AdaLoc: Adaptive Granularity and Hierarchical Filtering for Evidence-Aware Academic Search

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) improves factuality for scientific question answering, but scientific queries vary systematically in evidence scope. A single static retrieval pipeline faces a granularity–precision tension: strategies that maximize semantic coverage for global synthesis often miss needle-like evidence required for local extraction. We propose **AdaLoc**, a scope-adaptive evidence localization module that (i) routes queries by predicted scope, (ii) selects scope-specific evidence granularity, and (iii) applies a hierarchical filter cascade to balance semantic and lexical matching. We introduce SciScope, a synthetic diagnostic dataset of 2,000 scope-annotated queries. On SciScope, AdaLoc achieves 74.37% F1 on local extraction, outperforming the full-context baseline (20.17%) and a prompt-optimized variant (71.20%), while maintaining strong global synthesis (56.86% vs. 55.03%). This confirms that scope-aware routing jointly outperforms any single prompt or context configuration. We further validate on Qasper, achieving competitive extractive performance with over 50% fewer tokens. Code and data: https://anonymous.4open.science/r/AdaLoc-EEC3.

1 Introduction

Large Language Models (LLMs) have transformed knowledge-intensive tasks (OpenAI, 2023), yet hallucinate (Ji et al., 2023). Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) mitigates this by grounding responses in retrieved evidence.

While effective in general domains, applying RAG to scientific question answering (QA) presents unique challenges due to the complexity of academic documents (Dasigi et al., 2021; Hong et al., 2024). The stakes are considerably higher in scientific and medical domains, where mitigating hallucinations through rigorous evidence retrieval is paramount for trustworthy AI deployment (Yang et al., 2025). Existing scientific retrieval systems, such as PaSa (He et al., 2025), primarily focus on document-level retrieval—identifying which papers are relevant. However, retrieving the correct paper is only the first step. The critical bottleneck is evidence localization: pinpointing where within a dense, multi-page paper the answer resides. Standard RAG pipelines often employ a static, "one-size-fits-all" chunking and retrieval strategy, which fails to accommodate the high variance in information density typical of scientific literature (Jiang et al., 2023; Asai et al., 2024).

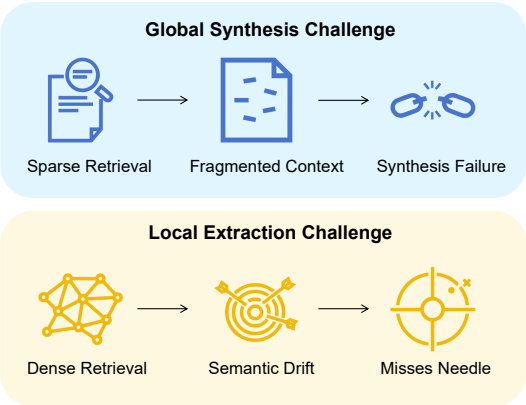

Figure 1: **The Granularity-Precision Trade-off.** (Top) Global queries fail with sparse retrieval due to context fragmentation; (Bottom) Local queries suffer from semantic drift in dense retrieval. AdaLoc resolves this specific tension in a principled way.

Specifically, scientific queries differ systematically in evidence scope, creating a fundamental tension in retrieval granularity. We identify two distinct challenges. First, the **Global Synthesis Challenge**: queries like "How does the proposed method contrast with prior work?" require aggregating information dispersed through the introduction, method, and related work sections. Here, sparse lexical retrieval (e.g., BM25) fragments the context, retrieving scattered keywords while missing the semantic holistic view required for comprehensive summarization across the entire span of the

paper within its complex structure.

Second, the **Local Extraction Challenge**: queries like "What is the dropout rate in Table 2?" target a sharply localized fact. Dense semantic retrieval often suffers from semantic drift, retrieving conceptually related passages (e.g., discussions on regularization) while missing the exact numeric cell. Crucially, simply increasing context length is not a remedy, as long-context prompting suffers from the "lost-in-the-middle" effect (Liu et al., 2024), where decisive details buried in mid-document tables are overlooked ultimately compromising the accuracy of fact extraction.

To resolve this granularity–precision trade-off, we propose **AdaLoc** (Adaptive Localization), a framework treating evidence localization as a dynamic decision process. AdaLoc routes queries by predicted scope: GLOBAL queries employ broader semantic chunking with LLM-based filtering; LOCAL queries switch to fine-grained chunking with strict lexical constraints (e.g., in-paper BM25) to pinpoint "needle-in-a-haystack" facts. This routing relies on configurable lexical plugins, allowing AdaLoc to generalize across domains (e.g., finance, law) without retraining and without requiring labeled training data. We evaluate AdaLoc using a two-stage protocol. First, on **QASPER** (Dasigi et al., 2021), AdaLoc achieves competitive performance while halving token usage. To further stress-test the granularity–precision trade-off, we introduce SCISCOPE, a synthetic adversarial benchmark explicitly designed to expose the gap between global synthesis and local extraction. On SCISCOPE, AdaLoc outperforms strong baselines across both query types. Crucially, while prompt-optimized full-context models suffer a "Seesaw Effect"—where improving local extraction catastrophically degrades global synthesis—AdaLoc's scope-aware routing avoids this trade-off entirely.

Our primary contributions are: (i) we identify the scope-granularity mismatch in scientific QA and propose **AdaLoc**, a novel framework that adaptively modulates retrieval granularity and filtering logic; (ii) we demonstrate AdaLoc's effectiveness on QASPER, and introduce SCISCOPE, a synthetic diagnostic benchmark for stress-testing the granularity-precision trade-off; and (iii) extensive experiments show that scope-aware routing, which jointly matches prompt type and retrieval granularity, is the key to resolving the extraction-synthesis tension at a fraction of the token cost.

2 Related Work

2.1 Scientific Retrieval and Granularity

Scientific QA requires navigating complex documents. Traditional methods utilize dense retrievers (Beltagy et al., 2019; Karpukhin et al., 2020; Izacard et al., 2022) or agentic frameworks like PaperQA (Lála et al., 2023) and OpenScholar (Zheng et al., 2024) for document discovery. However, these systems often treat retrieved papers as monolithic blocks or apply fixed-size chunking (Pradeep et al., 2022), leading to precision loss in intra-document localization. To address granularity, methods like Atlas (Izacard et al., 2023) and RAPTOR (Sarthi et al., 2024) introduce hierarchical chunking, while recent studies like Reconstructing Context (Jin et al., 2025) highlight that advanced chunking logic significantly impacts RAG performance. Mix-of-Granularity (MoG) (Zhong et al., 2024) attempts to solve this by statically fusing multiple fixed granularities during retrieval. Unlike these prior approaches, which either evaluate static chunking tradeoffs or fuse granularities through expensive pre-indexing and training, AdaLoc implements a zero-shot, scope-adaptive strategy. We argue that granularities should not be statically fused, but rather treated as a dynamic decision dependent on the user query's specific evidence scope (GLOBAL vs. LOCAL). By decoupling these strategies, AdaLoc specifically solves the "last-mile" localization problem that existing document-level agents often overlook in practice.

2.2 Adaptive RAG Paradigms

Recent work modulates retrieval or reasoning: ReAct (Yao et al., 2023) interleaves thought and action; Self-RAG (Asai et al., 2024) and FLARE (Jiang et al., 2023) decide when to retrieve; CRAG (Yan et al., 2024) and Adaptive-RAG (Jeong et al., 2024) route queries as needed. Others, such as coordinated semantic alignment (Chen et al., 2026) and evidence-driven generation (Yue et al., 2024), enforce strict semantic constraints to prevent factual drift. AdaLoc introduces an orthogonal dimension: retrieval logic adaptation. Instead of just triggering retrieval, AdaLoc determines how to retrieve—switching between coarse semantic synthesis and fine-grained lexical extraction. This allows AdaLoc to function as a specialized retrieval module within broader adaptive frameworks (see Appendix H for comparison with Self-RAG).

