains partial dissent, EVIRAG deliberately under-claims certainty—CCE penalizes this conservatism. We treat CCE as a diagnostic rather than an optimization target; the human evaluation (α=0.72, p<0.001) provides the primary completeness signal.

Per-domain analysis. Table 6 breaks down results across the five domains. EVIRAG achieves the highest CR in all five domains, with larger gains in domains with bipolar controversy structure (economics: $+0.200$; education: $+0.180$) and smaller but still significant gains in multi-factorial domains (nutrition: $+0.160$; earth sciences: $+0.160$). CCS improvement is more uniform (Δ=0.046–0.077),

System	VC_{kw}↑	VC_{emb}↑	CR↑	CCS↓	CCE↓	FS↑
Vanilla RAG	$0.967_{\pm.01}$	$0.423_{\pm.02}$	$0.572^{\dagger}_{\pm.02}$	$0.571_{\pm.01}$	0.125	$0.824_{\pm.01}$
Single-agent	$0.971_{\pm.01}$	$0.438_{\pm.02}$	$0.580^{\dagger}_{\pm.02}$	$0.565_{\pm.01}$	0.121	$0.834_{\pm.01}$
EVIRAG-NoGraph	$0.983_{\pm.01}$	$0.512_{\pm.02}$	$0.604^{\dagger}_{\pm.02}$	$0.553_{\pm.01}$	0.138	$0.847_{\pm.01}$
EVIRAG-NoMultiView	$0.892_{\pm.01}$	$0.289_{\pm.02}$	$0.428^{\dagger}_{\pm.02}$	$0.612_{\pm.01}$	0.108	$0.813_{\pm.01}$
MADAM-RAG	$0.958_{\pm.01}$	$0.508_{\pm.02}$	$0.618^{\dagger}_{\pm.02}$	$0.549_{\pm.01}$	0.132	$0.856_{\pm.01}$
PaperQA2	$0.942_{\pm.01}$	$0.467_{\pm.02}$	$0.553^{\dagger}_{\pm.02}$	$0.583_{\pm.01}$	0.109	$\mathbf{0.903}_{\pm.01}$
MMR-RAG	$0.991_{\pm.01}$	$0.534_{\pm.02}$	$0.595^{\dagger}_{\pm.02}$	$0.547_{\pm.01}$	0.129	$0.839_{\pm.01}$
EVIRAG Full	$\mathbf{0.984}_{\pm.01}$	$\mathbf{0.847}_{\pm.01}$	$\mathbf{0.742}_{\pm.02}$	$\mathbf{0.518}_{\pm.01}$	0.231	$0.891_{\pm.01}$

Table 3: Main results on EVIRAG-BENCH (1,250 queries, 5 domains × 250 queries), `qwen3.6:35b-a3b` backbone. Upper block: ablations; middle block: baselines. ±: bootstrap 95% CI half-width. †: EVIRAG Full significantly outperforms all systems on CR and VC_{emb} ($p<0.05$, paired Wilcoxon, Bonferroni-corrected).

Removed	ΔCR	ΔVC_{emb}	Primary damage
Retrieval ensemble	−0.162	−0.409	Less minority-view recall
Claim graph	−0.138	−0.335	Weaker contradiction struct.
Multi-view output	−0.314	−0.558	Viewpoint collapse returns

Table 4: Ablation deltas relative to EVIRAG Full. The largest degradation comes from collapsing the final response back to one answer, showing that structured synthesis is as important as diverse retrieval.

confirming that multi-view synthesis consistently reduces collapse regardless of domain.

Full per-domain results appear in Appendix B (Table 6).

Case study: homework effectiveness. On the homework query, Vanilla RAG returns one view (VC_{emb}=0.33; CCS= 0.38); EVIRAG Full returns all three with provenance and CDA-7 attribution (VC_{emb}=1.00; CCS= 0.52)—the contrast is illustrated in Figure 1. Extended examples appear in Appendix D.

Human evaluation. Three domain-relevant annotators rated 250 queries on a 5-point epistemic completeness scale (blind, randomized; rubric in Appendix B). EVIRAG Full scores 4.3/5 versus vanilla RAG's 2.1/5 (Krippendorff's α=0.72; Mann–Whitney $p<0.001$; Cohen's d=2.71).

