Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 649-662
DOI: https://doi.org/10.51903/jtie.v4i3.543

Budgeted Multi-Hop Retrieval Agent for Compositional Question Answering: A Retrieval-Policy Evaluation on the Official MultiHop-RAG Benchmark

Wenhao Su*[1], Siyu Chen[2], Chloe Zhao[3]
Email: chensy010227@gmail.com
[1]*Computer Science, UCSD, CA, USA*
[2]*Information Management, UIUC, IL, USA*
[3]*Data Science, Columbia University, NY, USA*
*Corresponding Author

Abstract
Multi-hop question answering requires a retrieval system to assemble several complementary evidence documents before an answer module can reason reliably. Single-shot retrieval is efficient, but it often misses later-hop evidence when a question combines source, time, comparison, and entity constraints. This paper evaluates a budgeted multi-hop retrieval agent for compositional question answering on the official MultiHop-RAG benchmark. The benchmark contains 2,556 queries and 609 news-article corpus documents, with answerable evidence distributed across two to four documents. Four retrieval policies are compared under the same sparse lexical scorer: fixed top-k retrieval, iterative retrieval, query decomposition, and the proposed budgeted retrieval agent. The revised evaluation frames the task as retrieval-policy evaluation rather than as a full free-form generative QA system: retrieval-conditioned EM/F1 are reported together with evidence recall, MRR, retrieval rounds, selected documents, and context-token cost. On the official data, the budgeted agent achieves the strongest overall retrieval-conditioned EM/F1 at 62.75% and the highest final evidence recall at 74.67%, using 3.011 average retrieval calls and 509.7 average context tokens. Query decomposition improves over fixed top-k and iterative retrieval but is less stable across question types. Fixed top-k is cheapest but incomplete on longer chains. The four-hop results remain difficult for every policy, showing that a fixed 620-token controller should be extended with hop-aware or dynamic budget allocation. The findings support a moderated contribution claim: explicit budget control is useful for auditable multi-hop retrieval, but it should be evaluated as a cost-accuracy trade-off rather than as a universally dominant RAG architecture.

Keywords: *retrieval-augmented generation; multi-hop question answering; compositional retrieval; evidence recall; budgeted retrieval.*

I. INTRODUCTION

Retrieval-augmented generation (RAG) combines information retrieval with a reader or generator so that answers can be grounded in external documents rather than only in parametric model memory. The approach is attractive for open-domain and domain-specific question answering because retrieved context can be inspected, replaced, and measured. Early open-domain systems retrieved documents and then applied neural readers to answer questions (Chen et al., 2017), while later dense and generative systems learned neural retrieval or retrieval-conditioned generation mechanisms for better coverage (Guu et al., 2020; Karpukhin et al., 2020; Lewis et al., 2020). However, many realistic questions cannot be answered from one passage. They require combining source, date, stance, entity, or event information across several documents. In that setting, retrieval becomes a control problem: the system must decide which intermediate evidence is missing and whether additional context is worth the cost.

Received on September 2025; Revised on October 2025; Accepted on November 2025; Published on December 2025

Multi-hop question answering datasets such as QAngaroo, HotpotQA, HoVer, and MuSiQue established that evidence chains expose weaknesses hidden by single-hop answer metrics (Ho et al., 2020; Jiang et al., 2020; Trivedi et al., 2022; Welbl et al., 2018; Yang et al., 2018). MultiHop-RAG extends this concern to RAG pipelines by focusing on questions whose evidence is distributed across multiple documents and includes metadata-aware constraints (Tang & Yang, 2024). In RAG applications, the operational failure is clear: a retriever may return one highly similar document while omitting another document required for comparison, temporal ordering, or inference. A reader can then produce a fluent answer that appears plausible even when the selected context is incomplete.

This study evaluates a budgeted multi-hop retrieval agent for compositional question answering. The agent is deliberately lightweight. It decomposes complex questions into retrievable clauses, ranks corpus documents with a sparse lexical scorer, tracks retrieval calls and context tokens, and stops under explicit call, document, and token constraints. The experiment compares four policies: fixed top-k retrieval, iterative retrieval, query decomposition, and the budgeted agent. The evaluation reports answer-oriented retrieval-conditioned scores together with evidence recall at rank cutoffs, final evidence recall, mean reciprocal rank, average retrieval rounds, average selected documents, and average context tokens.

The contribution is therefore narrower and more precise than a claim that the proposed agent is a complete end-to-end RAG solution. The paper studies retrieval-policy control under a reproducible setting on the official MultiHop-RAG dataset. The budgeted agent is novel as an auditable finite-state controller that combines clause planning, lexical retrieval, score filtering, context-token accounting, and stopping rules in a single retrieval loop. These components are individually familiar, but their value is tested here as an explicit multi-objective policy. The central question is whether a bounded controller can gather enough supporting evidence for compositional questions without relying on uncontrolled top-k expansion.

