BanditRAG: A Bandit Scheme for Learning Multi-hop Scoring function for Optimal RAG

Anonymous ACL submission

Abstract

Recent advances in Large Language Models has lead to their widespread adoption as generative answer engines for complex information access needs. However, addressing complex information needs, in tasks like complex Question Answering, requires a multi-stage document retrieval-rerank pipeline that has to capture relevant documents at lower depths to ensure efficient generation with manageable context length limits. While recent neighborhood-aware retrieval approaches aim to bridge the recall boundedness issue of first-stage retrieval, they are heuristic and do not explore the space of possible relevant documents. We propose a novel exploration algorithm to construct a subset of relevant documents from the initial retrieved pool and the pool of neighborhood documents. Our results demonstrate gains in QA performance while ensuring efficiency by minimizing number of LLM calls by **77.4%**, requiring **5 x** fewer calls on average per query, compared to state-of-the-art in complex QA tasks. The code and data can be found at [1].

1 Introduction

Open-domain complex Question Answering (CQA) is a challenging task that requires retrieval and reasoning around multiple aspects of a query and multiple pieces of evidence to address the complex information needs. Such systems employ query understanding, retrieval and reasoning or generation using LLMs. With the advances in large Language models, query decomposition and iterative retrieval approaches that interleave retrieval and reasoning, like Self-Ask (Press et al., 2023), IRCOT (Trivedi et al., 2023), Searchain (Xu et al., 2024) among others (Qi et al., 2019; Yao et al., 2023). However, a fundamental limitation of these existing systems is the *bounded recall* problems where documents that are missed during the initial retrieval stage cannot be recovered later, overall affecting the QA performance. This is further exacerbated in CQA where failing to retrieve relevant documents for one of the intermediate sub-questions (Press et al., 2023) can result in cascading failures. More recently, SUNAR (Venktesh et al.) was proposed to bridge this recall gap by proposing a Neighborhood Aware retrieval (NAR) strategy, which explores related documents of the initially retrieved documents, leveraging the feedback of the LLM-based answering system.

While SUNAR demonstrated that this helps bridge the *bounded recall* issues it still suffers from two fundamental limitations. **Firstly**, apart from iterative query decomposition, reliance on cross-encoder based ranker for guiding NAR, SUNAR also relies on LLM feedback for each batch of documents, leading to huge latency and computational cost during inference for each query. This renders such pipelines infeasible for real-world latency and compute constraints. **Secondly**, the NAR component in SUNAR relies on the clustering hypothesis (Jardine and van Rijsbergen, 1971), which states that documents that are similar to each other help address the same query. Hence, while NAR aids in adaptively recovering related documents that were missed during first-stage retrieval, it does not account for the multi-aspect nature of the original complex query. The LLM feedback-based prioritization to select top-5 or top-10 documents from candidate documents retrieved using NAR across sub-questions may result in homogenous documents. This is because the LLM-based semantic uncertainty signal for document prioritization relies only on self-consistency of answers generated given a batch of documents from the candidate pool, which might lead to prioritizing documents leading to homogenous answers and not necessarily the right answer. For instance, for a query "Who is the mother of director of the film Polish-Russian war?", SUNAR may prioritize documents only related to the director of the film and may miss out

[1] https://anonymous.4open.science/status/Multi-hop-Bandit-RAG-51FD

documents related to the mother, impacting downstream performance.

Our goal here is to primarily **reducing the computational cost during inference** *without deteriorating CQA performance*. Hence, we propose BanditRAG, a novel RAG framework to train a lightweight multi-hop scoring function as a retriever for question adaptive selection of subsets of documents from a candidate pool. The retriever is trained offline based on LLM feedback. Our retrieval method includes a new optimization formulation for selection of a diverse set of paths in the multi-hop setting, thus resulting in document covering diverse aspects of the CQA problem. We also develop a multi-armed bandit-based algorithm for sample-efficient learning, where the sparse arm-selection is achieved by optimization of the multi-hop surrogate function. We demonstrate that the proposed method improves the recall of selected documents by $\sim 4.5\%$ accounting for relevance to the query and diversity to cover multiple aspects of the query, and reducing LLM calls by **77.4%**.

2 Related Work

Context Retrieval Approaches: The typical retrieval approaches for QA systems follow a retrieve-and-read framework (Chen et al., 2017), where the retriever can be either sparse (Robertson and Zaragoza, 2009; Formal et al., 2021b) or dense (Karpukhin et al., 2020; Xiong et al., 2020; Song et al., 2020). The reader or the answering system can be either fine-tuned (Mavi et al., 2022) or employ few-shot or zero-shot reasoning (Brown et al., 2020; Lewis et al., 2021). Pseudo Relevance Feedback (PRF) approaches aim to bridge query-document vocabulary/representation mismatch and improve retrieval effectiveness by recomputing query representations for re-retrieval/re-ranking using relevant documents from first-stage retrieval (Li et al., 2022). However, PRF methods are known to suffer from *drift* resulting in suboptimal performance due to retrieving irrelevant contexts.

While adaptive retrieval approaches have been proposed (MacAvaney et al., 2022; Kulkarni et al., 2023; Rathee et al., 2024) for improving recall, Our approach differs by adapting LLM feedback through bandit-based optimization to learn a scoring function offline. This obviates the expensive ranking of all documents at inference time.

Complex QA Approaches: With the recent advances in LLMs (Wei et al., 2022a; Brown et al., 2020; Wei et al., 2022b), Retrieval Augmented Generation (RAG) approaches have been proposed which retrieve to incorporate external knowledge mitigating hallucination to generate accurate answers (Khot et al., 2023; Xu et al., 2024; Schick et al., 2023). These RAG pipelines usually comprise query understanding, retrieval and reasoning stages that leverage the capabilities of a LLM to interleave retrieval and reasoning in a multi-turn fashion(Press et al., 2023; Xu et al., 2024; Khot et al., 2023; Yao et al., 2023; Trivedi et al., 2023; Dua et al., 2022). However, the existing RAG pipelines for complex QA are limited by the performance of the retriever (Asai et al., 2024; Li et al., 2024). An additional challenge is the context limit of LLMs, making it necessary to capture the relevant documents within the limited budget. Additionally, the inherent noise in retrieved documents, could result in a wrong reasoning path and (Shi et al., 2024) and derive a wrong answer. While recent approaches (Venktesh et al.; Rathee et al., 2024), introduce neighborhood-aware retrieval to address this issue, they rely on expensive LLM self-consistency based feedback, which also causes redundancy in the prioritized document pool. Hence, in this work, we focus on improving the recall of top-ranked documents at lower depths while ensuring the diversity of the pool to ensure coverage of multiple aspects of a complex query.

3 Optimal Multi-hop Scoring Function for RAG

In this section we describe *BanditRAG*, a Multi-armed Bandit (MAB)-based algorithm for learning a multi-hop scoring function for a set of documents in a RAG approach to complex QA.