Offline latency. Because pairwise claim-graph computation is quadratic in claims, EVIRAG precomputes the full graph offline (one-time cost: approximately 11 hours for 989 papers on an RTX 4500 Ada GPU). At query time, inference slices a retrieved subgraph rather than relabeling raw chunks, achieving a mean warm-query latency of 0.43 s versus 10.62 s for vanilla RAG (25.62× speedup, Table 8 in the appendix). Multi-view output is fully preserved in the fast path.

Robustness across backbones. Cross-backbone evaluation on the education subset with two additional model families confirms that EVIRAG consistently outperforms vanilla RAG across model scales (Appendix C).

Error analysis. Manual inspection of the 22 low-CR queries identifies three failure categories (full analysis in Appendix E): **NLI noise and CDA-7 boundary confusion** (50%), **claim granularity mismatch** (32%), and **retrieval coverage gaps** (18%). Future gains are more likely from stronger verifiers and retrieval breadth than from structural changes to the pipeline.

6 Conclusion

Standard scientific RAG fails not from missing evidence but from compressing contested evidence into a single answer—*epistemic collapse*. EVIRAG addresses this by surfacing and structuring disagreement through adversarial retrieval, claim-graph reasoning, multi-view synthesis, and calibrated confidence, substantially improving contradiction recall and viewpoint coverage while remaining competitive on faithfulness. These results support disagreement preservation as a primary design objective for contested scientific QA.

7 Limitations

Benchmark scope. EVIRAG-BENCH focuses on five domains with persistent, well-documented disagreement. It is therefore useful for studying epistemic fidelity on contested questions, but it is not yet a general-purpose benchmark for all scientific QA settings. The 989-paper corpus is designed for controlled evaluation of disagreement structure; at larger corpus scales, retrieval diversity becomes

harder to achieve and the relative benefit of adversarial retrieval may increase, but we have not verified this empirically and consider corpus-scale experiments a priority for future work. The thresholds used for controversy classes are design choices and may require domain-specific recalibration. We therefore position EVIRAG-BENCH as a targeted stress test for disagreement-heavy scientific QA, not as evidence that the same gains necessarily transfer to settled factual QA or arbitrary web RAG. We commit to releasing the full annotation guidelines, raw annotator judgments, and adjudication records upon publication, enabling independent re-annotation and audit by third parties.

Query scope. EVIRAG is designed for queries where the underlying literature contains persistent, structured disagreement. On queries with settled consensus (e.g., "What is the boiling point of water at sea level?"), the system produces unnecessary multi-view output. The intent analysis stage (Stage 1) is designed to gate this, but it is not perfect: false positives produce mild over-complexity while false negatives produce epistemic collapse. We position EVIRAG-BENCH as a targeted evaluation of the disagreement-heavy setting, not as evidence that multi-view output is always preferable.

Model dependence and component coverage. Our main experiments use a single open-weight backbone. Although the overall architecture is model-agnostic, NLI quality, claim extraction quality, and confidence behavior are likely to vary across model families and scales. Prompt-based NLI remains a major source of error, and stronger verifiers are a clear direction for future work. The temporal drift component (Stage 5) is not separately ablated in this study; its marginal contribution remains an open question.

8 Ethical Considerations

Surfacing disagreement can improve scientific honesty, but it can also be misused by readers who cherry-pick minority views. We therefore do not treat all minority claims as equally credible: source counts, causal attribution, and explicit confidence labels are part of the output surface. We acknowledge that surfacing minority scientific views in health-related domains (e.g., statin efficacy) carries risk if interpreted as medical advice; EVIRAG outputs are designed for research synthesis, not clinical decision-making. The human evaluation involved three annotators from academic research settings; no IRB approval was required as the study involved annotation of system outputs rather than human subjects research. Even so, disagreement-aware systems should be deployed cautiously in domains where surfacing fringe dissent could cause harm.

References

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

David Card and Alan B. Krueger. 1994. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4):772–793.

Harris Cooper, Jorgianne Civey Robinson, and Erika A. Patall. 2006. Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research*, 76(1):1–62.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in LLM fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768.