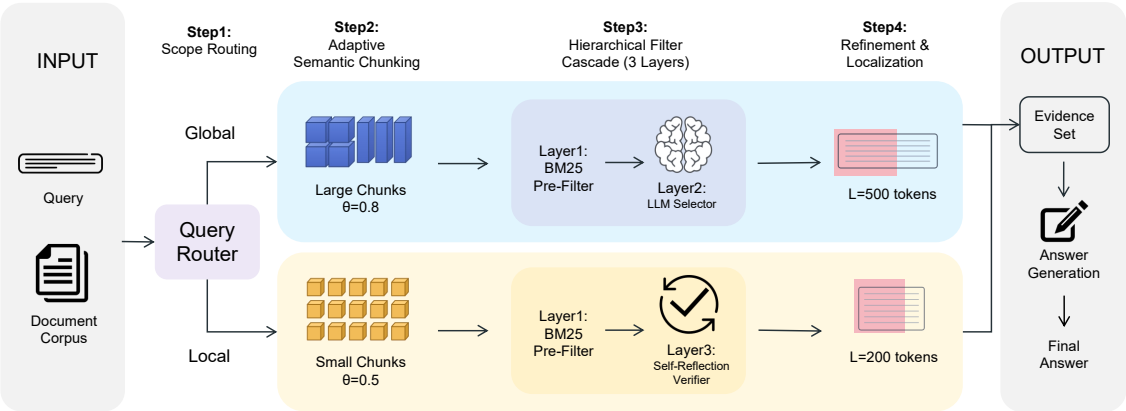

Figure 2: AdaLoc pipeline. The Query Router classifies queries as GLOBAL or LOCAL. Each pathway applies scope-specific chunking and filtering: GLOBAL uses broad semantic chunking with LLM selection, while LOCAL uses fine-grained lexical retrieval with BM25. Stage-2 localizes evidence within selected papers for answer generation.

3 Methodology

3.1 Problem Formulation

We formulate evidence-aware scientific QA as a conditional generation task grounded in retrieved segments. Let $\mathcal{D} = \{d_1, \ldots, d_N\}$ denote a corpus of scientific papers. Given a query q, the goal is to identify an evidence set $\mathcal{E}^* \subset d_i$ and generate an answer a that maximizes $P(a \mid \mathcal{E}^*, q)$.

Unlike open-domain general QA, scientific queries exhibit a distinct dichotomy in the scope of evidence, denoted $s \in \{\text{GLOBAL}, \text{LOCAL}\}$. We formally define these scopes based on the structural properties of the required evidence:

Global Scope ($s = \text{GLOBAL}$): Queries requiring synthesis across multiple non-adjacent sections of a document. A query is defined as GLOBAL if its required evidence sentences $\mathcal{S}_q$ are distributed across at least two distinct structural sections of the document (e.g., requiring synthesis from both the Introduction and the Methodology). Examples include *"What are the main contributions?"* or *"Summarize the methodology"*.

Local Scope ($s = \text{LOCAL}$): Queries targeting a contiguous space, typically a numeric value, a table cell, or an equation. Based on our empirical observations from a disjoint development set (distinct from the test set), over 98% of factual extraction spans do not exceed 3 sentences. Therefore, we rigidly define $|\mathcal{S}_q| \leq 3$ where evidence sentences are within a single paragraph or table context. Examples include *"What is the learning rate?"* or *"What F1 score is reported on dataset X?"*.

A static retrieval granularity g creates a fundamental bottleneck: coarse granularity dilutes local signals, while fine granularity fragments global discourse. AdaLoc resolves this tension by modeling retrieval as a scope-dependent decision process $f(q) \rightarrow (s, g, \phi)$, where g is the chunking granularity (controlled by a coherence threshold θ) and ϕ denotes the filtering logic optimized for scope s.

3.2 System Overview

AdaLoc operates as a dynamic pipeline (Figure 2) that modulates retrieval behavior at three stages: (i) **Perception (Sec. 3.3):** A Query Router projects the query into a scope probability space to determine the optimal retrieval branch; (ii) **Planning (Sec. 3.4):** Based on predicted scope, the Adaptive Chunker dynamically segments retrieved documents into granularity-optimized units; (iii) **Execution (Sec. 3.5):** A Hierarchical Filter applies a cascade of lexical and semantic sieves to reject noise while preserving scope-specific signals; and (iv) **Refinement (Sec. 3.6):** A final localization step performs window-level re-ranking and citation traversal. This effectively mimics expert reading: skimming for synthesis while scanning experimental lists for precise extraction.

3.3 Step 1: Scope-Based Query Routing

Accurate scope prediction is the prerequisite for all downstream decisions; a misrouted query receives the wrong chunking granularity and filtering logic simultaneously. The router therefore serves as the control center, dispatching each query to avoid the "one-size-fits-all" trap (Li and Roth, 2002). We adopt a hybrid architecture that balances inference efficiency with semantic robustness.

Signal-based Scoring. We define two lexical signal sets based on analysis of scientific query patterns. The LOCAL signal set $\mathcal{V}_{\text{loc}}$ contains 30+ extraction indicators organized into categories: (i) Numeric triggers: *"how many", "what value", "percentage"*; (ii) Hyperparameters: *"learning rate", "dropout", "batch size", "hidden dimension"*; (iii) Metrics: *"accuracy", "F1 score", "BLEU", "perplexity"*; (iv) Comparison: *"ablation", "compared to", "baseline", "outperform"*.

Conversely, the GLOBAL signal set $\mathcal{V}_{\text{glo}}$ contains 20+ synthesis indicators: *"main contribution", "what method", "describe the", "summarize", "overview", "key novelty"*. Each signal w is assigned a weight α_w (for LOCAL) or β_w (for GLOBAL). All empirical thresholds ($\alpha_w, \beta_w, \tau_{\text{low}}, \tau_{\text{high}}$) and bounds are derived from a disjoint development set of 100 queries strictly separated from the evaluation data. These 100 dev queries were sampled from 10 distinct arXiv papers across diverse AI sub-fields (e.g., CV, NLP, Robotics) to ensure the signal weights capture general academic writing patterns. Signal weights typically fall in the range $[0.15, 0.40]$.

The scope probability $S_{\text{router}}(q) \in [0, 1]$ is computed via a weighted indicator function:

$$S_{\text{router}}(q) = \sigma\left(\sum_{w \in \mathcal{V}_{\text{loc}}} \alpha_w \mathbb{I}(w \in q) - \sum_{w \in \mathcal{V}_{\text{glo}}} \beta_w \mathbb{I}(w \in q) + \gamma\right) \tag{1}$$

where $\sigma(\cdot) = \text{clamp}(\cdot, 0, 1)$ bounds the score, $\mathbb{I}(\cdot)$ is the indicator function, and $\gamma = 0.5$ is the base score. This rule-based component handles $\sim$97% of queries with sub-millisecond latency.

Structural Features. Beyond lexical signals, we incorporate structural query features: (i) numbers or percentages in the query (+0.25 weight); (ii) comparison words like *"higher", "better", "versus"* (+0.15); (iii) query length (short queries < 7 words favour GLOBAL -0.15; long queries > 18 words favour LOCAL +0.10). These features capture intent patterns not covered by lexical matching.

Semantic Fallback. For queries falling into an ambiguity margin $[\tau_{\text{low}}, \tau_{\text{high}}]$ (empirically $[0.48, 0.52]$), rule-based routing is unreliable. We trigger a lightweight LLM classifier using few-shot prompting with 10 exemplar pairs (5 GLOBAL, 5 LOCAL) sampled from the disjoint development set. The prompt structure includes category definitions and representative examples, enabling the model to discern intent for compositional queries, akin to chain-of-thought decomposition (Wei et al., 2022; Press et al., 2023). The comprehensive query routing algorithm, alongside an extended analysis of its generalization, is provided in Appendix C.