3.1 Problem Formulation

A common approach to open-domain QA typically follows a two-step process: retrieval of relevant contexts, followed by a reasoning or QA step using those retrieved contexts (Lewis et al., 2021). In our setup, we are given a document corpus $\mathcal{D} = \{d_1, ..., d_N\}$, and a corpus of complex questions and their final answers $\mathcal{X} = \{(x_m, y_m), m = 1, ..., M\}$. We are interested in developing a system for constructing a subset $s(x_m)$ of the documents such that $s(x_m)$ can be used in a RAG framework along with an LLM for answering the question x_m. However, the size of $s(x_m)$ is usually constrained to be a small number (denoted as K), since process-

ing large contexts using LLMs can be expensive and confusing in a RAG QA setup (see e.g. (Cuconasu et al., 2024)). Hence, $|s(x_m)| \leq K$.

For complex QA, $s(x_m)$ must include multiple documents needed to answer each sub-question of the complex question x_m. Following SOTA methods, we assume to have access to the a set of sub-questions $\{x_{m1}, ..., x_{mk_m}\}$ corresponding each complex question x_m, obtained using a question decomposition method e.g. those used in (Khot et al., 2023; Venktesh et al.). For each document $d_i \in \eta(x_m)$, the a learned cross-encoder-based score function $\pi(x_m, d_i) \in [0, 1]$ is typically used tp predict the relevance of document d_i towards containing the answer of question x_m. A simple baseline is to sort the documents d_i in $\eta(x_m)$ using the score $\pi(x_m, d_i)$, and use the top-K documents. However, many gold documents do not contain the answers to the original question x_m but a sub-question, which is needed to arrive at the final answer. For example, in the situation where we are interested in answering the question *"Who is the mother of director of the movie Polish-Russian War?"*; the document about the movie contains the name of the director, but not the mother (which is the final answer) In order to promote the inclusion of such documents in the subset s given to the LLM, we must discard repeated similar documents to a given question.

Formally, given a question x_m, a retrieved candidate document pool $\eta(x_m) \subseteq D$, and an LLM-based answering function f, our goal is to construct a subset of documents $s(x_m) \subseteq \eta(x_m)$ of bounded size K, such that the resulting answer accuracy is maximized. Let

$$A(s(x_m), x_m) = \mathbf{1}[y_m = f(s(x_m), x_m)] \quad (1)$$

denote the answer correctness obtained by prompting the LLM with the selected subset $s(x_m)$, where y_m is the ground-truth answer. The retrieval objective can therefore be written as:

$$s^*(x_m) = \arg \max_{s \subseteq \eta(x_m),\, |s| \leq K} A(s, x_m) \quad (2)$$

That is, we seek the subset of retrieved documents that maximizes downstream QA accuracy under a constrained context budget. However, directly optimizing this objective is challenging because the reward signal $A(s, x_m)$ is observed only after expensive LLM inference and depends jointly on the interaction between multiple documents rather than individual document relevance alone. Moreover, existing iterative retrieval and decomposition-based approaches suffer from two important limitations. First, methods such as Self-Ask (Press et al., 2023) and SUNAR (Venktesh et al.) rely generated sub-questions for accurate retrieval. In complex QA settings, this can be expensive in-terms of LLM calls, does not provide a wholistic framework since the final answer depends on combining evidence across multiple reasoning hops, and failure to retrieve relevant documents for even a single intermediate sub-question can significantly deteriorate downstream reasoning performance. Second, while methods such as SUNAR employ semantic uncertainty signals from LLM-generated answers during inference, they do not explicitly learn from LLM feedback to optimize the retrieval policy itself. To address these limitations, we propose learning a multi-hop document scoring function that directly incorporates LLM-derived feedback signals into an offline optimization framework.

3.2 Learnable Multi-hop Document-Set Score

From the above discussion, it is clear that the utility of a document cannot be determined independently, since answering a multi-hop question often requires combining evidence distributed across multiple related documents. Hence, we define a multi-hop scoring function over an entire subset of documents, that combines semantic similarities between the question and candidate documents, as well as two candidate documents. Let $\pi(x_m, z_i) \in [0, 1]$ denote the relevance score assigned by a cross-encoder scoring to an intermediate document $z_i \in \eta(x_m)$. To model interactions between documents, we define a pairwise similarity function $\mathcal{S}(z_i, d_j)$ between an intermediate evidence document z_i and a candidate context document d_j. Given a subset of selected documents $s \subset \eta(x_m)$ of size at most k, we define the multi-hop document-set score as:

$$\sigma(s, x_m, \Theta) = \sum_{d_j \in s} \sum_{z_i \in \eta(x_m)} \pi(x_m, z_i) \cdot \mathcal{S}_\Theta(z_i, d_j) \quad (3)$$

The above objective measures the total utility of the selected subset by aggregating pairwise interactions between the selected documents and all retrieved intermediate evidence documents. The problem of selecting the best set s can be posed as a optimization problem over binary indicator variables $\alpha_j \in \{0, 1\}$ for each candidate document

$d_j \in \eta(x_m)$, where: $\alpha_j = 1$ if $d_j \in s, 0$otherwise. However, this flat formulation results in selection of many similar documents, ignoring the diverse aspects of a complex QA problem. For example, in the question "Who is the mother of the director of the movie Polish-Russian War?", we expect to select one set of documents corresponding to the "director" aspect and another set corresponding to the "mother" aspect. To this end, we propose a path selection formulation that helps in selecting diverse aspects.

Let $\eta(x_m)$ denote the retrieved candidate pool for question x_m. For every intermediate evidence document $z_i \in \eta(x_m)$ and candidate document $d_j \in \eta(x_m)$, we define $\beta_{ij} \in \{0, 1\}$ – an indicator denoting whether the reasoning path $z_i \rightarrow d_j \rightarrow x_m$ is activated, and $\gamma_i \in \{0, 1\}$ – an indicator denoting whether intermediate document z_i participates in at least one active reasoning path. Using these variables, the optimal multi-hop retrieval structure is obtained by solving:

$$\begin{aligned} \beta^*_{ij}, \gamma^*_i = \arg\max_{\beta_{ij},\gamma_i} \quad & \sum_i \sum_j \beta_{ij}\ \pi(x_m, z_i)\ S_\Theta(z_i, d_j) \\ \text{s.t.} \quad & \sum_i \beta_{ij} \leq 1, \qquad \forall d_j \in \eta(x_m), \\ & \gamma_i \geq \beta_{ij}, \qquad \forall i, j, \\ & \sum_i \gamma_i \leq K, \\ & \beta_{ij} \in \{0, 1\}, \qquad \gamma_i \in \{0, 1\}. \end{aligned} \tag{4}$$

The objective selects a set of high-utility multi-hop reasoning paths by jointly considering (i) the relevance of intermediate evidence documents to the original query and (ii) the semantic compatibility between connected evidence documents. Note that, the similarity function $\mathcal{S}_\Theta(z_i, d_j)$ is parameterized by learnable parameters Θ, which are optimized using LLM feedback signals described in the next section.