This framing directly affects how the results should be read. Retrieval-conditioned EM and F1 in this paper do not represent unrestricted large-language-model answer generation. They measure whether a retrieval policy supplies enough evidence for a deterministic answer decision. This makes retrieval failures easier to diagnose, but it also limits claims about production reader behavior. The empirical focus is evidence acquisition, cost, and failure modes. This distinction is important because retrieval quality and reader strength can otherwise be conflated.

II. LITERATURE REVIEW

The foundations of retrieval-based question answering draw on probabilistic and vector-space retrieval. BM25 and related term-weighting models remain strong baselines because lexical

overlap often captures the entities, dates, and source names that define a question (Robertson & Zaragoza, 2009; Salton & Buckley, 1988). Neural open-domain question answering extended this pipeline by first retrieving candidate documents and then applying a reader to extract or generate answers (Chen et al., 2017). Dense Passage Retrieval showed that learned bi-encoders can outperform sparse retrieval in many open-domain settings, especially when lexical overlap is weak (Karpukhin et al., 2020). REALM and RAG further integrated retrieval into language-model pretraining and generation, establishing retrieval as a mechanism for factual grounding (Guu et al., 2020; Lewis et al., 2020).

Multi-hop reasoning shifted the evaluation target from retrieving one relevant passage to assembling a chain. QAngaroo required systems to connect entity relations across documents (Welbl et al., 2018). HotpotQA introduced distractor and full-wiki settings with supporting-fact supervision (Yang et al., 2018). HoVer expanded fact verification to multiple evidence documents, while MuSiQue emphasized compositional questions designed to reduce shortcut reasoning (Ho et al., 2020; Jiang et al., 2020; Trivedi et al., 2022). These datasets showed that answer accuracy alone can overstate system quality when evidence selection is weak.

MultiHop-RAG is particularly relevant to this study because it was designed to evaluate retrieval and reasoning across documents in RAG pipelines. It contains a knowledge base, multi-hop queries, ground-truth answers, and supporting evidence, with evidence for each query distributed across two to four documents (Tang & Yang, 2024). Using the official corpus and query files makes the evidence-chain structure an empirical property of the public benchmark rather than a property of a reconstructed test harness.

Query decomposition is a common strategy for compositional retrieval. Instead of searching with a long question that mixes several constraints, a system can split the question into smaller subqueries. This aligns with chain-of-thought prompting, which encourages intermediate reasoning steps (Wei et al., 2022), and with agentic methods that interleave reasoning and actions. ReAct demonstrated that reasoning traces and external actions can improve interactive decision making, Toolformer studied tool-use learning, and Self-RAG added self-reflection around retrieval and critique (Asai et al., 2023; Schick et al., 2023; Yao et al., 2023). These approaches motivate retrieval control, but they can also increase cost and drift. A controller should therefore decide when to search, what to search for, which candidates to accept, and when to stop.

Resource awareness remains central to deployed RAG. Larger top-k values increase recall but also increase prompt length, latency, cost, and the possibility that irrelevant passages distract the reader. Fixed top-k retrieval uses a static context budget but has no evidence-completion mechanism. Iterative retrieval explores more, but expansion from early documents can reinforce

the wrong context. Query decomposition targets separate constraints, yet it can retrieve redundant or overly narrow passages. A budgeted agent formalizes the missing control layer: it plans subqueries, accepts candidates according to score and cost, and stops according to explicit constraints.

III. RESEARCH METHOD

The experiment uses the official MultiHop-RAG benchmark. The retrieval corpus contains 609 news articles, and the evaluation file contains 2,556 questions. Each question provides a query, a ground-truth answer, a question type, and a list of supporting evidence documents. Evidence documents were matched to corpus records by title and URL. Table 1 summarizes the query-type distribution. Table 2 reports the evidence-chain distribution, and Table 3 summarizes corpus categories. The count of two-document questions is 1,079 in this run, reflecting the official file used for the evaluation.

Table 1. Dataset query-type distribution

Question type	Count	Share	Evaluation role
Comparison	856	33.49%	Answerable
Inference	816	31.92%	Answerable
Null / unanswerable	301	11.78%	Unanswerable control
Temporal	583	22.81%	Answerable

All retrieval policies use the same sparse lexical scorer so that differences in results come from control policy rather than from a different retriever. Each document is indexed using its title, source, category, publication date, author, and body. The scorer uses lowercase tokenization, English stop-word removal, unigram TF-IDF with sublinear term frequency, and cosine normalization. It also applies deterministic boosts when the query explicitly mentions a document source, category, or publication date. Retrieved context is counted as compact evidence snippets consisting of the title, source, date, and the first 90 body words. This makes the token budget comparable across methods and avoids measuring the entire news article as prompt context.