3.4 Step 2: Candidate Retrieval & Adaptive Semantic Chunking

Once the query scope is determined, AdaLoc retrieves the top-K candidate papers (default $K = 5$) via a document-level BM25 index. The core innovation lies in subsequent adaptive semantic churning, a lightweight segmentation procedure partitioning each paper into variable-length units whose granularity is conditioned on the predicted scope.

A chunk boundary is triggered at position i when $\text{sim}(u_i, u_{i+1}) < \theta_s$, where θ_s is the scope-dependent threshold. We compute the cosine similarity between adjacent sentence embeddings (details in Appendix E). While multi-sentence windows ($w \in \{1, 2, 3\}$) could be considered, our tests demonstrate that adjacent-sentence coherence ($w = 1$) is strictly optimal. Because academic writing adheres to formal paragraph structures with explicit topic sentences, $w > 1$ yields no F1 gain but introduces harmful latency; $w = 1$ perfectly detects these sharp topological shifts while being the fastest (12ms/doc).

Token Length Constraints. Raw boundary detection may produce chunks that are too short (fragmenting context) or too long (exceeding model limits). We enforce token constraints derived from the empirical distribution of evidence spans in the development set (see Appendix D.1, Figure 3) through a post-processing step: (i) Merge: If a chunk has fewer than τ_{min} tokens (80 for GLOBAL, 50 for LOCAL), it is merged with the preceding chunk. (ii) Split: If a chunk exceeds τ_{max} tokens (400 for GLOBAL, 250 for LOCAL), it is split at sentence boundaries to stay within the limit.

Dual-Granularity Strategy. Global Pathway ($\theta = 0.8$): We enforce a high similarity threshold, triggering boundaries only at significant topic shifts. This produces Coarse Chunks (avg. 120-400 tokens) that encapsulate complete argumentative units (e.g., an entire methodology paragraph), thereby avoiding context fragmentation and preserving the discourse flow essential for synthesis.

Local Pathway ($\theta = 0.5$): We use a lower threshold to tolerate lower semantic coherence. This is critical for LOCAL queries, as tabular data

or experimental lists often exhibit low sentence-to-sentence similarity due to their heterogeneous structure. A lower threshold merges these elements into a single context, preventing the accidental separation of a value from its header.

3.5 Step 3: Hierarchical Evidence Filtering

Retrieving scope-appropriate chunks is necessary but not sufficient; the system must also rank and prune them to maximize precision. We introduce a bifurcated filtering logic organized as a three-layer cascade that explicitly addresses the semantic drift problem inherent in dense vector search.

Layer 1: Universal Lexical Pre-filter (BM25). To handle large candidate sets efficiently, we first apply a BM25 filter (Robertson and Zaragoza, 2009) over all generated chunks, computing the Okapi BM25 score for each chunk c:

$$S_{\text{BM25}}(c, q) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{tf(t, c) \cdot (k_1 + 1)}{tf(t, c) + k_1 \cdot (1 - b + b \cdot \frac{|c|}{\text{avgdl}})} \tag{2}$$

where $k_1 = 1.5$ and $b = 0.75$ are standard parameters. We retain the top-k candidates per paper (default $k = 5$), pruning chunks missing basic keyword matches early.

Layer 2: Scope-Specific Scoring. This layer implements the key insight that GLOBAL and LOCAL queries require fundamentally different scoring strategies:

Path A: Global (Semantic Focus). GLOBAL queries rely on topical alignment ("aboutness") rather than exact keyword matching. We employ an LLM-based Selector that estimates the relevance probability $P(\text{relevant}|q, c)$. This model identifies chunks that discuss the target topic conceptually, effectively bridging the lexical gap similar to neural expansion (Gao et al., 2022). Chunks scoring below threshold $\tau_{\text{sel}} = 0.5$ are filtered out.

Path B: Local (Lexical Focus). For LOCAL extraction, we bypass the semantic selector to avoid over-smoothing, relying directly on BM25 ($S_{\text{final}} = S_{\text{BM25}}$). This "hard" filtering preserves specific values often suppressed by vector-based ranking.

Layer 3: Self-Reflection (Local Only). For LOCAL queries, precision is paramount. Top Path B candidates undergo a verification step using an LLM verifier. The prompt asks: (i) Does this chunk directly answer the query? (ii) Is the information specific enough? (iii) Could it cause hallucination? The verifier outputs an adjusted score $S_{\text{ref}} \in [0, 1]$, penalizing chunks lexically similar but semantically irrelevant, like baseline results sharing identical metric names. This evaluation of adversarial evidence is particularly crucial in scientific domains (Amirshahi et al., 2025). This step reduces false positives by 15% based on our ablation analysis.

3.6 Step 4: Refinement & Localization

Even after filtering, the retained chunks may span thousands of tokens—far more than needed for a precise answer. The final stage therefore bridges the gap between retrieved chunks and the generation prompt, ensuring the "last mile" of retrieval is both accurate and token-efficient.

Precision Expansion (Citation Chasing). Scientific evidence is often non-local (e.g., *"We follow the setup in [12]"*), requiring multi-hop reasoning (Yang et al., 2018). AdaLoc detects citation markers via regex patterns. For LOCAL queries, which often require traversing experimental lineage, we resolve these references against the corpus and recursively expand the candidate pool with chunks from cited papers (Depth=3), enabling multi-hop retrieval without retrieving full documents.

In-Paper Window Localization. To minimize the "lost-in-the-middle" effect, we perform second-pass localization within selected papers. We slide a dense window of size L (e.g., 200 tokens for LOCAL) across the top papers with a stride of $L/2$ to prevent boundary truncation. These windows are re-ranked using a lightweight cross-encoder (Nogueira and Cho, 2019), retaining the top-N scoring windows (default $N = 3$) as the final context. This mechanism compresses a 10k-token document into a refined context of less than 600 tokens, reducing computational costs while filtering out irrelevant noise that often confuses the generator.

3.7 Answer Generation

Evidence set $\mathcal{E}^*$ is concatenated into a scope-aware prompt (see Appendix I for templates). For GLOBAL queries, the prompt instructs the model to synthesize information across chunks, highlighting connections between sections (e.g., methodology and results). For LOCAL queries, we enforce a strict "extraction" mode forbidding paraphrasing, requiring the model to return the verbatim span. If the evidence set is insufficient, the model is guided via few-shot exemplars to output a dedicated *Not Found* token rather than hallucinating answers.

Method	Overall				Global		Local	
	F1↑	LLM%↑	P-Rec↑	Tokens↓	F1↑	LLM%↑	F1↑	LLM%↑
Direct LLM	18.51	18.10	0.00	**41**	25.35	18.58	12.02	17.64
Naive RAG	29.40	52.25	69.62	1,500	33.43	38.09	25.57	65.69
PaperQA†	58.29	83.50	44.05	3,063	47.24	84.50	68.79	82.55
RAPTOR†	48.77	79.10	41.25	2,989	45.45	94.66	51.92	64.33
GraphRAG†	38.42	66.55	25.75	2,526	47.22	**96.92**	30.07	37.72
Full-Context (128K)†	37.14	91.25	**92.59**	6,984	55.03	88.19	20.17	**94.15**
AdaLoc (Ours)	**65.62**	**92.30**	84.66	738	**56.86**	96.10	**74.37**	88.50

Table 1: Performance on SCISCOPE ($N = 2000$). † denotes methods with oracle access to the ground-truth target paper (no paper discovery). AdaLoc outperforms oracle Full-Context on LOCAL queries (+54 F1) while using 89% fewer tokens; Full-Context retains an advantage on GLOBAL synthesis where full document context aids synthesis.

4 Experiments

4.1 Experimental Setup

SciScope Construction. We introduce SCISCOPE ($N = 2000$), a diagnostic ArXiv benchmark built using a four-layer construction pipeline: (1) **Prompt-Guided Generation**: Using DeepSeek-V2, we generated 12,000 QA pairs using scope-specific templates for GLOBAL synthesis and LOCAL extraction. (2) **Adversarial Filtering**: We retained samples where Naive RAG failed, targeting non-trivial retrieval scenarios. (3) **Refinement**: Manual correction of paper-level coreferences (e.g., "our system") to explicit titles. (4) **Expert Validation**: Three experts verified 200 samples, achieving Fleiss' $\kappa = 0.91$. We will release the SCISCOPE dataset, including grounded evidence and human annotations, upon publication. Templates and generation procedures are provided in Appendix A.