Query-token-sensitive Score: Unlike standard document similarity measures, we require $S_\Theta(z_i, d_j)$ to be *query-token-sensitive*, since different aspects or facets of a complex query may correspond to different reasoning hops. To achieve this, we employ contextualized token-level embeddings inspired by the late-interaction mechanism of ColBERT (Khattab and Zaharia, 2020). However, the classic ColBERT similarity function is a parameter-free and gives equal emphasis on all query tokens, resulting in homogeneous document selection. To address this issue, we propose a novel parameterized query-sensitive pairwise similarity function $S_\Theta(d, z\ \ x_m)$. We first cluster the contextualized query token embeddings into semantically coherent groups, where each cluster approximately corresponds to a latent reasoning aspect or sub-question. Let $\mathcal{C}(k)$ denote the cluster assignment of query token t_k, and let $\theta_{\mathcal{C}(k)}$ denote the learnable importance weight associated with that cluster. The parameters $\Theta = \{\theta_1, \ldots, \theta_{|\mathcal{C}|}\}$ therefore control the relative importance assigned to different semantic aspects of the query. Formally, let $d_{j(k)} = \arg\max_{d_{j()} \in d_j} \text{sim}(d_{ju}, t_k)$ and $z_{i(k)} = \arg\max_{z_{iv} \in z_i} \text{sim}(z_{i()}, t_k)$, where $\text{sim}(\cdot, \cdot)$ denotes the token embedding similarity. Intuitively, $d_{j(k)}$ and $z_{i(k)}$ correspond to the most relevant semantic concepts present in documents d_j and z_i respectively for the reasoning aspect represented by token t_k. We then define the parameterized query-sensitive similarity between documents d_j and z_i as:

$$S_\Theta(d_j, z_i, x_m) = \frac{1}{|x_m|} \sum_{k=1}^{|x_m|} \theta_{\mathcal{C}(k)} \left(\mathbf{d}^\top_{j(k)} \mathbf{z}_{i(k)} \right). \tag{5}$$

The above formulation can be interpreted as a weighted late-interaction similarity operator. Unlike standard ColBERT similarity, which independently measures query-document relevance, our formulation measures whether two documents contain semantically complementary evidence with respect to the same query aspect. The cluster-specific weights $\theta_{\mathcal{C}(k)}$ further allow the model to suppress redundant semantic facets while emphasizing reasoning-critical aspects of the query. Finally, we exponentiate the similarity score to ensure positivity and amplify highly compatible document pairs. Our final formulation for the selection problem can be obtained by replacing $\mathcal{S}_\Theta(z_i, d_j)$ with $S_\Theta(z_i, d_j, x_m)$ in equation 4.

3.3 A Bandit Scheme for Learning

The multi-hop scoring formulation introduced in the previous section depends on the learnable parameter vector $\Theta = \{\theta_1, \ldots, \theta_{|\mathcal{C}|}\}$, which controls the relative importance assigned to different semantic aspects of the query. However, directly optimizing Θ is challenging because the reward signal is observed only after expensive LLM inference over the selected document subset. Multi-armed bandit schemes (Lattimore and Szepesvári, 2020) provide sample-efficient optimization techniques for such expensive functions. We formulate the learning procedure as a linear contextual

bandit optimization problem, where the selected document subset corresponds to an action, A, and the downstream QA correctness obtained from the LLM acts as the reward signal. The overall scoring function can therefore be written compactly as: $\sigma(x_m, \alpha^* \mid \Theta) = \sum_{l=1}^{|\mathcal{C}|} \theta_l A_l(x_m) = \Theta^\top A(x_m)$. Hence, considering a question x_m and reitrived document set $\eta(x_m)$ ($L = |\eta(x_m)|$ documents), and the final score function:

$$\sigma(x_m, \alpha^* \mid \Theta) = \sum_{d_j, z_i \in \eta} \beta^*_{ij}\, \pi(x_m, z_i)\, S_\Theta(d_j, z_i, x_m). \tag{6}$$

the contextual feature vector can be defined as: $A(x_m) \in \mathbb{R}^{|\mathcal{C}|}$, whose l-th component is:

$$A_l(x_m) = \left(\sum_{d_j \in \eta} \sum_{z_i \in \eta} \beta^*_{ij}\, \pi(x_m, z_i)\, \frac{d^\top_{j(l)} z_{i(l)}}{|x_m|} \right) \tag{7}$$

if cluster l appears in the query, and 0 otherwise. Algorithm 1 describes the steps for our bandit learning scheme. Intuitively, the bandit learner gradually shifts probability mass toward semantic query aspects that consistently contribute to successful multi-hop reasoning chains, while suppressing redundant or noisy reasoning paths. Since rewards are observed only for the selected document subsets, the resulting formulation enables sample-efficient optimization using a small number of expensive LLM feedback calls. Additionally, we maintain a cache of previously evaluated document subsets and reuse stored rewards whenever identical subsets reappear during training, substantially reducing the total number of LLM evaluations required during optimization. A key distinguishing aspect of our scheme is that arm-selection uses a sparse-optimization approach similar to lasso-bandits (Ariu et al., 2022), different from confidence-based exploration in LinUCB (Li et al., 2010).

4 Experimental Results

We aim to answer the following research questions:
RQ1: Can LLM feedback-optimized lightweight scorer achieve higher recall at lower depths for context retrieval in CQA at low computational cost?
RQ2: Does (a small) LLM feedback based context retrieval lead to better (All or bigger) LLM accuracy?
RQ3: How can the 2-hop surrogate function ensure retrieval of a diverse subset of documents covering the important aspects?

Algorithm 1 BanditRAG: Bandit-based Learning of Multi-hop Scoring Function

Input: Training set $\mathcal{X} = \{(x_m, y_m)\}_{m=1}^{M}$
Output: Learned parameter vector Θ

Initialize Θ, $V_0 = \lambda I$, and $b_0 = \mathbf{0}$
for $t = 1$ to T **do**
Context-sampling step:
    Sample question-answer pair (x_m, y_m)
    Retrieve candidate document pool $\eta(x_m)$, $L = |\eta(x_m)$
Arm-selection step:
    Compute the $L \times L$ similarity matrix $\mathbf{S}_{x_m}$ using Eq. 5
    Solve problem in Eq. 4 to obtain β^*_{ij} and γ^*_i
    Construct feature vector A_t using Eq.7
Parameter Update-step:
    Query the LLM using selected documents using β^* and observe reward $\hat{v}_t$
    Update: $V_t \leftarrow V_{t-1} + A_t A_t^\top$, $b_t \leftarrow b_{t-1} + A_t \hat{v}_t$, $\Theta_t \leftarrow V_t^{-1} b_t$
end for
return Θ_T

RQ4: Does the resulting bandit scheme lead to sample efficiency and lower LLM calls?

4.1 Experimental Setup

Datasets: We experiment on two well-known multi-hop complex QA benchmarks, namely 2WikiMultihopQA (WQA) and MusiqueQA (MQA). Unlike HotpotQA, which does not mandate multi-hop reasoning and can be answered through single-hop reasoning (Min et al., 2019; Trivedi et al., 2022), MQA, WQA mandates multi-hop reasoning to arrive at the correct answer, making them ideal for benchmarking. For evaluation, we use 2-hop questions in the test set from MQA which accounts for 1252 questions. For WQA, we evaluate on the sampled subset of 1200 questions following the same setup as prior works Self-Ask (Press et al., 2023) and SUNAR (Venktesh et al.). The corpus used for retrieval is common across WQA and MQA, comprising 569,461 documents.