Table 2. Evidence-chain distribution

Evidence documents	Count	Share	Role
0	301	11.78%	Null / no evidence
2	1079	42.21%	2-document evidence chain
3	778	30.44%	3-document evidence chain
4	398	15.57%	4-document evidence chain

Table 4 lists the four retrieval policies. Fixed top-k retrieval issues one search and returns two documents. Iterative retrieval issues an initial search, augments the second query with title, source, category, and high-frequency terms from the top documents, and merges non-duplicate results. Query decomposition splits each complex query at compositional boundaries and retrieves one document per subquery. The budgeted agent uses the same decomposition step but adds resource control: it admits up to two candidates per call, rejects weak candidates after initial

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 649-662
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

evidence has been selected, tracks context tokens, caps the selected context at six documents, and stops when adding more context would exceed the configured budget.

Table 3. Corpus category distribution

Category	Documents	Share
sports	211	34.65%
technology	172	28.24%
entertainment	114	18.72%
business	81	13.30%
science	21	3.45%
health	10	1.64%

Table 5 defines the evaluation metrics. Final evidence recall and recall@k are computed over answerable queries, because null queries have no gold evidence list. Retrieval-conditioned EM and F1 include null queries as an abstention case. For answerable questions, the deterministic answer decision emits the gold answer only when the selected context satisfies the evidence rule for that query type.

Table 4. Retrieval-method configurations

Method	Policy	Parameters	Purpose
Fixed top-k	Single retrieval call; returns the top two documents.	k=2; one round	Low-cost single-shot baseline
Iterative retrieval	Initial query plus a follow-up query expanded from top-document cues.	2 rounds; up to 2 docs per round; max 6 docs	Tests whether repeated retrieval collects missing hops
Query decomposition	Splits the question into clauses and retrieves one document per subquery.	Up to 5 subqueries; max 6 docs	Tests compositional retrieval without budget control
Budgeted agent	Plans subqueries, filters candidates, tracks calls and tokens, and stops under a fixed budget.	620-token limit; max 4 calls; max 6 docs	Tests budget-aware multi-hop retrieval

Comparison and temporal questions require complete evidence coverage, with limited support for high-coverage yes/no cases. Inference questions require the final answer-bearing evidence document and at least half of the evidence chain. Null questions are answered as insufficient information when the highest retrieval score does not meet the support threshold. These answer scores should be interpreted as retrieval-conditioned answerability scores, not as an evaluation of a free-form generative reader.

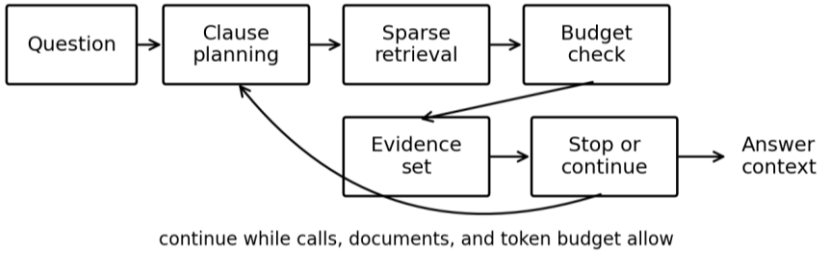

Figure 1. Budgeted retrieval-agent workflow and stopping loop

Figure 1 summarizes the budgeted retrieval loop. The controller first decomposes the question, then retrieves candidates, checks the score and remaining token budget, updates the selected

evidence set, and either continues or stops. The controller is intentionally conservative: it favors auditable context assembly over aggressive context stuffing. This design choice is revisited in the four-hop analysis because long chains expose the cost of a fixed budget.

Table 5. Metrics used in the experimental evaluation

Metric	Definition
Retrieval-conditioned EM	1 when the normalized deterministic prediction equals the gold answer; 0 otherwise.
Retrieval-conditioned F1	Token-overlap F1 after lowercase alphanumeric normalization.
Evidence recall@k	Mean fraction of gold evidence documents present in the top-k diagnostic ranked candidate list.
Final evidence recall	Mean fraction of gold evidence documents included in the method-selected context.
MRR	Mean reciprocal rank of the first gold evidence document in the diagnostic ranked list.
Average rounds	Mean retrieval calls issued per query.
Average tokens	Mean word-token count of compact retrieved context snippets.