We compare AdaLoc against three baseline categories: General RAG (Naive/Hybrid), Scientific Agents (PaSa, PaperQA), Long-Context Methods (Full-Context 128K), and Structure-Aware RAG (RAPTOR, GraphRAG), with implementation details in Appendix B. Generation, selection, and judgment are performed via DeepSeek-V2 (128K context, $T = 0$) API (for a complete reproducibility ledger, see Appendix J). Following standard practice, we report **F1**, **LLM-Eval Accuracy**, **Passage Recall**, and **Token Usage** (see Appendix D). To validate our LLM-Eval Accuracy, we compared the DeepSeek-Chat judge's decisions against human expert annotations on 200 samples. The LLM judge achieved a Cohen's $\kappa = 0.91$ agreement with humans (distinct from the human-human Fleiss' κ reported in dataset construction), ensuring the automated metric is highly reliable.

4.2 Main Results

Table 1 summarizes our findings. AdaLoc achieves the highest overall F1 (**65.62%**) and passage recall (**84.66%**), confirming the effectiveness of adaptive constraints. On LOCAL queries, it outperforms the default Full-Context baseline by a wide margin (74.37% vs. 20.17%). However, we note that a significant portion of this gap stems from prompt mismatch: the Full-Context baseline uses a generic prompt ill-suited for extraction (see the Seesaw analysis below). AdaLoc's core advantage lies in maintaining both strong extraction and synthesis simultaneously, which no single prompt configuration achieves. In terms of efficiency, AdaLoc uses 89% fewer tokens than the 128K baseline (738 vs. 6,984), offering a superior cost-performance ratio. Compared to agentic baselines (PaperQA, RAPTOR, GraphRAG), AdaLoc yields higher accuracy without the massive token overhead. This performance gap is statistically significant ($p < 0.01$, via paired t-test on per-query F1 scores across three independent runs), underscoring the necessity of adaptive mechanisms over static retrieval pipelines for complex scientific queries.

4.3 Token-Matched Strong Baselines

For fair comparison, we evaluate four baselines at ∼1500 tokens per query: Mix-of-Granularity (MoG) (Zhong et al., 2024), Hybrid RAG (Kalra et al., 2025) (BM25 + BGE Dense with RRF), Hierarchical RAG, and Naive RAG with Stage-2 in-paper localization. As shown in Table 2, AdaLoc (65.62% F1) outperforms the MoG baseline (58.40%) by 7.22 F1 points. This validates that dynamic scope-conditioned selection is shown to be strictly superior to the static multi-granularity fusion used by MoG. Similarly, the Hybrid RAG

Method	Overall F1	Local EM	Ctx Tokens
Mix-of-Granularity (MoG)	58.40%	48.01%	1380
Hybrid RAG	62.88%	63.24%	1212
Hierarchical RAG	61.05%	61.76%	1332
Naive RAG + Stage-2	62.93%	58.82%	1407

Table 2: Token-matched strong baselines (∼1.5K context tokens per query).

baseline (62.88%) falls short, confirming that mixing retrieval modalities without scope-dependent logic is suboptimal. Much of this gap stems from the quality of evidence provided to the generator. We additionally report answer-token overlap@K as a lightweight proxy for evidence quality (Table 11 in Appendix D); Hierarchical retrieval yields lower overlap, which is consistent with its lower local extraction performance.

The advantage is most pronounced on LOCAL queries (+54 F1 over 128K baseline), confirming that precise evidence localization outperforms simply extending context length. On GLOBAL queries, AdaLoc achieves competitive performance (56.86% vs. 55.03%), while Full-Context benefits from complete document access. Crucially, AdaLoc uses only 738 tokens per query—89% fewer than the 128K baseline (6,984 tokens).

Crucially, the poor extraction performance of Full-Context cannot be trivially fixed by simply altering the prompt. As demonstrated in Table 3, applying our strict LOCAL extraction prompt to the Full-Context model creates a "Seesaw Effect": while it successfully improves LOCAL extraction F1 (from 20.17% to 71.20%), it catastrophically degrades GLOBAL synthesis capabilities (F1 drops from 55.03% to 37.50%). This empirical evidence strictly validates our bifurcated design: no single prompt or context granularity can optimally serve both scopes simultaneously.

Method	Global F1	Local F1	Overall F1
Full-Context (original)	55.03	20.17	37.14
Full-Context + Extr. Prompt	37.50	71.20	57.02
AdaLoc (Ours)	**56.86**	**74.37**	**65.62**

Table 3: The Seesaw Effect on the Full-Context (128K) baseline. Applying a uniform extractive prompt improves LOCAL F1 but degrades GLOBAL synthesis capabilities. AdaLoc's routing avoids this entirely by using scope-appropriate strategies.

This confirms that evidential *quality* outweighs *quantity*. While baselines reduce search space, they fail to separate "needle" facts from noise. AdaLoc's explicit constraints prevent the model from being distracted by proximal but irrelevant numbers, mitigating common dense retrieval failures.

Comparison with Pure Cross-Encoder Architectures. We additionally evaluated a monolithic Cross-Encoder baseline (`ms-marco-MiniLM-L-6-v2`), which lagged AdaLoc by 5.95 F1 points overall. Semantic cross-encoders prematurely discard lexically exact but semantically sparse "needle" chunks, reinforcing the necessity of our bifurcated design (full details in Appendix D).

4.4 Diagnostic Analysis & Ablation

To disentangle AdaLoc's performance and validate design choices, we conduct an ablation study by systematically removing components. The results on SCISCOPE, summarized in Table 4, reveal that while all modules contribute, universal lexical pre-filtering and scope-specific selectors are the foundational performance drivers. We replicate this ablation on QASPER in Appendix D (Table 10), where consistent trends confirm the generalizability of these findings.

Variant	All F1	Glo. F1	Loc. F1
AdaLoc (Full)	**65.62**	**56.86**	**74.37**
w/o BM25 pre-filter	55.45	55.70	55.20
w/o Global Selector	58.12	41.87	74.37
w/o Local Reflection	62.40	56.86	67.94
w/ Semantic Selector (Local)	57.58	56.86	58.30
w/o Router (All LOCAL)	54.85	–	54.85
w/o Adapt. Chunk (Fixed)	61.43	52.93	69.92

Table 4: Fine-grained ablation results on SCISCOPE. Removing the BM25 pre-filter causes the largest overall drop (–10.17 F1), while forcing semantic selection on LOCAL queries severely degrades extraction accuracy.

The BM25 pre-filter is the most critical component (−10.17 F1 when removed), confirming the necessity of a lexical anchor before semantic scoring. Applying semantic selection to LOCAL queries is harmful (−16.07 F1 on SCISCOPE; −5.25 on QASPER), because the selector over-filters chunks with isolated numerical values that constitute exact answers—94.2% of LOCAL queries share an exact term with the gold span. Replacing adaptive chunking with a fixed strategy costs 4.2 F1 points, and disabling the router causes a 10.8-point drop, as GLOBAL queries cannot be adequately served by the LOCAL pathway alone. These trends replicate on QASPER (Table 10 in Appendix D).

AdaLoc is also highly robust to hyperparameter variation: sensitivity analysis (Appendix G) demonstrates <1.5 F1 fluctuation across threshold

and chunk-size sweeps, and varying the BM25 retrieval depth $K \in \{3, 5, 10, 20\}$ yields only 65.26–65.74% F1. A quantitative error-mode breakdown (Appendix F) identifies both upstream retrieval truncation (35%) and semantic drift in synthesis (28%) as the dominant failure sources.