Setup for bandit-based training: BanditRAG

The bandit-based training of scoring function / surrogate run was conducted using a training split of 408 questions and a validation split of 300 questions from both the WQA and MQA datasets. For obtaining relevant feedback during training, we em-

Table 1: Performance Comparison on MQA and WQA . LLM calls are in the format (SubQ generation call + Training call + Testing call)

Dataset	Method	LLM Calls ($\downarrow 77\%$)	CoverEM ($\uparrow 18.57, 23.92$)	Recall@1	Recall@2	Recall@5
MQA	Searchain (Xu et al., 2024)	5.64	35.00	0.158	0.203	0.267
	SUNAR (Venktesh et al.)	16.94	40.75	0.286	0.382	0.458
	BanditRAG + Splade	3.81	41.50	0.288	0.385	0.494
	BanditRAG + BM25 w/o diversity	4.62	16.08	0.002	0.004	0.009
	BanditRAG + BM25	4.62	43.58	0.297	0.385	0.470
	BanditRAG + BM25 + own SubQ	2	43.50	0.297	0.385	0.465
WQA	Searchain (Xu et al., 2024)	7.37	54.00	0.206	0.283	0.378
	SUNAR (Venktesh et al.)	22.13	62.42	0.287	0.382	0.584
	BanditRAG + Splade	4.68	66.92	0.284	0.458	0.601
	BanditRAG + BM25 w/o diversity	5.49	29.75	0.000	0.001	0.002
	BanditRAG + BM25	5.49	70.17	0.290	0.460	0.601
	BanditRAG + BM25 + own SubQ	2	68.75	0.284	0.447	0.581

Table 2: CoverEM Results (Top-5) using GPT-4.1

Method	MQA	WQA
w/o Query Understanding		
ZERO-SHOT-COT (Kojima et al., 2023)	32.17	52.42
FEW-SHOT-COT (Wei et al., 2022a)	31.67	54.08
FEW-SHOT-COT + PRF (Li et al., 2023)	20.5	45.58
w/ Query Understanding		
Self-RAG (Asai et al., 2024)	17.80	35.25
ReAct (Yao et al., 2023)	31.25	55.25
DecomP (Khot et al., 2023)	27.67	50.83
Searchain (Xu et al., 2024)	35.00	54.00
Self-Ask (Press et al., 2023)	31.83	49.33
NAR (w/ Query Understanding)		
SUNAR (Venktesh et al.)	40.75	62.42
BanditRAG + SPLADE	**41.50**	**66.92**()
BanditRAG + BM25	43.58	70.17
BanditRAG + BM25 + SubQ	43.50	68.75

Table 3: Performance Comparison using Qwen2.5-7B-Instruct-AWQ

Method	MQA	WQA
ZERO-SHOT-COT	09.58	23.17
SUNAR	26.25	43.00
BanditRAG + SPLADE	27.42	43.50

ploy the Qwen-2.5-3B model as the feedback generator. Our method produces binary $\{0, 1\}$ feedback indicating the correctness of the answer generated using the top-5 documents selected by the 2-hop surrogate scoring function. To compute π scores, we utilize the *cross-encoder/ms-marco-MiniLM-L12-v2*. For computing S, we use contextualized embeddings from *bert-base-uncased*. During training, we additionally maintain a cache of previously evaluated documents to reduce LLM calls. As illustrated in Figure 2, the number of new LLM calls progressively decreases across epochs. Overall, the complete training process for 14 epochs requires approximately 970 LLM calls. Furthermore, for query decomposition, we employ GPT-4.1. During inference, we utilize sub-question-based cross-encoder scores along with pairwise similarity scores as input to the learned scoring function. Based on thes signals, the multi-hop surrogate /scoring function ranks the candidate documents and selects the top-5 documents.

Baselines: We first compare with CoT based approaches without any query understanding like Zero-shot-CoT (Kojima et al., 2023), Few-shot CoT (Wei et al., 2023) where evidences are retrieved using SPLADEV2 with respect to original question. We compare SUNAR with state-of-the-art iterative retrieval and planning based approaches like Self-Ask (Press et al., 2023), Searchain (Xu et al., 2024), query-decomposition based dynamic retrieval approaches like DecomP (Fu et al., 2021) and LLMs fine-tuned using self-correction like Self-RAG (Asai et al., 2024). We use a temperature value of 0.3 following setup of prior works (Venktesh et al.) to ensure diversity in generated decomposition and reasoning. For all baselines and our approach during inference we use SPLADEV2 (Formal et al., 2021a) as the retriever.

4.2 Retrieval, Complex QA performance and efficiency

To answer **RQ 1**, we first report the retrieval performance in Table 1, of BanditRAG and compare it to SUNAR to ensure that removing the expensive LLM feedback in BanditRAG does not affect retrieval performance. From Table 1, we observe that BanditRAG achieves higher recall at lower depths.

For instance, on WQA for BanditRAG achieves Recall@2 of **0.458** outperforming Self-Ask which achieves Recall@2 of **0.283** by **61.8%**. We also observe similar gains for Recall@5.

The significance of our approach BanditRAG lies in it's efficiency, achieving higher recall at lower depths, without the expensive LLM feedback at inference time, unlike SUNAR. Hence, for instance, BanditRAG reduces the total number of LLM calls by up to **77.44%** on MQA during inference on all test samples. When comparing the average number of LLM calls per query, in Figure 1, we observe that on average BanditRAG requires **5 x** fewer calls than SUNAR per query across both datasets, while achieving similar or better recall on both WQA and MQA as observed by Recall@{1,2,5} in Table 1.

This is primarily because our bandit-based approach captures multi-hop document-document relatedness and also models the multi-aspect nature of the query. This enables the scoring function to model document-document relatedness (clustering hypothesis), while ensuring the documents that are ranked higher also cover diverse aspects of the query. This sample-efficient learning mechanism, which happens once offline obviates the need for expensive LLM feedback at inference time.

To answer **RQ 2**, we observe the answer generation performance on MQA and WQA in Table 2. We observe that BanditRAG significantly outperforms non query-understanding based approaches like zero-shot and few-shot CoT. BanditRAG also significantly outperforms other iterative retrieval and query understanding based approaches like React, Searchain, Self-Ask, as we observe from Table 2, that BanditRAG even achieves a performance gain of **7.21 %** on answer accuracy as measured by Cover EM on WQA. We observe that the scoring function BanditRAG preserves the diversity of the final top-k (k=5) prioritized documents during inference. While, SUNAR sometimes may prioritize homogeneous documents that are related to each other based on LLM self-consistency-based feedback, impacting QA performance.

We also perform several ablations, such as BanditRAG + BM25, where instead of SPLADEV2 for first-stage retrieval, we use BM25. We observe that with BM25 docs BanditRAG demonstrates similar performance gains.

Additionally, we also perform inference using smaller language models like Qwen-2.5:7B on Table 3, demonstrating that BanditRAG achieves similar performance or gains compared with SUNAR with less computational cost across diverse LLMs.