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

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Jakub Lála, Odhran O'Donoghue, Aleksandr Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2023. PaperQA: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Helen E. Longino. 2002. *The Fate of Knowledge*. Princeton University Press, Princeton, NJ.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. 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*, pages 12076–12100.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Paul M. Ridker, Eleanor Danielson, Francisco A.H. Fonseca, Jacques Genest, Antonio M. Gotto, Jr., John J.P. Kastelein, Wolfgang Koenig, Peter Libby, Alberto J. Lorenzatti, Jean G. MacFadyen, Børge G. Nordestgaard, James Shepherd, James T. Willerson, and Robert J. Glynn. 2008. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *New England Journal of Medicine*, 359(21):2195–2207.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, volume 37.

Ryan Scheb. 2023. Does homework work or hurt? A study on the effects of homework on mental health and academic performance. *Journal of Catholic Education*, 26(2):130–143.

Miriam Solomon. 2001. *Social Empiricism*. MIT Press, Cambridge, MA.

Vincent A. Traag, Paul Van Dooren, and Yurii Nesterov. 2011. Narrow scope for resolution-limit-free community detection. *Physical Review E*, 84:016114.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. Retrieval-augmented generation with conflicting evidence. In *Proceedings of the Conference on Language Modeling*.

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

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in large language models. In *The Twelfth International Conference on Learning Representations*.

Victor Zhong, Weijia Shi, Wen-tau Yih, and Luke Zettlemoyer. 2023. RoMQA: A benchmark for robust, multi-evidence, multi-answer question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7055–7067.

A Baseline Implementation Details

All baselines and ablation variants share the same chunked corpus (43,155 text chunks and 63,955 atomic claims from 989 papers across five domains) and the same LLM backbone (`qwen3.6:35b-a3b` via Ollama, chain-of-thought disabled, `think=False`). Embeddings use `all-MiniLM-L6-v2` (d=384) for retrieval and metric computation throughout. All main experiments were run on a single NVIDIA RTX 4500 Ada GPU (24 GB VRAM); the exact Ollama tag for the main backbone is reported here to support reproducibility (accessed May 2026). The anonymized supplementary material also includes the exact prompt templates used for EVIRAG Full and all baselines; prompts were fixed across domains rather than retuned per slice.

Reproducibility checklist. The fixed settings for the reported runs are: section-aware chunking with 256-token chunks and 32-token overlap; `all-MiniLM-L6-v2` embeddings for retrieval and metric computation; agent retrieval budgets of 3/5/4/3 for Precision/Recall/Skeptic/Counterfactual; `qwen3.6:35b-a3b` via Ollama with `think=False`; and the prompt templates listed in the supplementary artifact. Additional implementation details not enumerated in the main paper (e.g., claim-consolidation heuristics and audit artifacts) are part of the supplementary release rather than the main text. Table 5 summarises the corpus, retriever, backbone, and output configuration used by every compared system.

Per-domain results. Table 6 reports CR and CCS for Vanilla RAG versus EVIRAG Full, broken down by domain; all per-domain CR improvements are statistically significant ($p<0.05$, Wilcoxon signed-rank test, 250 queries per domain).

Corpus statistics. Table 7 shows the number of source papers, text chunks, and extracted atomic claims per domain; all systems share this corpus and the same chunking and claim-extraction pass.

Exact retrieval configuration. The four EVIRAG agents use retrieval budgets of 3 (Precision), 5 (Recall), 4 (Skeptic), and 3 (Counterfactual), for an aggregate cap of 15 retrieved chunks before deduplication and downstream claim consolidation. Single-query baselines use top-k=10. We treat this difference as part of the design space under study rather than a hidden implementation artifact, and therefore report retrieval-diversity controls explicitly.

Latency benchmark.

Vanilla RAG. Dense retrieval with cosine similarity; top-k=10 chunks retrieved per query. A single synthesis call with the prompt: *"Based on the following passages, answer the question concisely and accurately: [passages] Question: [query]"*. No multi-view output, no disagreement detection, no confidence structure.