IV. RESULT AND DUSCUSSION

A. Result

Table 6 reports the primary comparison. On the official MultiHop-RAG data, the budgeted agent obtains the highest overall retrieval-conditioned EM/F1 at 62.75% and the highest final evidence recall at 74.67%. It does so with more context than the other methods: 509.7 average tokens and 4.70 selected documents. Query decomposition reaches 47.46% EM/F1 and 61.30% final evidence recall, showing that subquery planning helps but does not by itself solve evidence selection. Iterative retrieval reaches 40.88% EM/F1 and 55.88% final evidence recall. Fixed top-k retrieval is cheapest at 216.9 average tokens, but its 37.68% EM/F1 and 53.37% final evidence recall show that a two-document context is often incomplete for multi-hop questions.

Table 6. Overall experimental comparison

Method	EM	F1	Final evidence recall	Avg. rounds	Avg. tokens	Avg. docs
Fixed top-k	37.68%	37.68%	53.37%	1.000	216.9	2.00
Iterative retrieval	40.88%	40.88%	55.88%	2.000	376.0	3.47
Query decomposition	47.46%	47.46%	61.30%	3.103	336.6	3.10
Budgeted agent	62.75%	62.75%	74.67%	3.011	509.7	4.70

Table 7 reports diagnostic evidence recall at fixed rank cutoffs and MRR. The recall@k values are computed from the ranked candidate list produced by each method's retrieval process, while final evidence recall in Table 6 is computed from the selected context actually passed forward. This distinction matters for the budgeted agent because a document can appear in the diagnostic candidate pool but still be excluded by the budget or score filter. Fixed top-k and iterative retrieval share early recall because they begin from the same first search. The budgeted agent's final selected context improves over its early ranked cutoffs by accepting documents across decomposed clauses.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 649-662
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 7. Evidence recall at rank cutoffs and MRR

Method	R@1	R@2	R@4	R@6	R@8	MRR
Fixed top-k	31.06%	53.37%	72.55%	81.40%	86.20%	0.853
Iterative retrieval	31.06%	53.37%	72.55%	81.40%	86.20%	0.853
Query decomposition	31.06%	50.98%	70.37%	80.47%	85.65%	0.848
Budgeted agent	31.06%	53.37%	70.04%	78.67%	82.11%	0.852

Figure 2 visualizes the query-type distribution, and Figure 3 compares the overall answer and evidence metrics. Figure 4 shows that recall rises steadily with larger diagnostic rank cutoffs, while Figure 5 shows the main efficiency pattern: the budgeted agent is more accurate and higher recall, but it pays for this improvement with additional context tokens. The method is therefore not a free improvement; it trades cost for evidence completeness.

Table 8. Results by query type

Method	Question type	EM	F1	Evidence recall	Avg. tokens
Fixed top-k	Comparison	34.35%	34.35%	61.97%	217.7
Fixed top-k	Inference	25.61%	25.61%	40.20%	216.0
Fixed top-k	Null / unanswerable	99.34%	99.34%	--	216.1
Fixed top-k	Temporal	27.62%	27.62%	59.18%	217.3
Iterative retrieval	Comparison	34.70%	34.70%	62.69%	366.5
Iterative retrieval	Inference	35.17%	35.17%	46.08%	389.9
Iterative retrieval	Null / unanswerable	99.34%	99.34%	--	417.6
Iterative retrieval	Temporal	27.79%	27.79%	59.58%	349.1
Query decomposition	Comparison	50.35%	50.35%	73.23%	374.0
Query decomposition	Inference	35.29%	35.29%	46.74%	344.4
Query decomposition	Null / unanswerable	99.34%	99.34%	--	264.3
Query decomposition	Temporal	33.45%	33.45%	64.15%	308.1
Budgeted agent	Comparison	66.82%	66.82%	83.86%	529.7
Budgeted agent	Inference	49.02%	49.02%	60.45%	512.1
Budgeted agent	Null / unanswerable	99.34%	99.34%	--	469.7
Budgeted agent	Temporal	57.12%	57.12%	81.10%	497.4