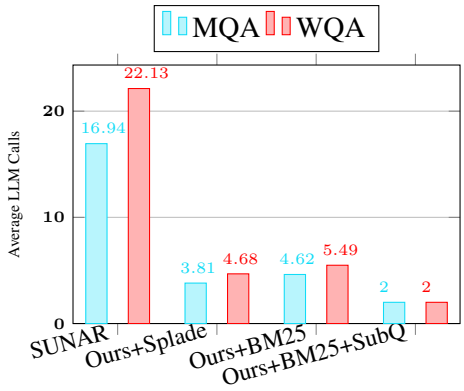

Figure 1: Comparison of total LLM calls across retrieval methods on MQA and WQA datasets.

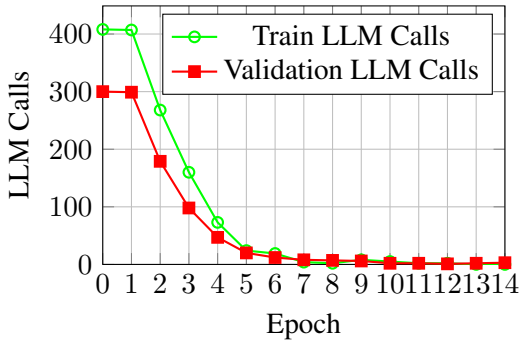

Figure 2: Reduction in LLM calls across epochs due to caching.

4.3 Impact of BanditRAG on diversity and coverage of evidence

To answer **RQ 3**, we perform ablations in the training setup of the scoring function. We first test the ablation of removing the bandit-based algorithm and directly re-ranking the candidate evidences using cross-encoder scores as signal, as indicated by cross-encoder row in Table 4. The results, show that BanditRAG outperforms purely cross-encoder approach by **30.21%** on WQA and **29.75%** on MQA. Moreover, removing the diversity-inducing component in the bandit-based learning of the scoring function also results in severe degradation answer accuracy (table 4)

A qualitative example of the advantage of diversity induced through the bandit-based training of the scoring function is shown in Table 7. We observe that BanditRAG is able to retrieve the key documents covering the diverse aspects of the complex query as shown in Table 7. Here, Doc 1 an-

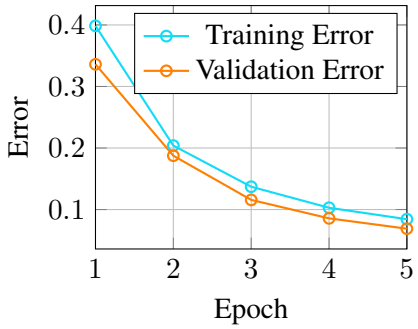

Figure 3: Training vs Validation Error

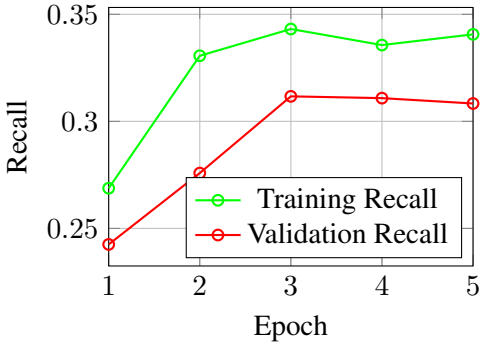

Figure 4: Training vs Validation Recall

swers the sub-query regarding "publisher of the newspaper North Star", Doc 4 and Doc 5 retrieve the necessary documents for answering the final subquery regarding the identity of the person who married the publisher. However, SUNAR only manages to recover redundant documents related to the publisher, which is only useful in answering a part of the complex question due to it's overreliance on self-consistency based LLM feedback during inference without accounting for the diverse aspects of the query during final document ranking and consolidation.

Table 4: Effect of Diversity in Bandit Retrieval using GPT-4.1

Method	WQA	MQA
Cross-Encoder (question based)	54.08	33.47
BanditRAG without diversity	29.75	16.08
BanditRAG + BM 25	70.17	43.58

4.4 Sample-efficiency of proposed approach when learning scoring function

To answer **RQ 4**, we analyze the bandit-based training dynamics for the scoring function. From Figure 3, we observe that training and validation errors decrease over rounds, demonstrating that the bandit-based optimization works and results in convergence of the scoring function. We also observe

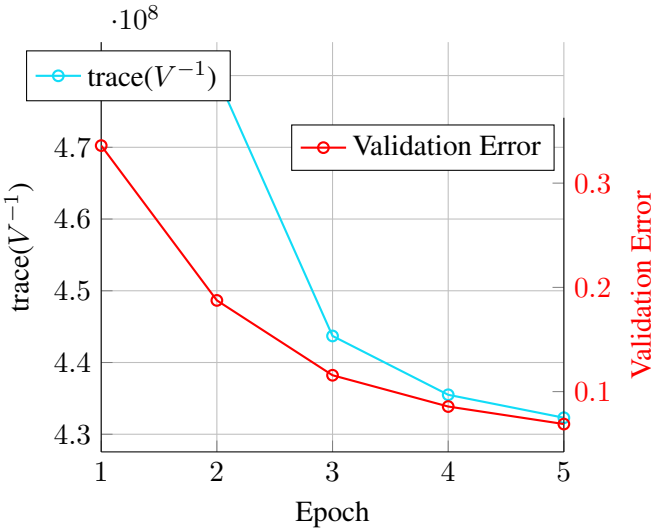

Figure 5: Confidence Shrinkage vs Validation Error

that the corresponding Recall of retrieved documents improves over the rounds as observed in Figure 4, demonstrating that the LLM feedback helps improve Recall which in turn positively impacts answer generation accuracy. To demonstrate that the bandit-based optimization converges and is correlated with answer generation performance, we compare the trends in the change of the confidence interval in the bandit algorithm and the validation error in Figure 5. We observe that confidence interval shrinks as the LLM feedback for different subsets of documents are sampled over time. This directly impacts with the validation error as shown in the Figure 5, where as confidence intervals shrinks, the validation error also decreases. This demonstrates that BanditRAG helps learn a useful scoring function that prioritizes documents, improving performance for CQA.

5 Conclusion

In this work, we tackle the challenge of improving context retrieval at lower depths, while ensuring diversity of the context pool, and without the need for expensive compute for complex QA. We observe that existing state-of-the-art approaches require expensive multi-turn LLM feedback and need to rank large number of documents. To tackle this, we propose a bandit-based learning scheme of a lightweight scoring function that learns to score documents accounting for multi-aspect nature of the query ensuring diversity and relevance of documents. This scoring function is trained offline and obviates the need for expensive LLM feedback.

6 Limitations

BanditRAG offer a modest preformance improvement than SUNAR(Venktesh et al.) and a significant amount of the other off-the-shelf RAG methods with significant computation reduction by alleviating the expensive LLM calls and still the feedback quality is dependent on the reasoning capability of the underlying LLM used during training, leading to more number of LLM calls. Employing a larger and more capable language model could potentially generate more accurate feedback signals, leading to faster improvements in both training and validation accuracy. This may also reduce the number of iterative weight updates and consequently decrease the total number of LLM calls required during optimization. Another drawback of our current approach is that it still uses cross-encoder-based signals between sub-question-document pairs as features for the scoring function. Though we sample fewer cross-encoder scores compared to state-of-the-art methods, alternative lightweight signals could further reduce computational demands. Exploring alternative lightweight signals is left for future work.