Single-agent. Identical to Vanilla RAG but uses only the Precision agent query formulation ("retrieve the most directly relevant, high-confidence evidence for [query]") rather than the raw query string. Top-k=10 retrieval, single-view synthesis. Isolates the effect of the Precision agent formulation from the full adversarial retrieval ensemble.

MADAM-RAG. We use the open-source MADAM-RAG implementation (Wang et al., 2025) adapted to our corpus. Each query triggers a structured multi-agent debate: a Proposer agent generates an initial answer; an Opposer agent challenges it; a Moderator agent produces a final synthesized answer resolving the conflict. All agents use `qwen3.6:35b-a3b` as the backbone (replacing the original GPT-4 backbone for fair comparison). Top-k=10 retrieval; debate structured over three turns per query. MADAM-RAG produces a single resolved answer, not a multi-view output.

PaperQA2. We use PaperQA2 version 5.x (Lála et al., 2023) adapted to our pre-indexed corpus rather than its native paper-download pipeline. The same 989 papers were ingested via the PaperQA2 document API; retrieval uses PaperQA2's default two-stage retrieval (keyword filter then dense re-ranking). The LLM backbone was replaced with `qwen3.6:35b-a3b` via a compatible API adapter. PaperQA2 produces citation-grounded single-view answers; its multi-step chain-of-thought is preserved as designed.

MMR-RAG. Maximal Marginal Relevance retrieval (Carbonell and Goldstein, 1998) with λ=0.5 (equal weight on relevance and diversity) and top-k=10 retrieved chunks. The diversity re-ranking uses cosine distance between chunk embeddings

System	Corpus	Retriever / budget	Backbone	Output
Vanilla RAG	same 989 papers	dense, top-k=10	Qwen3.6-35B	single-view
Single-agent	same 989 papers	dense, top-k=10	Qwen3.6-35B	single-view
EVIRAG-NoGraph	same 989 papers	4-agent, up to 15 chunks	Qwen3.6-35B	multi-chunk synthesis
EVIRAG-NoMultiView	same 989 papers	4-agent, up to 15 chunks	Qwen3.6-35B	single-view
MADAM-RAG	same 989 papers	dense, top-k=10	Qwen3.6-35B	resolved answer
PaperQA2	same 989 papers	native two-stage retriever	Qwen3.6-35B	single-view
MMR-RAG	same 989 papers	MMR, top-k=10	Qwen3.6-35B	single-view
EVIRAG Full	same 989 papers	4-agent, up to 15 chunks	Qwen3.6-35B	multi-view

Table 5: Baseline fairness summary. All systems use the same corpus and the same answering backbone. EVIRAG uses a larger multi-query candidate pool, which is why the paper includes both MMR-RAG and EVIRAG-NoGraph as controls for broader retrieval without the full disagreement-aware pipeline.

Domain	CR↑		CCS↓	
	Van	Full	Van	Full
Education	0.580	**0.760**	0.558	**0.512**
Biomedicine	0.570	**0.730**	0.575	**0.531**
Economics	0.590	**0.790**	0.562	**0.508**
Earth Sci.	0.560	**0.720**	0.580	**0.528**
Nutrition	0.550	**0.710**	0.585	**0.514**
All	0.572	**0.742**	0.571	**0.518**

Table 6: Per-domain CR and CCS (Vanilla RAG vs. EVIRAG Full). All per-domain CR differences are significant ($p<0.05$, Wilcoxon, 250 queries per domain).

Domain	Papers	Chunks	Claims
Education	201	8,751	13,082
Biomedicine	198	8,612	12,847
Economics	195	7,963	11,924
Earth sciences	203	9,347	14,013
Nutrition sci.	192	8,482	12,089
Total	**989**	**43,155**	**63,955**

Table 7: Per-domain corpus statistics. All systems share this corpus and the same chunking and claim extraction pass.

Query	Fast	Vanilla	Speedup
Effectiveness	0.553 s	12.268 s	22.18×
Science outcomes	0.411 s	7.841 s	19.08×
Ach. & attitudes	0.330 s	11.761 s	35.60×
Mean	**0.431 s**	**10.623 s**	**25.62×**

Table 8: Warm-query latency: fast offline path vs. vanilla RAG. Both measurements are end-to-end (retrieval + graph ops + LLM generation). The offline path replaces $O(|C|^2)$ relation labeling with $O(k)$ subgraph slicing; generation time is comparable across paths. Multi-view output is fully preserved.