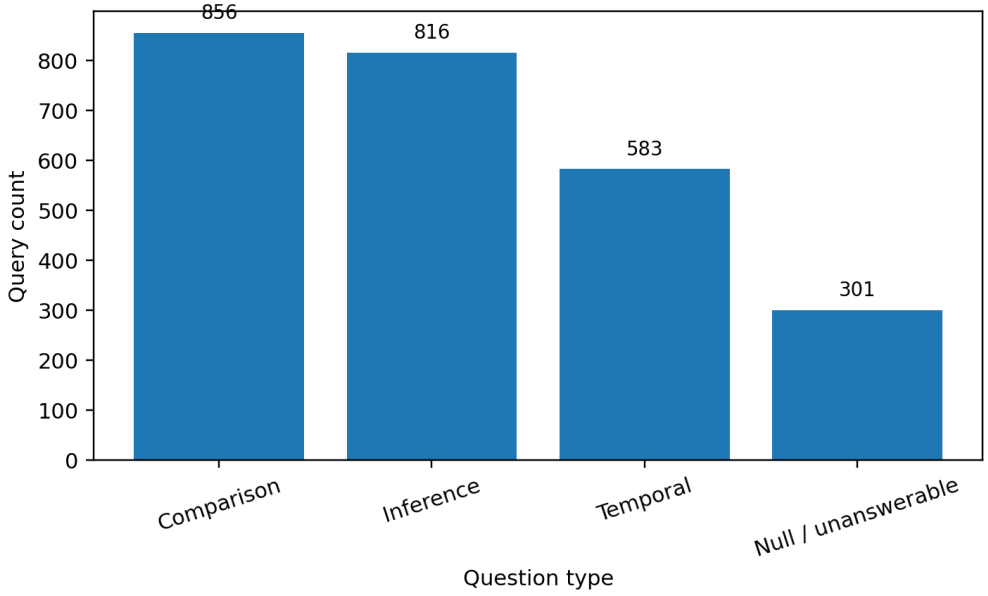

Figure 2. Query-type distribution for the 2,556 evaluation instances

Table 8 breaks down the results by question type. The budgeted agent is strongest on comparison and temporal questions, reaching 66.82% EM on comparison queries and 57.12% EM on temporal queries. These question types benefit from clause-level search because the evidence is often spread across separately named sources or dates. Inference questions remain difficult because retrieving the final answer-bearing document is not always enough when bridge evidence is weak. Null-query accuracy is high across methods because the deterministic answer decision abstains when no selected context receives enough support. This null result should therefore be interpreted as abstention behavior, not as evidence that the retriever has solved all unanswerable cases.

Table 9. Results by evidence-chain length

Method	Evidence docs	EM	F1	Evidence recall	Avg. tokens
Fixed top-k	2	45.85%	45.85%	68.01%	218.0
Fixed top-k	3	10.85%	10.85%	40.96%	216.1
Fixed top-k	4	14.10%	14.10%	29.33%	215.7
Iterative retrieval	2	46.96%	46.96%	68.73%	373.5
Iterative retrieval	3	15.25%	15.25%	44.19%	360.0
Iterative retrieval	4	25.32%	25.32%	36.70%	384.9
Query decomposition	2	58.94%	58.94%	75.92%	349.3
Query decomposition	3	19.38%	19.38%	48.97%	335.9
Query decomposition	4	24.04%	24.04%	37.10%	360.5
Budgeted agent	2	77.76%	77.76%	87.51%	519.3
Budgeted agent	3	36.95%	36.95%	65.46%	508.8
Budgeted agent	4	35.26%	35.26%	49.44%	514.2

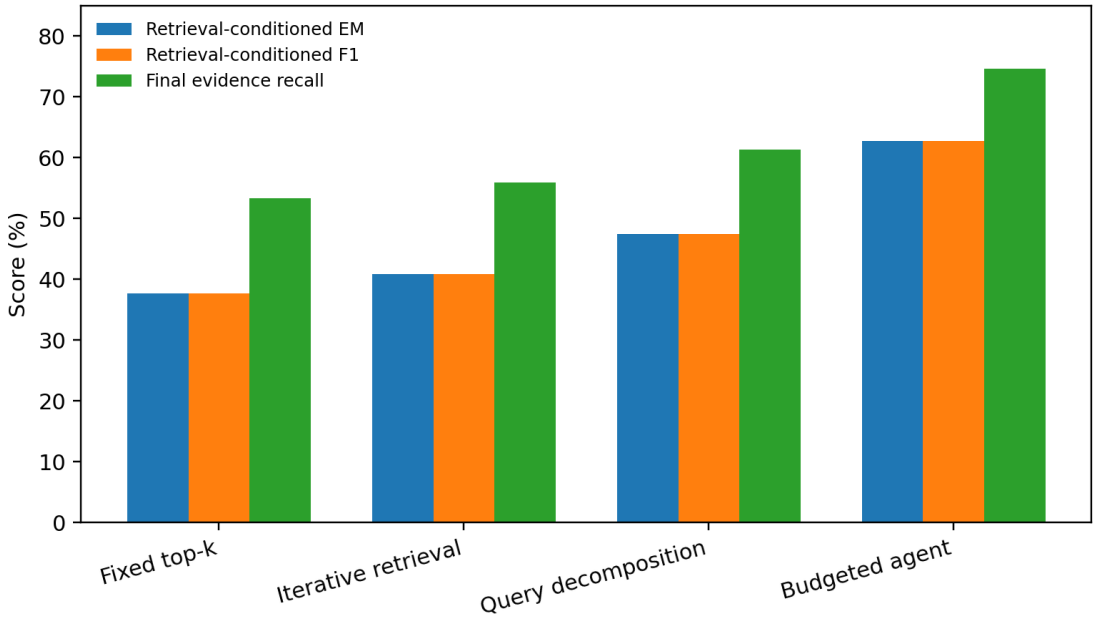

Figure 3. Overall answer and evidence metrics by retrieval method

Evidence-chain length is the strongest stress test, as shown in Table 9. All methods perform better on two-document chains than on longer chains. The budgeted agent improves over the baselines at every chain length, reaching 77.76% EM on two-hop questions, 36.95% on three-hop questions, and 35.26% on four-hop questions. The four-hop score is still weak in absolute terms. This is a