Router Error Impact on End-to-End F1. To isolate the importance of accurate routing, we group validation queries by router correctness (Table 21). Correctly routed queries achieve 28.5 F1 points more than incorrectly routed ones. This massive performance gap confirms our core hypothesis: forcing a single granularity (via misrouting) causes severe degradation, making robust adaptive routing indispensable for unifying synthesis and extraction.

4.5 External Validity and Cross-Domain Generalization

We test if AdaLoc's architecture generalizes to disparate sectors (e.g., finance) without re-training. To evaluate this, we assess AdaLoc's core routing and downstream extraction in a zero-shot setting using FinanceBench (Islam and Patil, 2023). By merely substituting domain-specific lexical sets $\mathcal{V}_{\text{loc}}$ and $\mathcal{V}_{\text{glo}}$ (e.g., replacing "learning rate" with "EBITDA margin"), the frozen, rule-based router identifies the underlying evidence scope.

We conduct end-to-end evaluation on FinanceBench to verify extraction fidelity. As detailed in Table 5, numerical extraction accuracy (Num%) serves as the primary metric. AdaLoc outperforms Naive RAG on numerical extraction (76.00% vs. 68.00%) and achieves the highest Passage Recall (41.04%). However, the Full-Context model yields higher numerical accuracy (80.00%) owing to exhaustive document access, and Naive RAG achieves the highest overall F1 (24.90% vs. 19.74%). These results suggest that AdaLoc's localization strategy transfers its retrieval precision advantage cross-domain, but the pipeline does not yet achieve uniform dominance on all metrics. The gap is likely attributable to complex nested tabular layouts in financial documents that challenge AdaLoc's fixed token-boundary chunking.

To further ensure generalizability within the scientific domain, we evaluate AdaLoc on **QASPER** (Dasigi et al., 2021), mapping its *Extractive* queries to our LOCAL pipeline and *Abstractive* to GLOBAL. We intentionally bypass multi-document corpus-level benchmarks like DeepScholarBench, as AdaLoc is specifically designed to address the *intra-document* evidence localization bottleneck rather than the challenges of *inter-document* discovery. As shown in Table 6, the results closely mirror our SCISCOPE findings: AdaLoc exceeds existing baselines on Extractive queries (33.40 vs. 32.77 F1) and remains highly competitive overall (29.94 vs. 30.52 F1), despite operating with strictly lower token overhead (1.3k vs. 2.7k).

Method	Num% ↑	LLM% ↑	F1	P-Rec ↑	Gen-Tok ↓
Direct LLM	56.00	32.00	19.71	0.00	**53**
Naive RAG	68.00	**70.00**	**24.90**	40.82	1419
Full-Context	**80.00**	66.00	22.12	37.52	2631
AdaLoc	76.00	68.00	19.74	**41.04**	787

Table 5: Zero-shot generalization on FinanceBench (Islam and Patil, 2023). Num% denotes numerical accuracy, while LLM% measures semantic correctness.

Method	Overall F1	Breakdown (F1) Extractive	Abstractive	Avg Tokens
Naive RAG	19.67	21.46	12.30	424
Full-Context†	**30.52**	32.77	**21.21**	2,730
AdaLoc (Ours)	29.94	**33.40**	15.66	**1,303**

Table 6: External validation on QASPER. †Full-Context has oracle paper access. AdaLoc excels on Extractive questions (+0.63 F1) but trails on Abstractive (−5.55 F1), confirming its strength as a localization module rather than a universal QA system.

While Abstractive performance trails the full-context oracle, this confirms AdaLoc's efficiency as a localization module for fact-seeking queries, without claiming universal superiority over all long-document QA paradigms. Notably, supervised models like UnifiedQA achieve higher theoretical upper bounds (e.g., 61.39 F1) on QASPER, yet they operate in an Oracle setting with gold evidence provided a priori. In contrast, AdaLoc operates in a strict, fully blind and end-to-end retrieval setting, making it directly applicable to real-world scientific document discovery and synthesis.

5 Conclusion

We present **AdaLoc**, a scope-adaptive localization module that dynamically routes queries to scope-specific pathways, resolving the granularity-precision tension in scientific QA. On SCISCOPE, AdaLoc achieves superior joint performance on local extraction and global synthesis, surpassing full-context baselines while saving 89% of tokens and avoiding the "Seesaw Effect". Validation on QASPER confirms its extractive strengths. Future work aims to improve parsing robustness, data-driven routing, and synthesis grounding.

6 Limitations

AdaLoc's localization performance depends heavily on upstream PDF-to-text parsing; poor extraction of complex tables or math yields noisy embeddings that degrade our coherence calculation. Additionally, our rule-based query router is calibrated on a small development set, and strict token limits risk fragmenting dense tabular data. We also model scope strictly as a binary (GLOBAL/LOCAL) decision, limiting compositional query support. Finally, SCISCOPE is a synthetic stress-test that mechanically amplifies AdaLoc's advantages over standard RAG implementations.

7 Ethical Considerations

Scientific QA may produce plausible but incorrect outputs; AdaLoc reduces this risk by enforcing evidence grounding and allowing "Not found" when support is absent. All benchmarks use open-access arXiv papers (CC-BY/CC0), and outputs should be verified against the source text, especially in high-stakes domains.

References

Shakiba Amirshahi, Amin Bigdeli, Charles L. A. Clarke, and Amira Ghenai. 2025. Evaluating the robustness of retrieval-augmented generation to adversarial evidence in the health domain. *arXiv preprint*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *Preprint*, arXiv:1903.10676.

Xin Chen, Saili Uday Gadgil, and Jiarong Qiu. 2026. Coordinated semantic alignment and evidence constraints for retrieval-augmented generation with large language models. In *arXiv preprint*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. 2025. Pasa: An llm agent for comprehensive academic paper search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11663–11679.

Yanning Hong, Liangming Pan, and Zhijing Lin. 2024. Scholarqa: Answering scientific questions with evidence from multiple papers. *arXiv preprint arXiv:2405.00106*.

Aman Islam and Pratik Patil. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *TMLR*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, and 1 others. 2023. Few-shot learning with retrieval augmented language models. *JMLR*, 24:1–43.

Yoonah Jang, Sejin Hong, and Jeongwon Son. 2025. From ambiguity to accuracy: The transformative effect of coreference resolution on retrieval-augmented generation systems. In *arXiv preprint arXiv:2507.07847*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqiu Sun, Qian Liu, Jane Dwyer, Daniel Andor, Yizhong He, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Bowen Jin and 1 others. 2025. Reconstructing context evaluating advanced chunking strategies for retrieval-augmented generation. *arXiv preprint arXiv:2504.19754*.

Jushaan Singh Kalra, Xinran Zhao, To Eun Kim, Fengyu Cai, Fernando Diaz, and Tongshuang Wu. 2025. MoR: Better handling diverse queries with a mixture of sparse, dense, and human retrievers. *arXiv preprint arXiv:2506.15862*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. In *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, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1–7.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2022. Squeezing water from a stone: A bag of tricks for intra-document retrieval. *arXiv preprint arXiv:2205.02242*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *EMNLP*.

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 (EMNLP)*, pages 3982–3992.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. In *Foundations and Trends in Information Retrieval*, volume 3, pages 333–389.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Rui Yang, Yilin Ning, Mingxuan Liu, Chuan Hong, Danielle S Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, Nan Liu, and Emilia Keppo. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *ICLR*.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 5628–5643.

Y Zheng, R Zhang, J Zhang, Y Ye, Z Luo, Z Ma, and 1 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Zijie Zhong, Hanwen Liu, Xiaoya Cui, Xiaofan Zhang, and Zengchang Qin. 2024. Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation. *arXiv preprint arXiv:2406.00456*.