7 Ethical Considerations

We use only publicly available QA datasets from Wikipedia that do not contain private or harmful information. While we employ LLMs for QA systems in our experiments, we do not prompt them in a format that would elicit harmful or biased information.

References

Kaito Ariu, Kenshi Abe, and Alexandre Proutière. 2022. Thresholded lasso bandit. In *International Conference on Machine Learning*, pages 878–928. PMLR.

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 17 of *SIGIR 2024*, page 719–729. ACM.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021a. Splade v2: Sparse lexical and expansion model for information retrieval.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021b. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Ruiliu Fu, Han Wang, Xuejun Zhang, Jun Zhou, and Yonghong Yan. 2021. Decomposing complex questions makes multi-hop QA easier and more interpretable. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 169–180, Punta Cana, Dominican Republic. Association for Computational Linguistics.

N. Jardine and Cornelis Joost van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Inf. Storage Retr.*, 7(5):217–240.

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, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Hrishikesh Kulkarni, Sean MacAvaney, Nazli Goharian, and Ophir Frieder. 2023. Lexically-accelerated dense retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 152–162. ACM.

Tor Lattimore and Csaba Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Hang Li, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, Jimmy Lin, and Guido Zuccon. 2022. Improving query representations for dense retrieval with pseudo relevance feedback: A reproducibility study. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, page 599–612, Berlin, Heidelberg. Springer-Verlag.

Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2024. Self-prompting large language models for zero-shot open-domain QA. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 296–310, Mexico City, Mexico. Association for Computational Linguistics.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 661–670.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023. Few-shot in-context learning for knowledge base question answering.

Sean MacAvaney, Nicola Tonellotto, and Craig Macdonald. 2022. Adaptive re-ranking with a corpus graph. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 1491–1500. ACM.

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2022. A survey on multi-hop question answering and generation.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

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

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. 2019. Answering complex open-domain questions through iterative query generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2590–2602, Hong Kong, China. Association for Computational Linguistics.

Mandeep Rathee, Sean MacAvaney, and Avishek Anand. 2024. Quam: Adaptive retrieval through query affinity modelling. *arXiv preprint arXiv:2410.20286*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2024. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7339–7353, Bangkok, Thailand. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

V Venktesh, Mandeep Rathee, and Avishek Anand. Sunar: Semantic uncertainty based neighborhood aware retrieval for complex qa.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

A Dataset

We evaluate on MusiqueQA and 2WikimultihopQA released under CC-BY-4.0 LICENSE

2WikiMultihopQA (WQA): Wikipedia based multi-hop questions that consists of compositional queries and comparison based queries. We evaluate on 1200 compositional questions as done in self-ask (Press et al., 2023) for a fair comparison. Example Question: Who was born later, Gideon Johnson or Holm Jølsen?

MuSiQue (MQA) Comprises hard multi-hop compositional questions constructed from Wikipedia-based QA datasets. The dataset was constructed from multiple single-hop datasets to result in complex questions that mandate multistep reasoning. MQA is a challenging dataset created to avoid possible shortcuts in existing multi-hop datasets like HotPotQA, where models do not perform compositional reasoning. We evaluate on 1252 compositional questions as done in self-ask (Press et al., 2023) for a fair comparison. Example Question: What did the actress in My Fair Lady win a Tony for ?

B Training and Validation

Table 5 presents the training and validation performance across epochs. We observe that both training and validation accuracy improve comparatively low during the initial epochs and remain relatively stable thereafter, indicating consistent learning behavior. The improved accuracy is not very high due to the limitation of using the small LLM, where even with good context due to limitation of capability to reason over good documents, but later that has been improved by using a larger LLM. Recall improves progressively, suggesting that BanditRAG becomes increasingly effective at identifying relevant documents over training iterations. Additionally, both training and validation errors decrease monotonically across epochs, demonstrating stable optimization and convergence of the proposed feedback-driven training procedure. Table 6 reports the correlation between prediction error and confidence interval estimates for both training and validation sets. The strong Pearson correlations indicate a near-linear relationship between error and confidence estimates, while the perfect Spearman correlations demonstrate a highly consistent monotonic relationship. These results suggest that the confidence interval produced by the model is strongly aligned with retrieval error, making it a reliable signal for guiding adaptive feedback and weight updates during training.

Table 5: Training and Validation Performance Across Epochs

Epoch	Train Acc.	Val Acc.	Train Recall	Val Recall	Train Error	Val Error
1	0.4125	0.346667	0.268750	0.242500	0.398524	0.335954
2	0.4175	0.363333	0.330625	0.275833	0.203943	0.187538
3	0.4200	0.356667	0.343125	0.311667	0.137094	0.115733
4	0.4200	0.350000	0.335625	0.310833	0.102944	0.085707
5	0.4300	0.353333	0.340625	0.308333	0.084346	0.068921

Table 6: Correlation of Error with Confidence Interval

Correlation Type	Train Error	Validation Error
Pearson Correlation	0.9781	0.9871
Spearman Correlation	1.0000	1.0000

Case Study on MQA: Comparison Between BanditRAG and SUNAR	
Question	Who married the publisher of abolitionist newspaper *The North Star*?
BanditRAG	Gold: `Helen Pitts Douglass` Pred: `Helen Pitts Douglass.` Status: ✓
SUNAR	Gold: `Helen Pitts Douglass` Pred: `Anna Murray-Douglass.` Status: ×
Top-5 Documents Selected by BanditRAG	
Doc 1	The North Star was a nineteenth-century anti-slavery newspaper published from the Talman Building in Rochester, New York by abolitionist *Frederick Douglass*. The paper commenced publication on December 3, 1847 and ceased as *The North Star* in June 1851 when it merged with Gerrit Smith's Liberty Party Paper to form *Frederick Douglass' Paper*.
Doc 2	William Cooper Nell (1816–1874) was an African-American abolitionist, journalist, publisher, author, and civil servant. Writing for abolitionist newspapers *The Liberator* and *The North Star*, he helped publicize the anti-slavery cause.
Doc 3	W.F. Herman (died 1938) was a Canadian newspaper publisher and owner of the *Windsor Star*.
Doc 4	*Helen Pitts Douglass* (1838–1903) was an American suffragist and abolitionist, known for being the second wife of *Frederick Douglass*. She also created the Frederick Douglass Memorial and Historical Association.
Doc 5	*Helen Pitts Douglass* (1838–1903) was an American suffragist and abolitionist, known for being the second wife of *Frederick Douglass*.
SUNAR Evidences	
Doc 1	The North Star was a nineteenth-century anti-slavery newspaper published from the Talman Building in Rochester, New York by abolitionist Frederick Douglass.
Doc 2	William Cooper Nell (1816–1874) was an African-American abolitionist and journalist who wrote for *The North Star*.
Doc 3	Anna Murray-Douglass (1813–1882) was the first wife of Frederick Douglass.
Doc 4	The North Star was a nineteenth-century anti-slavery newspaper published by Frederick Douglass.

Table 7: Comparison between BanditRAG and SUNAR on a question where BanditRAG successfully retrieves evidence identifying Helen Pitts Douglass, while SUNAR predicts the incorrect answer Anna Murray-Douglass. Relevant supporting evidence is highlighted in blue.

Case Study: Comparison Between BanditRAG and SUNAR	
Question	What is the place of birth of Claudia Rodríguez De Castellanos's husband?
BanditRAG	Gold: `Bogotá` Pred: `Bogotá, Colombia` Status: ✓
SUNAR	Gold: `Bogotá` Pred: `Honduras` Status: ×
Top-5 Documents Selected by BanditRAG	
Doc 0	Claudia Yadira Inés Rodríguez is a Colombian Evangelical pastor and politician. She is married to *César Castellanos Domínguez*, an Evangelical pastor and leader of the International Charismatic Mission Church.
Doc 1	Fernando Rodríguez de Castro (1125–1185) was a Castilian nobleman, statesman, and military leader who made his career in León.
Doc 2	Antoinette Vilma Reyes Escalante is a former Honduran politician who was married to Cesar Castellanos Madrid, former mayor of the Central District of Honduras.
Doc 3	*César Castellanos* is a neo-charismatic pastor from *Bogotá, Colombia* and the founder of the Misión Carismática Internacional.
Doc 4	Claudia Alejandra Ledesma Abdala de Zamora is an Argentine politician and wife of Gerardo Zamora.
SUNAR Evidences	
Evidence 0	Claudia Yadira Inés Rodríguez is a Colombian Evangelical pastor and politician. She is married to *César Castellanos Domínguez*, an Evangelical pastor and leader of the International Charismatic Mission Church.
Evidence 1	Fernando Rodríguez de Castro (1125–1185) was a Castilian nobleman, statesman, and military leader.
Evidence 2	Antoinette Vilma Reyes Escalante was married to Cesar Castellanos Madrid, who became mayor of the Central District of Honduras.
Evidence 3	Augustus was the first emperor of the Roman Empire and founder of the Principate.
Evidence 4	Claudia Yadira Inés Rodríguez is married to César Castellanos Domínguez, leader of the International Charismatic Mission Church.
Evidence 5	Castellanos is a Spanish habitational surname associated with Castilians.

Table 8: Comparison between BanditRAG and SUNAR on a multi-hop question requiring entity linking and birthplace reasoning. BanditRAG successfully retrieves evidence identifying César Castellanos as being from Bogotá, Colombia, while SUNAR is distracted by an unrelated Honduras-based entity sharing a similar name.

BanditRAG and SUNAR Prompt

```
[Question]: When does monsoon season end in the state the area code 575 is
located?
[Answer]: The area code 575 is located in New Mexico. Monsoon season in New
Mexico typically ends in mid-September.
[Final Answer]: mid-September.
[Question]: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
[Answer]: Theodor Haecker was 65 years old when he died. Harry Vaughan Watkins
was 69 years old when he died. Hence Harry Vaughan Watkins lived longer.
[Final Answer]: Harry Vaughan Watkins.
[Question]: What is the current official currency in the country where
Ineabelle Diaz is a citizen?
[Answer]: Ineabelle Diaz is from Puerto Rico, which is in the United States of
America. The current official currency in the United States is the United
States dollar.
[Final Answer]: United States dollar.
[Question]: Where was the person who founded the American Institute of Public
Opinion in 1935 born?
[Answer]: The person who founded the American Institute of Public Opinion in
1935 is George Gallup. George Gallup was born in Jefferson, Iowa.
[Final Answer]: Jefferson.
[Question]: What language is used by the director of Tiffany Memorandum?
[Answer]: The director of Tiffany Memorandum is Sergio Grieco. Sergio Grieco
speaks Italian.
[Final Answer]: Italian.
[Question]: What is the sports team the person played for who scored the first
touchdown in Superbowl 1?
.....
[Question]: The birth country of Jayantha Ketagoda left the British Empire
when?
[Answer]: The birth country of Jayantha Ketagoda is Sri Lanka. Sri Lanka left
the British Empire on February 4, 1948.
[Final Answer]: February 4, 1948.
Follow the above examples carefully.
Existing decomposition reasoning path:
{decomposition_path}
Evidence:
{evidence_text}
Use the evidence carefully and generate your own reasoning path step by step.
Question:
{question}
Your response MUST follow this format:
[Answer]: <step by step reasoning>
[Final Answer]: <final short answer>
```

Subquestion Generator Prompt Template

```
SYSTEM: You are an expert at decomposing multi-hop questions into useful
subquestions.
Your task:
1. Read the question carefully.
2. Break the question into meaningful subquestions.
3. If you already know parametric/world knowledge related to the entities in
the
   question, use that knowledge while generating subquestions.
4. Generate ONLY the subquestions.
5. Keep the subquestions concise and retrieval-friendly.
6. Return subquestions separated by commas.

Example 1:
Question:
Which film came out first, The Love Route or Engal Aasan?
Subquestions:
What is the release date of The Love Route?,
What is the release date of Engal Aasan?

Example 2:
Question:
Do both films The Falcon (Film) and Valentin The Good have the directors from
the same country?
Subquestions:
Do both films "The Falcon" and "Valentin the Good" have directors from the
same country?,
Who directed the film "Valentin the Good"?,
What is Martin Fric's nationality?,
Who directed the film "The Falcon"?,
What is the nationality of William Clemens?,
What is the nationality of Maurice Geraghty?

Example 3:
Question:
What nationality is the director of film Wedding Night In Paradise (1950
Film)?
Subquestions:
What is the title of the 1950 West German musical comedy film directed by Geza
von Bolvary?,
Who directed the 1950 West German musical comedy film "Wedding Night in
Paradise"?,
What is Geza von Bolvary's nationality?

Example 4:
Question:
When is the composer of film Sruthilayalu's birthday?
Subquestions:
Who directed the film Sruthilayalu?,
What is the birth year of K. Viswanath?

Example 5:
Question:
What nationality is the director of film Borunbabur Bondhu?
Subquestions:
Who directed the film Borunbabur Bondhu?,
What is Anik Dutta's nationality?,
Where was Anik Dutta born?

Now decompose the following question into subquestions.
Question:
{question}
```

DecomP WikiMultiHop Prompt

```
SYSTEM: Solve a question answering task with interleaving retrieval.
Follow the given examples and given the question and context, decompose the
original question to follow-up questions if needed, use relevant evidence for
each subquestion from the given list of evidences, answer the subquestion and
output the final answer for the question using information in the context,
answers to subquestions, and give final answer strictly preceded by [Final
Answer]:

[Question]: In which country did this Australian who was detained in
Guantanamo
Bay detention camp and published "Guantanamo: My Journey" receive
para-military training?
Sub-question 1: (select) [retrieve odqa] Who is the Australian who was
detained
in Guantanamo Bay detention camp and published "Guantanamo: My Journey"?
Answer: {"titles": ["Guantanamo: My Journey", "Bismullah v. Gates",
"Guantanamo Bay detention camp"], "answer": ["David Hicks"]}
Sub-question 2: (select) [retrieve odqa] In which country did David Hicks
receive his para-military training?
Answer: {"titles": ["John Adams Project", "Camp Echo (Guantanamo Bay)",
"Guantanamo Bay Museum of Art and History", "David Hicks"],
"answer": ["Afghanistan"]}
Sub-question 3: (select) [multihop titleqa] Titles: ["Guantanamo: My Journey",
........

[Question]: How many awards did the "A Girl Like Me" singer win at the
American
Music Awards of 2012?
........
[Question]: What was the 2014 population of the city where Lake Wales Medical
Center is located?
Sub-question 1: (select) [retrieve odqa] Lake Wales Medical Center is located
in what city?
Answer: {"titles": ["Baylor College of Medicine", "Lake Wales Medical Center",
"Tufts University School of Medicine", "Hanford Community Medical Center"],
"answer": ["Polk County, Florida"]}
Sub-question 2: (select) [retrieve odqa] What was the population of Polk
County
in 2014?
Answer: {"titles": ["Banner University Medical Center Tucson",
"Lake Wales, Florida"], "answer": ["15,140"]}
Sub-question 3: (select) [multihop titleqa] Titles: ["Baylor College of
Medicine",
"Lake Wales Medical Center", "Tufts University School of Medicine",
"Hanford Community Medical Center", "Banner University Medical Center Tucson",
"Lake Wales, Florida"]. Question: What was the 2014 population of the city
where Lake Wales Medical Center is located?
[Final Answer]: 15,140
...........
Follow the above examples, and use the given Evidence:
{evidence_text}
Use the information from the evidence to answer the original question and give
the final answer strictly preceded by [Final Answer]: for the Question:
{question}
```

Searchain Prompt

```
SYSTEM_MSG = """Solve a question answering task with interleaving Thought,
Action, Observation steps. Thought can reason about the current situation, and
Action can be three types:
(1) Search[entity], which searches the exact entity on Wikipedia and returns
the first paragraph if it exists. If not, it will return some similar entities
to search.
(2) Lookup[keyword], which returns the next sentence containing keyword in the
current passage.
(3) Finish[answer], which returns the answer and finishes the task.
Follow the given examples and Given the question and context decompose the
original question to followup questions if needed, use relevant evidence for
each subquestion from given list of evidences, answer the subquestion and
output final answer for the question using information in the context, answers
to subquestions and give answer in form of [Final Answer]: \n"""

FEW_SHOT = """Construct a global reasoning chain for this complex [Question] :
" {} " You should generate a query to the search engine based on
what you already know at each step of the reasoning chain, starting with
[Query].
If you know the answer for [Query], generate it starting with [Answer].
You can try to generate the final answer for the [Question] by referring to
the [Query]-[Answer] pairs, starting with [Final Answer].
If you don't know the answer, generate a query to search engine based on what
you already know and do not know, starting with [Unsolved Query].
For example:
[Question]: "Where do greyhound buses that are in the birthplace of Spirit
If...'s performer leave from?"
[Query 1]: Who is the performer of Spirit If... ?
If you don't know the answer:
[Unsolved Query]: Who is the performer of Spirit If... ?
If you know the answer:
[Answer 1]: The performer of Spirit If... is Kevin Drew.
[Query 2]: Where was Kevin Drew born?
If you don't know the answer:
[Unsolved Query]: Where was Kevin Drew born?
If you know the answer:
[Answer 2]: Toronto.
[Query 3]: Where do greyhound buses in Toronto leave from?
If you don't know the answer:
[Unsolved Query]: Where do greyhound buses in Toronto leave from?
If you know the answer:
[Answer 3]: Toronto Coach Terminal.
[Final Content]: The performer of Spirit If... is Kevin Drew [1]. Kevin Drew
was born in Toronto [2]. Greyhound buses in
Toronto leave from Toronto Coach Terminal [3]. So the final answer is Toronto
Coach Terminal.
[Final Answer]: Toronto Coach Terminal

[Question]: "Which magazine was started first Arthur's Magazine or First for
Women?"
[Query 1]: When was Arthur's Magazine started?
[Answer 1]: 1844.
[Query 2]: When was First for Women started?
[Answer 2]: 1989
[Final Content]: Arthur's Magazine started in 1844 [1]. First for Women
started in 1989 [2]. So Arthur's Magazine was started first.
[Final Answer]: Arthur's Magazine.
"""
```

ReAct Prompt

```
SYSTEM_MSG = """Solve a question answering task with interleaving Thought,
Action, Observation steps. Thought can reason about the current situation, and
Action can be three types:
(1) Search[entity], which searches the exact entity on Wikipedia and returns
the first paragraph if it exists. If not, it will return some similar entities
to search.
(2) Lookup[keyword], which returns the next sentence containing keyword in the
current passage.
(3) Finish[answer], which returns the answer and finishes the task.
Follow the given examples and Given the question and context decompose the
original question to followup questions if needed, use relevant evidence for
each subquestion from given list of evidences, answer the subquestion and
output final answer for the question using information in the context, answers
to subquestions and give answer in form of [Final Answer]: \n"""

FEW_SHOT = """Question: What is the elevation range for the area that the
eastern sector of the Colorado orogeny extends into?
Thought 1: I need to search Colorado orogeny, find the area that the eastern
sector of the Colorado orogeny extends into, then find the elevation range of
the area.
Action 1: Search[Colorado orogeny]
Observation 1: The Colorado orogeny was an episode of mountain building (an
orogeny) in Colorado and surrounding areas.
Thought 2: It does not mention the eastern sector. So I need to look up
eastern sector.
...........
Action 3: Search[High Plains]
Observation 3: High Plains refers to one of two distinct land regions:

..........
Thought 3: Adam Clayton Powell (film) is a documentary about an
African-American politician, not Finnish rock groups. So the documentary about
Finnish rock groups must instead be The Saimaa Gesture.
Action 3: Finish[The Saimaa Gesture]
[Final Answer]: The Saimaa Gesture

Question: What profession does Nicholas Ray and Elia Kazan have in common?
Thought 1: I need to search Nicholas Ray and Elia Kazan, find their
professions, then find the profession they have in common.
Action 1: Search[Nicholas Ray]
Observation 1: Nicholas Ray (born Raymond Nicholas Kienzle Jr., August 7, 1911
- June 16, 1979) was an American film director, screenwriter, and actor best
known for the 1955 film Rebel Without a Cause.
Thought 2: Professions of Nicholas Ray are director, screenwriter, and actor.
I need to search Elia Kazan next and find his professions.
Action 2: Search[Elia Kazan]
Observation 2: Elia Kazan was an American film and theatre director, producer,
screenwriter and actor.
Thought 3: Professions of Elia Kazan are director, producer, screenwriter, and
actor. So profession Nicholas Ray and Elia Kazan have in common is director,
screenwriter, and actor.
Action 3: Finish[director, screenwriter, actor]
[Final Answer]: director, screenwriter, actor
......

"""
```