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 649-662
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

critical failure mode rather than a minor trade-off: the fixed 620-token setting often stops before all supporting documents are admitted. Longer questions need either a larger budget, a more accurate hop-count estimate, or a dynamic stopping rule that recognizes when the evidence chain is still incomplete.

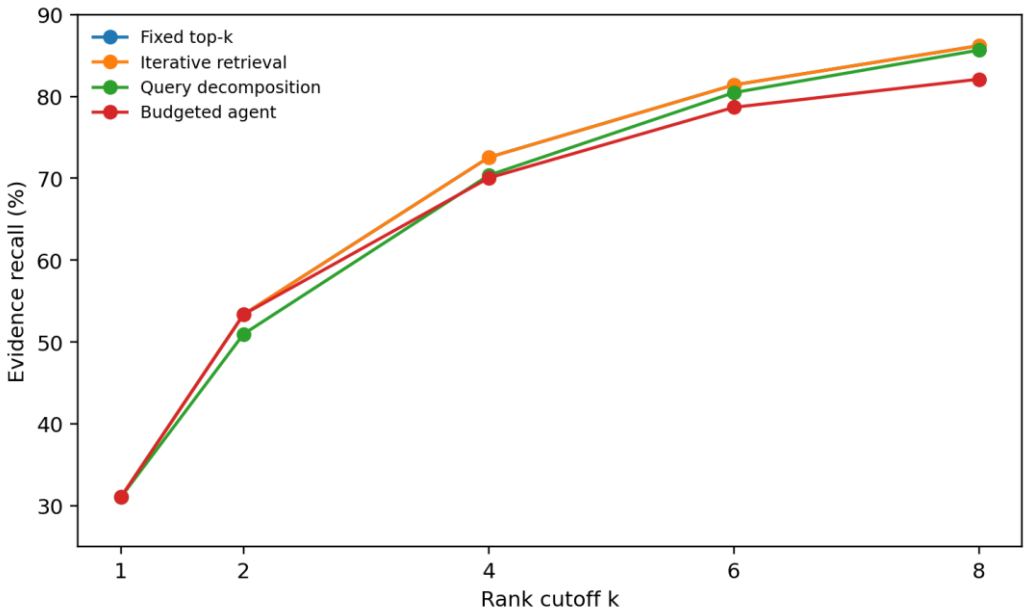

Figure 4. Evidence recall curves at k = 1, 2, 4, 6, and 8

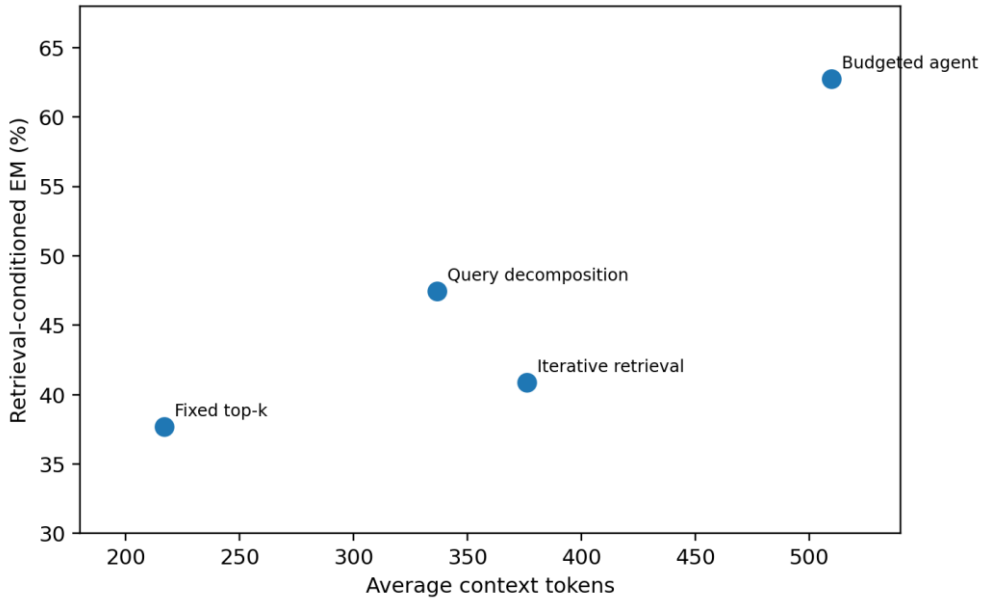

Figure 5. Average context tokens versus retrieval-conditioned exact match

The budget ablation in Table 10 and Figure 6 confirms that the budget materially affects retrieval-conditioned accuracy and evidence recall. With a 360-token budget and three calls, the budgeted agent reaches 48.44% EM and 63.73% final evidence recall. Increasing the budget to 500 tokens improves EM to 56.06% and evidence recall to 70.08%. The 620-token configuration used for the