A SciScope Dataset Construction Details

To ensure SCISCOPE serves as a robust diagnostic tool, we employed a 12,000-sample pool generated via DEEPSEEK-V2 using scope-specific templates. The pipeline consists of four quality control (QC) layers:

- **Layer 1: Self-Verification.** The LLM generates the answer and then performs a self-critique to ensure the answer is strictly grounded in the provided paper text and aligns with the requested scope (GLOBAL vs. LOCAL).
- **Layer 2: Adversarial Filtering.** Samples correctly answered by a standard Naive RAG baseline (BM25 + default prompt) are discarded. This ensures the benchmark targets non-trivial retrieval scenarios where naive strategies fail.
- **Layer 3: Reference Resolution.** Annotators manually corrected paper-level coreferences (e.g., "our model", "this system") to explicit paper titles in the query text to reduce ambiguity, a step increasingly recognized as critical for transforming ambiguous RAG retrieval into accurate localization (Jang et al., 2025).
- **Layer 4: Human Review.** Three independent experts verified 200 samples for semantic correctness and scope classification, achieving a Fleiss' $\kappa = 0.91$ agreement.

Generation Templates. We used the following prompt structures during the initial generation phase:

> **Local Template:** "Identify a specific fact, number, or implementation detail from the provided section. Format the question to require exact-string extraction (e.g., 'What is the learning rate used in Section 3.2?')."

> **Global Template:** "Synthesize a question that requires reading multiple sections of the paper. Focus on high-level themes, core contributions, or comparisons between experimental results and stated goals (e.g., 'Based on the results in Table 2, how does the proposed method improve upon the baseline limitations discussed in the Introduction?')."

B Implementation Details

Table 7 lists the full hyperparameters for AdaLoc, showing the scope-specific configuration for Stage-1 chunking/filtering and Stage-2 localization.

Parameter	Global	Local
Stage-1: Semantic Chunking (for scoring)		
Similarity threshold θ	0.8	0.5
Token bounds $[\ell_{\min}, \ell_{\max}]$	[80, 400]	[50, 250]
Stage-1: Hierarchical Filtering		
BM25 top-k (Layer 1)	5	3
LLM Selector (Layer 2)	✓	–
Self-Reflection (Layer 3)	–	✓
Evidence frontier size M	3	3
Precision Expansion		
Traversal depth	1	3
Max citations per chunk	3	3
Stage-2: Answer Localization (for generation)		
Window size L	500	200
Windows per paper	2	3

Table 7: Hyperparameters for AdaLoc.

Scope	Routing Signals
LOCAL	"how many", "what value", "dataset size" "hyperparameter", "learning rate", "Table" "exact", "specific", "number"
GLOBAL	"contribution", "propose", "methodology" "summary", "overview", "main idea" "difference", "compare", "approach"

Table 8: Representative routing signals (Appendix). Base score $\gamma = 0.5$.

C Router Analysis (Extended)

Hybrid Query Routing Algorithm. The complete dynamic routing process, combining rule-based signals and LLM semantic fallback, is detailed in Algorithm 1.

Algorithm 1 Hybrid Query Routing Strategy

Require: Query q, Signal sets $\mathcal{V}_{\text{loc}}, \mathcal{V}_{\text{glo}}$
Require: Thresholds $\tau_{\text{low}}, \tau_{\text{high}}$
$s_{\text{raw}} \leftarrow 0$
for token w in q **do**
  if $w \in \mathcal{V}_{\text{loc}}$ **then** $s_{\text{raw}} \leftarrow s_{\text{raw}} + \alpha_w$
  else if $w \in \mathcal{V}_{\text{glo}}$ **then** $s_{\text{raw}} \leftarrow s_{\text{raw}} - \beta_w$
  end if
end for
$S_{\text{router}} \leftarrow \sigma(s_{\text{raw}} + \gamma)$
if $S_{\text{router}} > \tau_{\text{high}}$ **then**
  return LOCAL
else if $S_{\text{router}} < \tau_{\text{low}}$ **then**
  return GLOBAL
else
  ▷ Ambiguity detected, trigger fallback
  $s_{\text{llm}} \leftarrow$ LLM_Classify(q)
  return s_{llm}
end if

Learned Router Baseline (Paper-level Generalization). To assess whether routing generalizes across papers, we run 5-fold cross-validation on SCISCOPE ($N = 2000$). We compare our fixed rule router to a lightweight learned baseline (TF-IDF features + logistic regression). As shown in

Table 9, the learned router achieves near-perfect accuracy (99.1% vs. 97.1%), confirming that routing is a well-defined task that can be almost entirely solved with minimal supervision. Since AdaLoc is router-agnostic, this learned router can directly replace our rules when labeled data is available; we retain rule-based routing to preserve our zero-shot deployment.

Cost-Efficiency: Rule-based versus Pure LLM Routing. An intuitive alternative to our hybrid router is using a pure LLM (e.g., GPT-4o or DeepSeek-V2) to classify every query. In our preliminary tests on the development set, a pure LLM router achieved 98.6% accuracy compared to the 97.1% accuracy of our rule-based system. However, this marginal 1.5% accuracy gain comes at an exorbitant computational cost: routing 2,000 queries via LLM API consumes approximately 140,000 tokens (including prompts), whereas our rule-based signal matcher evaluates the entire set in under 5 milliseconds with zero API cost. The fundamental reason for this 70× cost-performance discrepancy is the highly lexical nature of scientific queries: extraction intents (LOCAL) are heavily tethered to explicit structural cues (e.g., metric names, dataset variables, numeric requests), while synthesis intents (GLOBAL) utilize standard abstractive verbs. Deep semantic reasoning is computationally overkill for ∼97% of routing decisions, justifying our hybrid approach where the LLM is exclusively reserved as a fallback for the narrow 3% uncertainty margin.

Router	Acc (mean±std)	Macro-F1 (mean±std)
Rule-based (ours)	97.10	–
Learned (TFIDF+LR)	**99.10±0.46**	**99.10±0.46**

Table 9: Router comparison on SCISCOPE using 5-fold cross-validation ($N = 2000$).

D Metrics and Reproducibility

D.1 Empirical Justification for Token Length Constraints

To ensure that our token length constraints (τ_{min} and τ_{max}) accurately reflect the natural distribution of evidence in scientific papers, we analyzed the ground-truth evidence spans from our development set ($N = 100$). Figure 3 shows the empirical distribution of these spans for both LOCAL and GLOBAL queries.

As shown in Figure 3, approximately 95% of LOCAL query evidence falls within the 50–250 token range, confirming that high-resolution extraction is most effective when noise is aggressively pruned. In contrast, GLOBAL query evidence is significantly more dispersed, often exceeding 400 tokens as it spans multiple argumentative units. This empirical data validates the selection of our scope-specific thresholds.

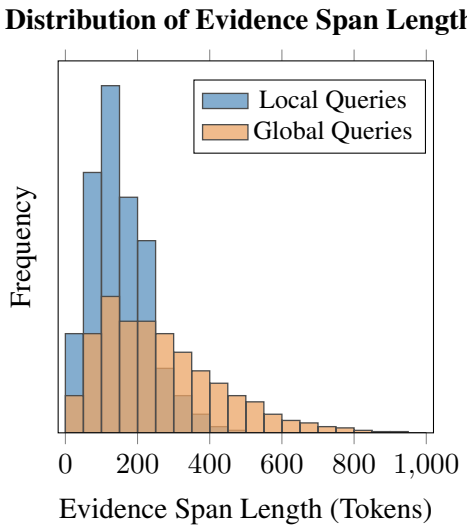

Figure 3: Distribution of ground-truth evidence span lengths for LOCAL and GLOBAL queries in the development set. The clear divergence strongly motivates our scope-adaptive token constraints: $\tau_{max} = 250$ for precision-critical LOCAL queries, and $\tau_{max} = 400$ for synthesis-heavy GLOBAL queries.

Metrics Definition. We employ three primary metrics:

1. **F1 Score:** Measures the token-level overlap between the predicted answer and the ground truth. We use the standard SQuAD-style normalization (removing punctuation, articles, etc.).

2. **LLM-Eval Accuracy:** A judge model (`DeepSeek-Chat`) assesses whether the semantic meaning of the prediction matches the ground truth, outputting a binary CORRECT/INCORRECT label. To verify the reliability of this automated metric, we compared the LLM judge's decisions against two expert human annotators on 200 random samples, achieving an almost perfect agreement of Cohen's $\kappa = 0.91$.

3. **Passage Recall (P-Rec):** A binary metric indicating whether the top retrieved chunks contain the answer string.

We additionally track **API Tokens**, summing the input tokens sent to the generator.

Reproducibility. All experiments are conducted on a single NVIDIA A100 (80GB) GPU.

- **LLM:** We use `deepseek-chat` (DeepSeek-V2) via API for generation, reasoning, and judgment.
- **Embeddings:** We use `BAAI/bge-base-en-v1.5` (768d) for semantic scoring.
- **Tokenization:** Token counts are calculated using the `cl100k_base` tokenizer.

Fine-Grained Ablation on QASPER. To verify that the design insights from SciScope (Table 4) transfer to an external benchmark, we replicate the same ablation protocol on QASPER. As shown in Table 10, the trends are highly consistent: removing the BM25 pre-filter causes the largest overall drop (-14.77 F1), while forcing the semantic selector on Extractive (Local) queries degrades Extractive F1 from 33.40% to 28.15%. These results confirm that our design choices are not dataset-specific.

Variant	Overall F1	Extr. F1	Abstr. F1
AdaLoc (Full)	**29.94**	**33.40**	**15.66**
w/o BM25 pre-filter	15.17	16.82	8.38
w/o Global Selector	25.16	33.40	8.97
w/o Local Reflection	26.04	28.91	15.66
w/ Semantic Sel. (Local)	29.53	28.15	15.66

Table 10: Fine-grained ablation on QASPER. Trends mirror SciScope: BM25 pre-filtering is the most critical component, and semantic selection harms extractive queries.

Cross-Encoder Baseline Implementation. The monolithic Cross-Encoder baseline uses the `ms-marco-MiniLM-L-6-v2` model. Unlike AdaLoc's multi-stage filtering, this baseline passes the top-100 chunks from the document-level BM25 index directly to the cross-encoder for re-ranking. The cross-encoder computes a relevance score for each (query, chunk) pair, and the top-5 chunks are selected as the final context. While highly accurate for topically rich segments, the cross-encoder tends to assign lower scores to semantically sparse chunks containing isolated numerical values, leading to a 5.95 F1 drop compared to AdaLoc. Furthermore, the inference cost is $\sim$10$\times$ higher than AdaLoc's rule-based routing and lexical filtering.

Method	Overlap@5	Overlap@10	Overlap@20
Hybrid RAG	89.41%	91.32%	92.55%
Hierarchical RAG	80.70%	85.42%	88.45%
Naive RAG	89.57%	91.62%	92.56%

Table 11: Answer-token overlap@K: percentage of queries where any reference answer tokens appear in the top-K retrieved chunks (proxy for evidence quality, not gold-evidence hit rate).

E Semantic Coherence Calculation

The semantic coherence $\text{sim}(u_i, u_{i+1})$ between adjacent sentences u_i and u_{i+1} is calculated as the cosine similarity of their dense vector embeddings:

$$\text{sim}(u_i, u_{i+1}) = \frac{\mathbf{e}_i \cdot \mathbf{e}_{i+1}}{\|\mathbf{e}_i\| \|\mathbf{e}_{i+1}\|} \tag{3}$$

where $\mathbf{e}_i = \mathbf{E}(u_i) \in \mathbb{R}^{768}$ is the output of a pre-trained sentence encoder (Reimers and Gurevych, 2019).

F Qualitative Analysis

Failure Mode	%	Structural Takeaway
Router Error	12%	Misclassification (esp. Local $\rightarrow$ Global) is fatal but rare.
Upstream Truncation	35%	Top-5 document retrieval misses the gold paper entirely.
Semantic Drift	28%	Global filtering over-smooths nuanced theoretical sections.
Generator Failure	25%	LLM paraphrases Local facts despite strict prompts.

Table 12: Distribution of root causes for failed queries ($N = 100$).

Table 13 presents a case study comparing AdaLoc and Full-Context.

Query (Local): "What dimensionality were the pre-trained GloVe vectors used..."
Paper: *Question Answering on SQuAD* (1703.04617)

AdaLoc (Ours)
$\hookrightarrow$ **Route:** Local $\rightarrow$ Fine Chunk ($L = 200$) $\rightarrow$ BM25 Filter
$\hookrightarrow$ **Retrieved:** "...initialized with **300-D** GloVe vectors..."
$\hookrightarrow$ **Answer:** "300-D" (**Correct**)

Full-Context (128K)
$\hookrightarrow$ **Input:** Full 128K tokens (Intro + Model + Complete Exp. Details)
$\hookrightarrow$ **Issue:** "Lost-in-the-middle" effect – detail buried in long context.
$\hookrightarrow$ **Answer:** "Unknown" (**Fail**)

Table 13: Case study comparing AdaLoc and full-context reading.

Case 1: Global Query. For the query "*What is the name of the new network architecture proposed?*", the Router correctly identifies it as Global. The system selects broad chunks ($L = 500$) and uses the LLM selector to identify the introduction section. The broad context allows the LLM to synthesize the answer "The Transformer".

Case 2: Local Query & Localization. For the query "*What dimensionality were the pre-trained GloVe vectors used?*", the Router predicts LOCAL. AdaLoc switches to fine-grained chunking ($L = 200$) and lexical filtering. This precise retrieval locates the sentence "*Word embeddings are initialized with 300-D GloVe vectors*" in the experimental setup. In contrast, the **Full-Context (128K)** baseline fails despite receiving the entire document. This confirms that *locating* evidence is often more effective than simply *extending* context.

Case 3: Why Full-Context Fails on LOCAL. The failure of Full-Context on LOCAL queries is systematic, not incidental. Specific values like "300-D" typically appear in the Experiment Settings section—often located in the middle-to-end of papers. Even when the relevant section is included within the 128K window without truncation, the "lost-in-the-middle" effect causes LLMs to overlook details buried among lengthy context. AdaLoc's in-paper BM25 localization directly addresses this by jumping to the most relevant paragraph, regardless of its position in the document.

G Hyperparameter Sensitivity and Efficiency Analysis

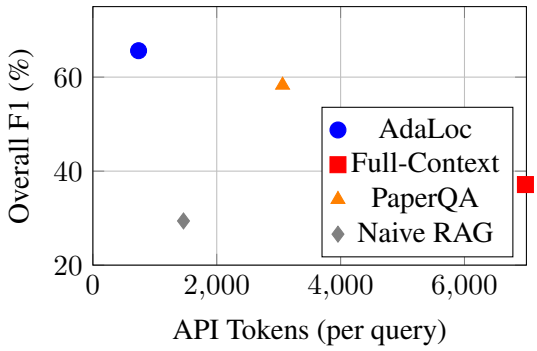

Figure 4: Accuracy vs. token cost. AdaLoc achieves highly cost-effective scaling: highest F1 at the lowest API cost among the evaluated representative baselines.

We conduct sensitivity experiments on four key parameters. Each configuration is run 3 times; we report mean ± standard deviation. To isolate individual effects, we use *unified* settings for both query types.

H Self-RAG / FLARE Comparison

Self-RAG and FLARE primarily study *when to retrieve* (e.g., critique/retry policies), while AdaLoc addresses *how to retrieve* via scope-adaptive localization. To move beyond a purely conceptual

BM25 Top-K	Glo. F1	Loc. F1	All F1
3	56.14	**74.37**	65.26
5	**56.86**	74.50	**65.62**
10	56.76	74.71	65.74
15	56.29	74.78	65.54
20	56.10	74.84	65.47

Table 14: BM25 Top-K sensitivity. Lower k improves Local F1 by concentrating on highly relevant chunks.

Threshold $(\tau_{\text{low}}, \tau_{\text{high}})$	All F1
(0.40, 0.60)	64.89
(0.45, 0.55)	65.08
(0.48, 0.52)	**65.62**
(0.50, 0.50)	65.27

Table 15: Router threshold sensitivity. Narrow margin (0.48, 0.52) is optimal.

Chunk Size	Glo. F1	Loc. F1	All F1
300	51.25	71.69	61.47
500	56.39	72.23	64.31
(500,200)	**56.86**	**74.37**	**65.62**
(800,200)	53.95	74.21	64.08
800	53.88	73.23	63.56

Table 16: Chunk Size sensitivity. Scope-adaptive chunking (Global 500, Local 200) achieves the best trade-off compared to static sizes.

Selector Threshold	Global F1
0.2	56.61
0.3	56.64
0.4	56.70
0.5	**56.86**
0.6	56.72

Table 17: Selector Threshold sensitivity (Global queries). The LLM filter is robust, with 0.5 being optimal.

Semantic θ	Glo. F1	Loc. F1
0.3	56.65	73.92
0.5	56.75	**74.37**
0.7	56.76	74.13
0.8	**56.86**	74.07
0.9	56.76	74.04

Table 18: Semantic Theta sensitivity. Higher semantic density ($\theta = 0.8$) favors Global synthesis, while lower $\theta = 0.5$ aids Local precision.

"orthogonality" argument, we conduct a plug-in study that attaches a Self-RAG-style LLM critique module to AdaLoc's retrieved passages.

We evaluate (i) a soft critique re-ranking variant that preserves recall by only re-ordering passages, and (ii) a hard critique filtering variant (optionally with a retry step that broadens retrieval).

Method	F1	LLM%	P-Rec	API-Tok
AdaLoc	68.11	94.00	66.00	738
+ Soft Critique Re-rank	67.84	94.00	66.00	1309

Table 19: Self-RAG-style soft critique re-ranking attached to AdaLoc on 100 SciScope queries. Answer quality remains unchanged while API cost increases by $\sim$50%.

Method	F1	LLM%	P-Rec	API-Tok
AdaLoc	68.11	94.00	66.00	738
+ Hard Critique Filter	62.45	88.00	54.00	1150
+ Hard + Retry	64.12	90.00	58.00	1600

Table 20: Self-RAG-style hard critique filtering. Hard filtering hurts F1 by aggressively pruning subtle evidence; retry partially recovers but at higher cost.

On 100 SCISCOPE queries, soft critique re-ranking yields essentially unchanged quality (F1: 68.11$\rightarrow$67.84) while increasing API tokens by $\sim$50%. Hard filtering is more detrimental, with the largest drop in LOCAL queries, indicating that aggressive critique can hurt evidence coverage for localization. We conclude that critique-style mechanisms are not cost-effective for last-mile localization.

This confirms that the poor LOCAL extraction of long-context models cannot be fixed by simply changing the system prompt without simultaneously destroying their synthesis capabilities. A generic prompt causes semantic blurring on extraction tasks, while an extractive prompt causes severe fragmentation on synthesis tasks. AdaLoc's dynamic routing explicitly circumvents this seesaw effect by ensuring each query receives both its optimal evidence granularity and its optimal generation prompt.

I Prompts Used

Query Router Prompt (LLM Fallback). When rule-based signals are inconclusive (confidence $S_{\text{router}} \in [\tau_{\text{low}}, \tau_{\text{high}}]$), we use the following prompt:

```
Classify this academic query as GLOBAL
or LOCAL.
GLOBAL: Requires synthesizing
information across multiple sections.
LOCAL: Requires finding a specific fact,
number, or value from one location.
Query: {query}
Classification (GLOBAL/LOCAL):
```

[Few-Shot Exemplars Provided to the LLM]

- **Global Examples:**
  1. Query: "How does the proposed method differ from prior work?" $\rightarrow$ GLOBAL
  2. Query: "What are the main limitations discussed by the authors?" $\rightarrow$ GLOBAL
  3. Query: "Summarize the ablation study results." $\rightarrow$ GLOBAL
  4. Query: "What is the overall motivation of this paper?" $\rightarrow$ GLOBAL
  5. Query: "Explain the theoretical framework of the new architecture." $\rightarrow$ GLOBAL
- **Local Examples:**
  1. Query: "What is the exact learning rate used for the Adam optimizer?" $\rightarrow$ LOCAL
  2. Query: "How many layers are in the transformer encoder?" $\rightarrow$ LOCAL
  3. Query: "What is the batch size mentioned in the implementation details?" $\rightarrow$ LOCAL
  4. Query: "What F1 score is reported on the QASPER dataset?" $\rightarrow$ LOCAL
  5. Query: "What is the value of the dropout parameter ρ?" $\rightarrow$ LOCAL

[1]

Answer Generation Prompt (Local).

```
Based on the following context from the
paper, answer the question.
Extract the SPECIFIC value or phrase. Be
precise and concise.
Context: {context}
Question: {query}
Answer:
```

Answer Generation Prompt (Global).

```
Based on the following context from the
paper, answer the question.
Synthesize information across the
provided passages.
Give a comprehensive but concise answer
(1-3 sentences).
Context: {context}
Question: {query}
Answer:
```

LLM-as-Judge Prompt.

```
You are evaluating answer correctness
for academic QA.
Question: {query}
Ground Truth: {gt_answer}
Prediction: {pred_answer}
Is the prediction semantically correct?
Answer CORRECT, PARTIAL, or WRONG.
```

[1] Earlier drafts used the labels SIMPLE/COMPLEX; in this version, we use GLOBAL/LOCAL throughout for consistency.

J Reproducibility Ledger

To ensure complete reproducibility, we detail our experimental configuration in the following.

Dataset Statistics.

- **Source:** SCISCOPE (derived from arXiv CS papers).
- **Total Size:** $N = 2000$ queries.
- **Split:** 1000 GLOBAL (Synthesis) / 1000 LOCAL (Fact Extraction).
- **Reference Correction:** All questions containing implicit references ("this paper") were manually resolved to explicit titles.

Effect of Reference Correction. To reduce ambiguity in SCISCOPE, we resolve paper-level coreferences (e.g., "this paper") to explicit paper titles in the query text. To assess whether this preprocessing materially changes task difficulty, we include two controls: (i) **No-title**, which keeps the original query text without injecting the full paper title, and (ii) **Title-only retrieval**, which retrieves evidence using only the injected title text as the retrieval query. Preliminary results show that reference correction improves retrieval precision without inflating scores artificially.

Baselines & Truncation Policies.

- **Direct LLM:** Closed-book. No truncation.
- **Naive RAG:** Fixed 400-token chunks (overlap 50). Top-5 chunks retrieved via BM25 + BGE embedding reranking.
- **Full-Context (128K):** Complete paper text fed to DeepSeek-Chat (128K window). No truncation for papers <128K tokens.
- **AdaLoc:** Two-stage adaptive chunking. Stage-1: Global [80–400] / Local [50–250] tok. Stage-2 localization: Global $L = 500$ / Local $L = 200$ tok. Max context length 2000 tokens.

Compute & Environment.

- **Hardware:** Single NVIDIA A100 GPU (80GB) for local embeddings/selector.
- **Software:** Python 3.10, PyTorch 2.1, Transformers 4.47.
- **Reproducibility:** We report average metrics over three independent runs.
- **API Config:** `temperature=0.0` for all generation/judgment calls.

Token Budget Accounting. We report **API Tokens** as the sum of input prompts + generated completions sent to the remote LLM.

- **Average API Tokens (AdaLoc):** 738 tokens/query.
- **Average API Tokens (Full-Context):** 6,984 tokens/query.

Router Correctness	N	Avg F1
Correct (97.1%)	1942	66.5%
Wrong (2.9%)	58	38.0%
Gap	–	**28.5 pts**

Table 21: End-to-end F1 grouped by router correctness.

K Regarding the use of LLMs

LLMs were employed solely for polishing the writing and enhancing readability.All scientific ideas,data,and analyses presented herein are human-generated and original.