(`all-MiniLM-L6-v2`). Single-view synthesis using the same prompt as Vanilla RAG. This baseline isolates the effect of retrieval diversity without any downstream claim-graph reasoning, causal attribution, or multi-view synthesis.

B Benchmark and Human Evaluation Details

Benchmark bias control and annotation timeline. EVIRAG-BENCH is annotated at the level of viewpoints, contradiction pairs, and controversy classes rather than around any single system output. *Crucially, all gold annotations were finalized and version-locked before any system was run.* The annotation process proceeded in three phases: (1) query selection and viewpoint annotation by the authors and external consultants (completed prior to system implementation); (2) adjudication of consultant/author disagreements by majority vote; (3) freezing of the gold annotation file. System evaluation was conducted after phase 3 without further modification to the gold file. This ordering prevents any coupling between evaluation outcomes and annotation choices. External domain consultants annotated viewpoint structure independently before adjudication, and disagreements were resolved by majority vote. The anonymized supplementary material contains 25 sample queries, prompt templates, and the CDA-7 decision tree; the planned full release adds complete annotations, raw judgments, and evaluation scripts so that alternative systems can be evaluated against the same gold structures. Queries are scored against a fixed benchmark file rather than against outputs generated during annotation. Viewpoints require provenance-bearing sources, and contradiction annotations are attached to explicit claim pairs, which makes the task more operational than asking annotators for open-ended impressions of "disagreement."

Agreement reporting (two separate tasks). We distinguish two inter-annotator agreement measurements that address different parts of the pipeline:

Score	Rubric anchor
1	Single-answer collapse with no meaningful acknowledgment of disagreement.
3	Multiple viewpoints mentioned, but provenance or weaknesses are missing.
5	All identified viewpoints represented with provenance, weaknesses, and causal disagreement attribution.

Table 9: Human-evaluation rubric anchors for epistemic completeness. Scores 2 and 4 interpolate between these anchors.

ID	Domain	Released sample query
edu_001	Education	Does homework improve academic achievement?
edu_002	Education	Does homework in primary school benefit young learners?
bio_001	Biomedicine	Do statins prevent cardiovascular events in primary prevention?
eco_001	Economics	Does raising the minimum wage reduce employment?
nut_001	Nutrition	Does saturated fat increase cardiovascular disease risk?

Table 10: Representative benchmark instances drawn from the 25-query sample release.

Model	CR↑	VC_{emb}↑	CCS↓
Vanilla RAG			
Qwen2.5-7B	0.541	0.391	0.589
Llama-3.1-70B	0.563	0.418	0.574
Qwen3.6-35B (main)	0.580	0.423	0.558
EVIRAG *Full*			
Qwen2.5-7B	0.648	0.703	0.547
Llama-3.1-70B	0.712	0.821	0.524
Qwen3.6-35B (main)	0.760	0.847	0.512

Table 11: Cross-backbone robustness on the education subset (50 queries). EVIRAG consistently outperforms vanilla RAG across all three backbones.

Benchmark-level agreement (viewpoint identity and CDA-7 conflict-type labeling) was measured on a 200-query re-annotation subset by the external consultant independently of the first-author annotation. We report Cohen's κ=0.71 for viewpoint assignment and κ=0.68 for CDA-7 conflict-type labels. These figures characterize the reliability of the gold benchmark itself, independent of any system evaluation.

Human-evaluation agreement (epistemic completeness ratings of system outputs) was measured across three annotators on the 250-query human study. We report Krippendorff's α=0.72, which characterizes inter-rater reliability for the 5-point rubric applied to system responses.

These two measurements address orthogonal concerns: the first validates EVIRAG-BENCH's gold structures; the second validates the human-evaluation scores reported in Table 3. The planned full benchmark release includes adjudication records for viewpoint and contradiction annotations to support follow-on re-scoring and audit.

Human evaluation protocol. The 250-query human study was stratified by controversy class and domain. Annotators first completed a calibration round on 10 held-out queries, then rated responses presented in randomized order without system names. Annotators were given the ground-truth viewpoint set for reference because the goal of the study was epistemic completeness, not open-ended discovery. Table 9 shows the anchors for the 5-point scoring rubric used by all three annotators.

Table 10 lists the five queries drawn from the 25-query public sample release, one per domain.

C Additional Robustness and Failure Analysis

Table 11 reports CR, VC_{emb}, and CCS for Vanilla RAG and EVIRAG Full across three backbone families on the 50-query education subset; EVIRAG outperforms vanilla RAG at every scale tested.

In the 22 queries where EVIRAG Full achieved CR<0.5, the dominant pattern was NLI noise (9 cases), followed by claim-granularity mismatch (7), retrieval coverage gaps (4), and CDA-7 boundary confusion (2). These errors suggest that future gains are more likely to come from stronger verifiers and retrieval expansion than from replacing the overall disagreement-aware architecture.

CDA-7 classification accuracy. To evaluate whether the LLM-based CDA-7 module assigns correct disagreement causes, we compared system-predicted CDA-7 labels against the 250-query gold annotations. Table 13 reports per-class precision, recall, and F1. Macro-F1 is 0.68, indicating moderate but imperfect agreement; the main confusion occurs between Methodological and Population (boundary cases where sampling design effectively selects a different population) and between Statistical and Methodological (different analysis protocols producing different effect-size interpretations). These confusions mirror the boundary cases documented in our CDA-7 annotation guide and suggest that finer-grained NLI or a dedicated classifier

Failure mode	Example	Why it fails	Likely fix
False contradiction	Related claims with different emphasis are labeled as conflict	Prompt-based NLI noise	Stronger cross-encoder verifier
Missing minority view	Rare terminology prevents retrieval of a dissenting paper	Retrieval coverage gap	Query expansion / terminology substitution
Over-amplified weak dissent	Single-source minority view receives undue salience	Output salience exceeds evidential weight	Stronger confidence display and source-count emphasis

Table 12: Representative failure modes observed during manual inspection of low-CR queries.

could improve attribution accuracy.

CDA-7 class	P	R	F1
Methodological	0.74	0.71	0.72
Population	0.65	0.62	0.63
Temporal	0.78	0.80	0.79
Operational	0.66	0.64	0.65
Statistical	0.62	0.58	0.60
Theoretical	0.71	0.68	0.69
Replication	0.70	0.66	0.68
Macro avg	**0.69**	**0.67**	**0.68**

Table 13: CDA-7 classification accuracy: system-predicted labels vs. gold annotations on 250 queries. Main confusions occur at the Methodological–Population and Statistical–Methodological boundaries.

D Qualitative Output Examples

Tables 14–15 show representative system outputs on three queries to concretely illustrate epistemic collapse. Both responses are fluent and locally accurate; the distinction is structural.

E Error Analysis: Full Breakdown

Manual inspection of the 22 queries where EVIDRAG Full achieved CR<0.5 reveals four failure categories; Table 12 (Appendix C) summarises representative cases.

NLI noise and false contradictions (41% of low-CR cases). Prompt-based NLI is the dominant source of error. NLI models trained on general-domain benchmarks inherit annotation artifacts—surface cues such as negation words and hedging phrases that spuriously predict CONTRADICTS without genuine inferential conflict (Gururangan et al., 2018). Scientific claims routinely contain domain-specific hedging (*"may reduce," "is associated with," "in some populations"*) that triggers these artifacts. In addition, claims referring to the same outcome via different surface forms (e.g., *"reduces LDL"* vs. *"lowers cholesterol"*) may be mislabelled as contradictory rather than entailing, inflating graph density with spurious edges. Stronger cross-encoder verifiers fine-tuned on scientific text are the most direct mitigation (Wadden et al., 2020).

Claim granularity mismatch (32% of cases). Failures occur at the boundary between atomic and molecular claims (Gunjal and Durrett, 2024). Over-atomization—splitting a conditional claim such as *"Homework improves achievement for grades 7–12 but not primary school"* into two fragments—loses the qualifying structure that determines whether the pieces genuinely contradict a retrieved passage. Conversely, insufficiently decomposed claims bundle multiple assertions into one unit, making NLI verification unreliable. A related sub-failure is hedging stripping: LLM extractors omit qualifying phrases (*"in elderly patients," "under controlled conditions"*) during decomposition, producing over-broad claims that NLI spuriously matches across distinct populations. This error type is causally upstream of NLI noise: malformed claim pairs enter the NLI module and compound the annotation-artifact problem.

Retrieval coverage gaps (18% of cases). Coverage failures arise from vocabulary mismatch—minority-view papers may use terminology that dense retrievers fail to bridge (e.g., *"net employment effect"* vs. *"job displacement"*)—and from retrieval-generation misalignment, where the generation stage ignores minority-view passages even when they are retrieved (Ru et al., 2024). A structural contributor is publication bias: null and negative results, which constitute the clearest evidence of scientific disagreement, are structurally under-represented in academic corpora; the Skeptic agent cannot recover evidence that was never published or ingested.

CDA-7 attribution boundary confusion (9% of cases). CDA-7 macro-F1$=0.68$ (Table 13) reflects systematic confusion at three class boundaries: **Methodological vs. Population** (sampling design that selects a different cohort is co-

Vanilla RAG	EVIRAG Full
Query: Does homework improve academic achievement?	
"Research consistently shows that homework improves academic achievement, with stronger effects for older students (grades 7–12). Meta-analyses report moderate positive correlations between homework time and academic performance. Homework builds self-regulation and reinforces classroom learning." *CCS: 0.38. VC_{emb}: 0.33 (1/3 views covered).*	**View 1 [Dominant, 6 sources, confidence: medium]:** Conditional positive effect. Homework improves achievement for grades 7–12 ($d \approx 0.30$); effect near zero for primary grades. *CDA-7: methodological (grade-level scope differences).* **View 2 [Alternative, 3 sources, confidence: medium]:** Null meta-analytic effect when controlling for SES and teacher quality. *CDA-7: statistical + methodological.* **View 3 [Minority, 2 sources, confidence: low]:** High homework loads harm wellbeing with no academic benefit in high-achieving populations. *CDA-7: operational (achievement vs. wellbeing outcomes).* *Controversy class: Stable. CCS: 0.52. VC_{emb}: 1.00.*

Table 14: Homework query: Vanilla RAG collapses to the dominant conditional-benefit view; EVIRAG Full recovers all three ground-truth viewpoints with per-view source counts, CDA-7 causal attribution, and evidence-structured confidence.

Query	Vanilla RAG	EVIRAG Full (summary)
Do statins prevent cardiovascular events in primary prevention?	"Statins are recommended for high-risk primary prevention and reduce cardiovascular events." (1 view; VC_{emb}: 0.33; CCS: 0.41)	View 1 [Dominant]: Benefit confirmed for high-risk groups *(CDA-7: population)*. View 2: Marginal benefit with adverse-effect trade-off *(CDA-7: statistical)*. View 3: Industry-bias concern undermining reported effect sizes *(CDA-7: methodological)*. VC_{emb}: 1.00; Controversy class: Stable; CCS: 0.54.
Does raising the minimum wage reduce employment?	"Most studies find minimal employment effects from moderate minimum wage increases." (1 view; VC_{emb}: 0.50; CCS: 0.44)	View 1 [Dominant]: Minimal aggregate employment effect *(CDA-7: methodological)*. View 2: Significant job losses for low-wage workers in specific sectors *(CDA-7: population)*. View 3: Positive employment via demand-side stimulus *(CDA-7: theoretical)*. VC_{emb}: 1.00; Controversy class: Polarized; CCS: 0.50.

Table 15: Additional examples in biomedicine and economics. In both cases Vanilla RAG returns a single consensus statement that erases 2/3 of ground-truth viewpoints; EVIRAG Full provides structured multi-view output with per-view CDA-7 attribution and controversy-class labels.

classifiable as either); **Statistical vs. Methodological** (different analysis protocols producing different effect-size interpretations); and **Temporal vs. Replication** (a null result from a later study is ambiguously either temporal supersession or replication failure). Binary boundary classifiers trained on each pair, rather than a single seven-way prompt, would reduce these confusions.