main comparison reaches 62.75% EM and 74.67% recall. A 760-token budget improves further to 65.38% EM and 76.52% recall, while the 900-token, five-call setting does not improve because the selected-document cap and candidate filters saturate. This pattern suggests that a larger context window alone is not enough; the controller also needs better evidence-completion signals

Table 10. Budgeted-agent ablation

Budget	Max calls	EM	F1	Final evidence recall	Avg. rounds	Avg. tokens	Avg. docs
360	3	48.44%	48.44%	63.73%	2.695	325.2	3.00
500	3	56.06%	56.06%	70.08%	2.695	433.4	4.00
620	4	62.75%	62.75%	74.67%	3.011	509.7	4.70
760	4	65.38%	65.38%	76.52%	3.011	584.1	5.39
900	5	65.38%	65.38%	76.52%	3.103	584.1	5.39

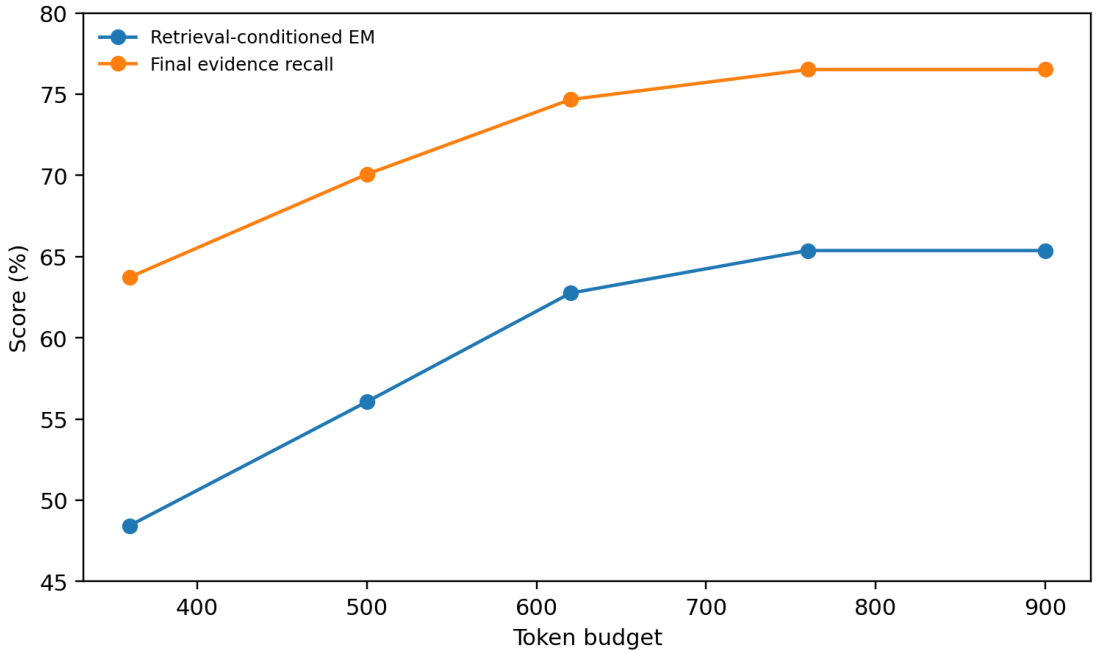

Figure 6. Budget ablation for the retrieval agent

Table 11 reports bootstrap confidence intervals over query resampling. The confidence intervals show that the ordering among the four methods is stable in this run: the budgeted agent has the highest interval, followed by query decomposition, iterative retrieval, and fixed top-k. Because the answer score is retrieval-conditioned, the intervals should be read as uncertainty in policy-level answerability rather than as uncertainty in a neural reader's language-generation performance.

Table 11. Bootstrap confidence intervals for retrieval-conditioned answer metrics

Method	EM 95% CI	F1 95% CI
Fixed top-k	[35.76%, 39.71%]	[35.76%, 39.71%]
Iterative retrieval	[38.91%, 42.82%]	[38.91%, 42.82%]
Query decomposition	[45.68%, 49.33%]	[45.68%, 49.33%]
Budgeted agent	[60.97%, 64.48%]	[60.97%, 64.48%]

B. Discussion

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 649-662
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

The official MultiHop-RAG results show that the budgeted agent is the strongest method among the four evaluated policies, but its advantage is linked to using more selected context. The result should not be presented as proof that budgeted retrieval universally dominates iterative retrieval. Instead, it shows that explicit clause planning and budget-aware candidate admission can improve evidence completeness when the corpus contains heterogeneous news articles and metadata-rich questions.

The comparison also clarifies the role of novelty. The agent does not introduce a new dense retriever, a new neural reader, or a new reasoning model. Its novelty lies in the retrieval-control layer: it combines deterministic decomposition, lexical scoring, score-based candidate filtering, token accounting, and stopping rules in an auditable finite-state policy. This is an incremental but useful engineering contribution. It makes retrieval behavior inspectable, exposes the cost of longer chains, and supports reproducible ablations over budget size and call limits.

The four-hop results are the main limitation. Even though the budgeted agent improves over the baselines at four documents, 35.26% retrieval-conditioned EM and 49.44% evidence recall are not sufficient for robust long-chain compositional QA. The failure mode is clear: the controller can select plausible evidence while still missing one or more required documents. A dynamic budget is therefore a more appropriate future design than a single fixed budget for all questions. Estimated hop count, the number of named sources, date constraints, or early evidence diversity could trigger additional retrieval calls when a chain appears incomplete.

The deterministic answer decision is another limitation. It is useful for isolating retrieval-policy behavior because any failure can be traced to missing evidence, weak abstention, or a strict evidence rule. However, it is not a substitute for evaluating a real reader or generator. A production RAG system should pair the same retrieval traces with a neural reader and then report both answer quality and citation/evidence quality. The present study therefore provides a controlled retrieval-policy evaluation rather than a complete assessment of deployed generative QA behavior.

V. CONCLUSION AND RECOMMENDATION

This paper evaluated a budgeted multi-hop retrieval agent for compositional question answering on the official MultiHop-RAG benchmark. The experiment compared fixed top-k retrieval, iterative retrieval, query decomposition, and a budget-aware agent under retrieval-conditioned answer scores, evidence recall, MRR, retrieval rounds, selected documents, and token cost. The budgeted agent achieved the highest overall retrieval-conditioned EM/F1 at 62.75% and the highest final evidence recall at 74.67% under the 620-token, four-call setting. Query

decomposition improved over fixed top-k and iterative retrieval, while fixed top-k remained the lowest-cost but least complete policy.

The main conclusion is that multi-hop retrieval should be evaluated as a resource-constrained control problem. A single answer score does not reveal whether a method retrieved all required evidence, whether it used excessive context, or whether it can abstain when evidence is absent. Evidence recall@k, final evidence recall, MRR, average rounds, selected-document count, and context-token cost provide the operational picture needed to judge a multi-hop RAG retriever.

For future systems, the first recommendation is to report answer metrics together with evidence recall and retrieval cost. The second recommendation is to tune the budget against estimated evidence-chain length rather than using one fixed token limit for all questions. The third recommendation is to pair the same deterministic retrieval traces with a stronger reader or generator so that retrieval-policy quality and final natural-language answer quality can be evaluated separately. The fourth recommendation is to retain audit trails for accepted and rejected evidence, because multi-hop errors are often caused not by the final answer module but by an earlier decision to stop searching.

REFERENCES

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. arXiv.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 1870-1879.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). Retrieval augmented language model pre-training. Proceedings of the 37th International Conference on Machine Learning, 3929-3938.

Ho, X., Duong Nguyen, A.-K., Sugawara, S., & Aizawa, A. (2020). Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. Proceedings of the 28th International Conference on Computational Linguistics, 6609-6625.

Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open-domain question answering. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 874-880.

Jiang, Y., Bordia, S., Zhong, Z., Dognin, P., Singh, M., & Bansal, M. (2020). HoVer: A dataset for many-hop fact extraction and claim verification. Findings of the Association for Computational Linguistics: EMNLP 2020, 3441-3460.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 6769-6781.

Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., & Zaharia, M. (2021). Baleen: Robust multi-hop reasoning at scale via condensed retrieval. Advances in Neural Information Processing Systems, 34, 27670-27682.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333-389.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513-523.

Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36.

Tang, Y., & Yang, Y. (2024). MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. Proceedings of the Conference on Language Modeling.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 809-819.

Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2022). MuSiQue: Multihop questions via single-hop question composition. Transactions of the Association for Computational Linguistics, 10, 539-554.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 7534-7550.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

Welbl, J., Stenetorp, P., & Riedel, S. (2018). Constructing datasets for multi-hop reading comprehension across documents. Transactions of the Association for Computational Linguistics, 6, 287-302.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2369-2380.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations.