FRAG: FILTERING NOISE USING SNIPPET-LEVEL QUERY RELEVANCE

Anonymous authors

Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation (RAG) augments large language models (LLMs) with external retrievals. Typically, expanding the retrieval window can improve RAG performance by retrieving more relevant content. However, it risks increased noise, which distracts the model's attention and degrades accuracy. To mitigate this, we propose Fine-Grained RAG (FRAG), which identifies key snippets from query and extracts relevant information while filtering noise from retrievals using snippet-level query relevance. Yet, a new challenge arises in addressing complex RAG queries, which require knowledge pieces with implicit multi-hop logical relationships. Failure to identify these relationships may lead to loss of inferencebased knowledge during filtering, degrading performance. To address this, we propose Self-Recognition, which extracts inference-based knowledge by leveraging historically extracted knowledge as contextual references. While FRAG notably improves performance, it incurs additional latency. To alleviate this, we present FRAG-ip, a fine-tuned framework which markedly accelerates FRAG by approximately 10×. Extensive experiments show that FRAG significantly boosts RAG, yielding average accuracy gains of 4.94%/13.44% on simple/complex tasks.

1 Introduction

Although the performance of LLM has seen breakthrough improvements, the issues of hallucination and outdated knowledge remain persistent challenges for LLM (Xu et al., 2024b). The RAG method (Lewis et al., 2020; Guu et al., 2020), by retrieving query-relevant knowledge from external knowledge bases as the generation context, effectively addresses these issues and significantly enhances LLM performance in knowledge-intensive tasks (Ram et al., 2023).

RAG typically employs a retriever to retrieve N passages (N denotes the retrieval window size) and combines them into a retrieval document as the generation context (Gao et al., 2023). Expanding the retrieval window is a common approach to enhance Retrieval-Augmented Language Model (RALM) performance by including more relevant knowledge (Lewis et al., 2020). However, this also introduces noise due to the text or vector similarity methods that retrievers rely on (Gao et al., 2023), which may degrade RALM accuracy (Yu et al., 2024b;c; Zhu et al., 2024), as excessive noise distracts RALM's attention and ultimately leads to incorrect responses (Theorem 2.2).

Existing works aim to reduce noise. SelfRAG mitigates it via multiple generations in shorter contexts (Asai et al., 2024), but neglects logical relationships across passages, limiting performance on complex RAG tasks (Zhang et al., 2024). RECOMP uses a compressor to compress the retrievals and filter noise (Xu et al., 2024a), yet fails to retain sufficient relevant information, resulting in suboptimal performance. RankRAG reranks the retrievals to select the top-k relevant ones (Yu et al., 2024c). Yet, it faces a trade-off: a small k may lose information while a large k retains noise.

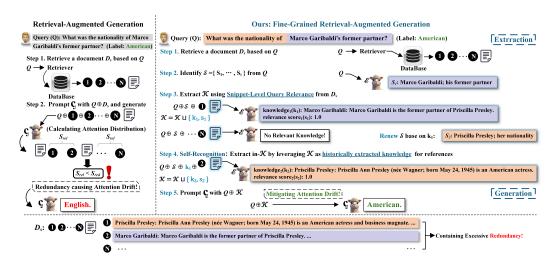


Figure 1: **Overview of FRAG.** \mathcal{K} represents the extracted relevant knowledge. \mathcal{S} denotes key sequence snippets of the query. \mathcal{E} and \mathcal{G} are the extractor and generator models, respectively. S_{rel} and S_{red} represent the total attention weights distributed to relevant information and noise in the generation context. Sentences highlighted with background color and color correspond to distinct snippets in \mathcal{S} or associated relevant information.

To mitigate the negative impact of noise, we propose *Fine-Grained RAG (FRAG)*, which uses an extractor model to construct the generation context by performing fine-grained extraction of relevant knowledge and filtering noise from initial retrievals. *Firstly*, we observe that certain sequence snippets in the query can serve as key indicators for extracting relevant knowledge, which we refer to as "*key sequence snippets*". To this end, we identify key sequence snippets from the query, incorporate them into the extraction process, and filter noise using snippet-level query relevance (Theorem 3.2). To alleviate the attention distraction caused by noise, we decompose the extraction into steps, each targeting a single, shorter-context passage (Appendix B.1).

Nevertheless, new challenges arise. Complex RAG tasks, especially multi-hop reasoning, require multiple pieces of knowledge with implicit multi-hop relationships (Mavi et al., 2022). Some of the relevant knowledge, e.g., *inference-based knowledge*, can only be extracted if corresponding prior knowledge is available (Li & Peng, 2023). However, decomposing extraction into multiple steps may omit prerequisite knowledge and corresponding relationships, hindering inference-based knowledge extraction. To address this, *secondly*, we propose *Self-Recognition* (Section 3.2), which leverages historically extracted knowledge as reference in extraction to restore missing logical relationships.

While FRAG markedly enhances RALM performance, it incurs high computational cost due to frequent LLM calls for extraction. To address this, we propose FRAG-ip, a wrapper framework that employs dual-stage fine-tuning and markedly accelerates FRAG by approximately $10 \times$ (Section 3.3).

Extensive experiments validate FRAG effectiveness in extracting relevant knowledge and filtering noise from initial retrievals. On the test dataset, FRAG retains over 90% of the golden passages and filters over 80% noise (Table 3). Compared to concurrent baselines, FRAG achieves *state-of-the-art* performance, yielding average accuracy gains of 4.94%/13.44% on simple/complex tasks over naive RAG for the best-performing test model (Table 2). Our contributions are as follows:

- We demonstrate the negative impact of noise on RAG performance (Theorem 2.2), motivating the promising direction of extracting relevant information and filtering noise from the initial retrieved documents to enhance RAG accuracy.
- We propose FRAG, which introduces snippet-level query relevance to effectively filter noise (Theorem 3.2) and incorporates Self-Recognition to enhance inference-based knowledge extraction in complex tasks. To mitigate the computational cost, we present a wrapper-based fine-tuning framework that improves efficiency by approximately 10×.
- Extensive experiments show that FRAG markedly boosts RALM performance, demonstrating state-of-the-art performance. For the best-performing model, FRAG yields average accuracy gains of 4.94%/13.44% on simple/complex tasks compared to naive RAG (Table 2).

2 FOUNDATIONS: HOW NOISE PRODUCES A NEGATIVE IMPACT

Preliminary. Given the embedding matrix of an input sequence for RAG generation: $X = X_Q \oplus X_d$, the sequence consists of a query X_Q and a retrieval document X_d . Here, $X_d = X_{\rm rel} \oplus X_{\rm red}$, and: $X \in \mathbb{R}^{n \times d}$, $X_Q \in \mathbb{R}^{n_q \times d}$, $X_{\rm rel} \in \mathbb{R}^{n_{\rm rel} \times d}$, and $X_{\rm red} \in \mathbb{R}^{n_{\rm red} \times d}$, where $X_{\rm rel}$ represents the portion of the retrieved document D_s that is relevant to Q, while $X_{\rm red}$ stands for the irrelevant portion (i.e., noise). $n, n_q, n_{\rm rel}$ and $n_{\rm red}$ denote the token counts of $X, X_Q, X_{\rm rel}$ and $X_{\rm red}$, respectively. d signifies the dimension of the embedding vectors. " \oplus " indicates the concatenation of different vectors along the vertical (row-wise) dimension in natural language order. $\mathfrak T$ refers to the first output token.

For simplicity, we consider: the RALM lacks parameterized knowledge to answer Q, relying solely on $X_{\rm rel}$ for the required information; the RALM follows the correlation paradigm where the average cosine similarity satisfies $\tilde{S}_{(X_Q,X_{\rm rel})} > \tilde{S}_{(X_Q,X_{\rm rel})}$; the RALM's attention focuses on X_Q and its correlation with other input sequence parts. Additionally:

Assumption 2.1. The Word Mover's Distance (WMD) between X_{rel} and X_{red} satisfies the inequality (Kusner et al., 2015): $WMD_{(X_{\text{rel}}, X_{\text{red}})} \gg \tau, \tau \in \mathbb{R}^+$.

This indicates that the semantic distinction between X_{rel} and X_{red} is sufficiently pronounced, allowing the computation of RALM's attention distribution across different segments of X_d (i.e., generating based on X_{rel} or X_{red}) to be approximated by the attention matrix $\mathbf{A}_{\mathfrak{T}}$ computation for \mathfrak{T} .

Theorem 2.2. In the input sequence of the retrieved document $X_d = X_{rel} \oplus X_{red}$, as the token count of irrelevant content X_{red} , i.e., n_{red} , increases, there exists $\tilde{n} \in \mathbb{N}^+$ such that when $n_{red} > \tilde{n}$, we have: $\mathfrak{A}_{rel} < \mathfrak{A}_{red}$, where \mathfrak{A}_{rel} , \mathfrak{A}_{red} denote the total attention distribution weights allocated to X_{rel} , X_{red} based on X_Q (the proof is provided in Appendix C.1).

Namely, noise negatively impacts RALM generation by distracting the model's attention, causing it to shift toward irrelevant content and fail to generate correct answers based on relevant knowledge. This theorem aligns with Zhu et al. (2024), who showed from the information bottleneck perspective that noise may degrade RALM accuracy by interfering with generation.

Remark 2.3. The negative impact motivates us to extract relevant knowledge from the initial retrieval document, thereby filtering out noise. Through reducing attention drift during RAG generation, FRAG mitigates noise-induced adverse effects and improves RALM accuracy.

3 FRAG: FILTERING NOISE USING SNIPPET-LEVEL QUERY RELEVANCE

FRAG finely extracts a set of relevant sequences $(\mathcal{R} = [x_1, x_2, \dots, x_m])$ from the initial retrieved document $(D_s = [d_1, d_2, \dots, d_N])$ based on their relevance to the query Q, using an extractor model \mathcal{E} . To enhance the extraction and ensure minimum information loss while filtering noise, FRAG identifies several key sequence snippets \mathcal{S} from Q, and extracts the relevant information using snippet-level query relevance. To further filter noise in \mathcal{R} , each sequence in \mathcal{R} is then evaluated for its relevance score to the query, and those sequences with scores above 0 are selected as the basic knowledge \mathcal{K} that is relevant to Q, and those above a given relevance threshold \mathcal{T}_G are selected as the basic knowledge for generation \mathcal{K}' . Furthermore, to reduce the negative effects of noise, FRAG conducts extraction within several steps, each focusing on extracting information from a single, shorter-context retrieval passage, which is proved to enhance extractor model's attention focusing on the relevant knowledge (see Appendix C.2 for proof). Additionally, to extract inference-based knowledge based on the corresponding prior information, we introduce Self-Recognition, a method for logical recognition. We provide an overview of FRAG extraction framework (Section 3.3), and furthermore, introduce FRAG-ip for better computational efficiency (Section 3.3).

3.1 HOW SNIPPET-LEVEL QUERY RELEVANCE BENEFITS THE NOISE FILTERING

Definition. Real-world user queries often include *lengthy background descriptions*. We define "key sequence snippets": certain sequence snippets of Q that serve as key indicators for extracting relevant knowledge; "non-key sequence snippets": the remaining irrelevant snippets. The input sequence is denoted as $X_Q = X_{Q^+} \oplus X_{Q^-}$, where X_{Q^+} and X_{Q^-} are the embedding matrices of key and non-key sequence snippets. X'_{Q^+} denotes the embedding matrix of the key sequence snippets decomposed from Q. $X_{\text{rel}, \subset}$ signifies the portion of extracted passage relevant to X_{Q^+} , while $X_{\text{red}, \subset}$ the irrelevant.

Lemma 3.1. The presence of non-key sequence snippets X_{Q^-} can diminish the LLM's attention on relevant knowledge during extraction. Namely, we have (proof is provided in Appendix C.3):

$$\frac{\mathfrak{A}_{rel}^+}{\mathfrak{A}_{red}^+} \ge \frac{\mathfrak{A}_{rel}^+ + \mathfrak{A}_{rel}^-}{\mathfrak{A}_{red}^+ + \mathfrak{A}_{red}^-}.$$
 (1)

Here, $\mathfrak{A}^+_{\mathrm{rel}}$, $\mathfrak{A}^-_{\mathrm{rel}}$ and $\mathfrak{A}^+_{\mathrm{red}}$, $\mathfrak{A}^-_{\mathrm{red}}$ signify the total attention distribution weights allocated to $X_{\mathrm{rel},\subset}$ and $X_{\mathrm{red},\subset}$, respectively, corresponding to X_{Q^+},X_{Q^-} . Thus, we have:

Theorem 3.2. By identifying key sequence snippets from Q, and explicitly appending the snippets to Q as query-augmented information during extraction, i.e., X'_{Q^+} —which indicates leveraging snippet-level query relevance—it mitigates the attention reduction caused by the non-key sequence snippets X_{Q^-} (Lemma 3.1), enhancing the extraction of relevant information. Namely, we have:

$$\frac{\mathfrak{A}_{rel}^{+} + \mathfrak{A}_{rel}^{-} + \mathfrak{A}_{rel}^{\prime+}}{\mathfrak{A}_{red}^{+} + \mathfrak{A}_{red}^{-} + \mathfrak{A}_{red}^{\prime+}} \ge \frac{\mathfrak{A}_{rel}^{+} + \mathfrak{A}_{rel}^{-}}{\mathfrak{A}_{red}^{+} + \mathfrak{A}_{red}^{-}}.$$
 (2)

Here, $\mathfrak{A}'^+_{\mathrm{rel}}$ and $\mathfrak{A}'^+_{\mathrm{rel}}$ denote the total attention distribution weights allocated to $X_{\mathrm{rel},\subset}$ and $X_{\mathrm{red},\subset}$, respectively, corresponding to X'_{Q^+} . Proof is provided in Appendix C.4.

Remark 3.3. As established in Theorem 3.2, FRAG identifies key sequence snippets from the query and explicitly appends them to the query as query-augmented information during extraction, thereby enhancing the extraction of relevant information from the retrievals using snippet-level query relevance and filtering noise with minimum information loss.

3.2 Self-Recognition: Enhancing Inference-Based Knowledge Extraction

While multi-step extraction increases the likelihood of retrieving relevant information, it poses a challenge for complex tasks like multi-hop reasoning, which require knowledge pieces with multi-hop relationships. In such cases, it is necessary for an inference-based sequence in one passage to be preceded by corresponding prior relevant sequences from other passages to correctly identify its relevance Li & Peng (2023). Yet, lacking access to these passages leads to the absence of multi-hop links and hinders correct identification of the inference-based sequences within the extracted passage.

To address this, we introduce *Self-Recognition*, which incorporates historically extracted knowledge into subsequent extraction as contextual references. This facilitates the extraction of inference-based knowledge exhibiting multi-hop contextual relationships within retrieved passages, formalized as:

$$\mathbb{1}_{k_{\text{In-B}}^i} = Q \wedge \mathbb{1}_{k_{\text{P-B}}^{(1,\dots l)}} \wedge d_i, \tag{3}$$

where $\mathbb{1}_k$ denotes the presence or generation of extracted knowledge; $k_{\text{In-B}}^i$ and $k_{\text{P-B}}^{(1,...l)}$ the inference-based knowledge within d_i and its l pieces of prior knowledge; " \wedge " means logical necessity.

3.3 OVERVIEW OF FRAG EXTRACTION FRAMEWORK

To extract the relevant sequences $\mathcal{R} = [x_1,\ldots,x_m]$ from the retrieval document $D_s = [d_1,\ldots,d_N]$, where d_i represents the i-th retrieval passage, and to construct the basic knowledge $\mathcal{K} = [k_1,\ldots,k_t]$, from which the most relevant sequences are selected to form the generation context \mathcal{K}' based on a relevance threshold for generation (\mathcal{T}_G), FRAG employs an opensourced extractor model \mathcal{E} and constructs an LLM-based Extraction Framework. The framework comprises six modules: Extractor \mathfrak{E} , Validator \mathfrak{V} , Prefixer \mathfrak{V} , Deduplicator \mathfrak{V} , Filter \mathfrak{F} , and Assessor \mathfrak{E} . Figure 2 illustrates the workflow of the FRAG extraction framework, while Table 1 outlines the function of each module.

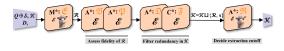


Figure 2: Workflow of FRAG extraction framework. $\mathfrak{E}, \mathfrak{V}, \mathfrak{P}, \mathfrak{D}, \mathfrak{F}, \mathfrak{R}$ represent *Extractor, Validator, Prefixer, Deduplicator, Filter, and Verifier,* respectively. M^*, C^*, A^* stand for main module, core module and auxiliary module, respectively. \mathcal{E} signifies the extractor model. \mathcal{R}, s denote the extracted relevant sequences and their corresponding relevance scores, respectively.

Table 1 outlines the function of each module. FRAG extraction framework leverages *Chain-of-Thought (CoT)* reasoning Wei et al. (2022) and *few-shot prompt engineering* Brown et al. (2020) to guide the LLM in reasoning for better extraction. Algorithm 1 illustrates the FRAG extraction.

Construction of FRAG extraction framework. The six modules are driven by \mathcal{E} , and operate sequentially to extract the relevant knowledge \mathcal{K} from D_s and filter noise. \mathfrak{E} , main module of FRAG identifies key sequence spinnets \mathcal{S} from

Return K

FRAG, identifies key sequence snippets S from Q, and extracts relevant sequences R from D_s using snippet-level query relevance:

$$x_i = (Q \oplus S) \otimes d_i = \begin{cases} d_i^{\text{rel}}, & \text{if } d_i^{\text{rel}} \text{ in } d_i, \\ \varnothing, & \text{otherwise;} \end{cases} \tag{4}$$

where " \otimes " signifies the semantic relevance matching operation and $d_i^{\rm rel}$ denotes the portion (if present) of di relevant to Q. It incorporates Self-Recognition, which integrates the historically extracted knowledge as references in subsequent extraction. $\mathfrak{D}, \mathfrak{D}$ verify the consistency of \mathcal{R} with D_s and \mathcal{K} , respectively: the former validates correctness, while the latter detects noise. Any sequence that is either incorrect or noisy will be discarded. \mathfrak{F} assesses the relevance score s between \mathcal{R} and Q, filtering noise in \mathcal{K} :

$$s_i = (Q \oplus S) \odot x_i = Q \odot x_i + S \odot x_i, \quad (5)$$

Table 1: **FRAG work modules. con, dup, ans** and **stop** represent *consistent, duplicate, QueryAnswered* and *stop* flag, respectively. \mathcal{K}_{prior} denotes the basic knowledge extracted in the last extraction round. (\mathcal{K}) means \mathcal{K} is an optional input.

Module	Input → Output	Function
Œ	$Q \rightarrow \mathcal{S}$ $Q \oplus \mathcal{S}, (\mathcal{K}), d_i \rightarrow x_i$	Identify S from Q . Extract x_i from d_i .
V	$x_i, d_i \rightarrow \mathbf{con}$	Discard x_i if x_i is inconsistent with d_i (con is false).
Ŗ	$x_i, d_i \rightarrow x_i$	Prepend the topics or subjects of d_i to x_i .
D	$x_i, \mathcal{K} \rightarrow \mathbf{dup}$	Discard x_i when x_i duplicates knowledge in \mathcal{K} (dup is $true$).
\mathfrak{F}	$x_i, Q, (\mathcal{K}) \rightarrow s_i$	Assess relevance s_i of $\{x_i, Q\}$.
E	$Q, \mathcal{K} \rightarrow \mathbf{ans}$ $\mathcal{K}, \mathcal{K}_{prior} \rightarrow \mathbf{stop}$	Stop extraction if Q is answerable or no more x_i is extracted (ans/stop is $true$).

where " \odot " denotes the semantic relevance computation between sequences. \mathcal{K} also acts as optional references, enabling accurate scoring of inference-based knowledge. s_i is evaluated on a scale from 0 to 1, and extracted sequences with $s_i > 0$ are included in basic knowledge \mathcal{K} :

$$\mathcal{K} = \mathcal{K} \cup \begin{cases} \{(x_i, s_i)\}, & \text{if } s_i > 0, \\ \varnothing, & \text{otherwise;} \end{cases}$$
 (6)

and the extracted sequences are sorted by their relevance in positive order:

$$\mathcal{K} = [(x_u, s_u), \dots, (x_v, s_v)], \quad s_u \ge s_{u+1} \ge \dots \ge s_v > 0.$$
 (7)

 \mathfrak{S} decides when the extraction process should be terminated, thereby minimizing the extraction time. An extraction round is defined as a process in which all passages from the initial retrieval document are processed once by FRAG extraction framework to extract relevant knowledge. The workflow iterates until no further relevant knowledge can be extracted or the maximum extraction rounds, $L_{\text{Max-Rounds}}$, is reached. More details of FRAG implementation are provided in Appendix B.

Algorithm 1 FRAG Extraction

```
Require: Query Q; Q's key sequence snippets S; Retrieved documents D_s = [d_1, \ldots, d_N]; FRAG extraction framework F; Basic knowledge K; Extracted relevant sequence x_i; Relevance score s_i of \{Q, x_i\}; Max extraction rounds L_{\text{Max-Rounds}}; Check K \leq K_{\text{prior}}: verify if K does not exceed K_{\text{prior}}. Initialize ans = false; stop = false; r = 1;
```

```
\triangleright Decompose S from Q (Section 3.1)
while {ans is false and stop is false and r \leq L_{\text{Max-Rounds}}} do
      \mathcal{K}_{\text{prior}} \leftarrow \mathcal{K}
      for i=1,\ldots,n do
            if K is \varnothing then
                  (x_i, s_i) \stackrel{\mathcal{F}}{\leftarrow} \{Q \oplus \mathcal{S}, d_i\}
                                                                            \triangleright Extract x_i given Q \oplus S & Assess s_i (Section 3.1)
                   (x_i, s_i) \stackrel{\mathcal{F}}{\leftarrow} \{Q \oplus \mathcal{S}, d_i, \mathcal{K}\}
                                                                                                                ▶ Self-Recognition (Section 3.2)
            \mathcal{K} \leftarrow \mathcal{K} \cup \{(x_i, s_i)\}, \text{ with } s_i > 0
      \mathcal{S} \xleftarrow{\mathcal{F}} Q \oplus \mathcal{K}; r \leftarrow r+1
                                                                                                                          \triangleright Reidentify S if necessary
      Assess {ans: true \text{ or } false} \stackrel{\mathcal{F}}{\leftarrow} \{Q, \mathcal{K}\}
                                                                                                    \triangleright Assess if Q is answerable based on \mathcal{K}
      if \mathcal{K} \leq \mathcal{K}_{prior} then
                                                                                                       No more basic knowledge extracted
            stop \leftarrow true
```

FRAG limitation in computational efficiency. Although FRAG significantly improves RALM performance by effectively extracting relevant information and filtering noise from the initial documents, it requires a large number of LLM calls for extraction, leading to high computational cost. The extraction process can take 15.07s per query on an A100 GPU on HotPotQA (Table A.6).

To reduce computational cost, we propose FRAG-ip, a wrapper framework that performs a dual-stage fine-tuning process on the Qwen2-7B-Instruct model to accelerate the extraction. In the first stage, the extractor model \mathcal{E} is fine-tuned to (1) identify key sequence snippets \mathcal{S} from the query Q, and (2) extract relevant sequences \mathcal{R} and filter noise from tretrieved passages in a fine-grained manner using snippet-level query relevance, while maintaining consistency with the source passages (to support \mathcal{E}). The post-training of the extraction module further incorporates the Self-Recognition method, leveraging historically extracted knowledge as contextual references to enhance the extraction of inference-based knowledge. In the second stage, \mathcal{E} is fine-tuned to (1) assess the relevance score of a given sequence with respect to \mathcal{S} and \mathcal{Q} (to support \mathcal{S}), and (2) determine whether a query is answerable given a set of relevant or irrelevant knowledge (to support \mathfrak{A}).

Furthermore, with improved extraction performance and more relevant, consistent extracted sequences, $\mathfrak{V}, \mathfrak{D}$ are transformed to be purely driven by an program executor \mathcal{P} . These optimizations significantly reduce the extraction runtime by approximately by $10 \times$ (Table A.6). More details, including the fine-tuning process, results of the experiments utilizing FRAG-ip and a comparison of computational cost between FRAG-ip and FRAG are provided in Appendix B.4.3.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

We evaluate FRAG on three simple single-hop datasets — PopQA (Mallen et al., 2023), PubHealth (Akhtar et al., 2022), and ARC-Challenge (Clark et al., 2018) — and on four complex datasets, including three multi-hop datasets (HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022)) and one long-context task (LongBench-v2 (Bai et al., 2024)). For single-hop and long-context datasets, we use accuracy (acc) as the evaluation metric; For multi-hop datasets, we evaluate using Exact Match (em) (Rajpurkar et al., 2016). However, since em may yield inaccurate assessments by simply matching answer text with the ground truth (Wang et al., 2023a), we use LLM-em (Yu et al., 2024a), which leverages LLMs to assess the correctness of RALM-generated answers based on ground-truth, thereby providing more reliable evaluation. Experimental details, including the conduction of LLM-em evaluation, are present in Appendix D.

4.2 Baselines

Baselines without retrievals. We evaluate the most advanced models, DeepSeek-v3 (Liu et al., 2024a) and GPT-4o (Achiam et al., 2023), as well as advanced open-source models, including Llama3-8B-Instruct (Meta AI, 2024), Qwen2-7B-Instruct (Yang et al., 2024), and the advanced fine-tuned question answering models for RAG, ChatQA-1.5-8B/70B (Liu et al., 2024b). We also compare FRAG to other advanced models simulated by GPT-4, Alpaca7B/13B (Dubois et al., 2023).

Baselines with retrievals. We test the baselines using the initial retrieval passages from our method for generation. We evaluate Llama3-8B-Instruct, Qwen2-7B-Instruct and ChatQA-1.5-8B/70B using naive RAG across all datasets. We also include the following representative RAG baselines: RankGPT (Sun et al., 2023), which relies on GPT-3.5 to rerank the retrievals and select the top-k (k=5 for test) passages for generation; SelfRAG (Asai et al., 2024), which employs a self-reflection mechanism to critique RALM responses and select the best one. Results of Alpaca 7B/13B were reported in SelfRAG method; RECOMP (Xu et al., 2024a), which relies on a compressor to condense the retrieval passages into short summaries; Rewrite-Retrieve-Read (Ma et al., 2023), which inserts a query rewriting step to enhance retrieving; RQ-RAG (Chan et al., 2024), which fine-tunes a model to rewrite, decompose, and disambiguate queries; ActiveRAG (Xu et al., 2024c), which uses four agent-driven strategies to integrate external evidence with LLM memory; SAIL (Luo et al., 2023), a method that instruction-tuning a model using the top retrieved documents; RA-DIT (Lin et al., 2024), which augments RALMs with a tunable retrieval model; Results of LLAMA-65B REPLUG were reported in the RA-DIT method; LongRAG (Zhao et al., 2024), which utilizes a hybrid retriever

and a long-context retrieval chunks refinement method; HippoRAG (Jimenez Gutierrez et al., 2024), which mimics hippocampal memory via an LLM-built knowledge graph and Personalized PageRank for efficient single-step multi-hop retrieval; Search-o1 (Li et al., 2025), which enhances RAG via a reasoning model and agentic search with document refinement for multi-step knowledge integration.

4.3 IMPLEMENTATION

Extraction settings. We separately employ the advanced instruction fine-tuned models, Llama3-8B-Instruct and Qwen2-7B-Instruct, as extractor models to extract relevant knowledge \mathcal{K} from the initial retrieval documents. $L_{\text{Max-Rounds}}$ is set to a default of 1 for simple tasks and 3 for complex ones.

Generation and evaluation settings. We separately use Llama3-8B-Instruct and Qwen2-7B-Instruct as the generator models. Additionally, we utilize ChatQA-1.5-8B/70B as generators using the basic knowledge extracted by Qwen2-7B-Instruct to test the robustness of FRAG when transferring extracted knowledge from one extractor model to another generator. We test four set values as \mathcal{T}_G during the experiments: 0.2/0.45/0.7/0.95 to analyze the impact of different \mathcal{T}_G values on RALMs' performance. We use GPT-40 as the evaluator to assess LLM-em scores across the multi-hop datasets.

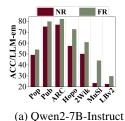
Table 2: **Overall experiment results. Bold** numbers indicate the best result across all models, <u>underlined</u> numbers denote the second-best, and **gray-colored bold** numbers signify the best result among retrieval-based models. "-" indicates values not reported in the original papers or inapplicable. "*" signifies that a detailed analysis is provided in Appendix E.1 for the results. "†" suggests potential data leakage to achieve such performance without retrieving any documents (Zhou et al., 2023).

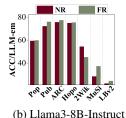
		Single-h	op			Multi-l	10р			Long-Contex
	Pop Pub ARC		Hotpo	tQA	2Wil	ki	MuSi(Que	LongBench-v2	
LMs	(acc)	(acc)	(acc)	(llm-em)	(em)	(llm-em)	(em)	(llm-em)	(em)	(acc)
					Baseline	es without r	etrieval			
DeepSeek-v3	31.17	71.02	95.74†	-	31.9	-	43.9	-	9.2	33.6
GPT-40	43.6	56.43	95.74†	-	36.7	-	47.4	-	14.5	32.41
Alpaca-7B	23.6	49.8	45	-	-	-	-	-	-	-
Alpaca-13B	24.4	55.5	54.9	-	-	-	-	-	-	-
ChatQA-1.5-8B	27.66	62.21	58.52	-	15.2	_	36	-	0.7	24.06
Llama3-8B-Instruct	26.16	67.88	74.11	-	0.8	_	0	-	0	1.79
Qwen2-7B-Instruct	27.38	62.92	78.36	-	0	-	0	-	0	20.87
ChatQA-1.5-70B	33.17	66.97	86.46†	-	19.5	-	35.3	-	2.3	<u>32.41</u>
					Baseli	nes with ret	rieval			
Alpaca-7B	46.7	40.2	48	-	-	-	-	-	-	-
Alpaca-13B	46.1	51.1	57.6	-	-	-	-	-	-	_
SAIL-7B	-	69.2	48.4	-	-	-	-	-	-	_
SelfRAG-7B	54.9	72.4	67.3	-	12.9	-	16.8	-	1.2	22.86
SelfRAG-13B	55.8	74.5	73.1	-	13.2	-	6.2	-	1.5	2.58
RECOMP-20B	-	-	-	-	30.4	_	-	-	-	_
RA-DIT-65B	-	-	60.5	-	40.7	_	-	-	-	_
REPLUG-LLAMA-65B	-	-	-	-	41.1	_	-	-	-	_
RankGPT-7B	55.68	78.52	82.28	-	42.9	_	35.6	-	15.9	24.45
Rewrite-Retrieve-Read	-		-	-	34.38	_	-	-	-	_
RQ-RAG-7B	57.1	-	68.3	-	0*	_	0*	-	0*	27.04
LongRAG-6B	-	-	-	-	40.5	_	37.5	-	17.5	_
ActiveRAG-8B	46.46	32.22	46.34	-	23.6	_	29	-	7.5	17.89
HippoRAG	-	-	-	-	41.8	_	46.6	-	19.2	_
IRCoT+HippoRAG	-	-	-	-	45.7	_	47.7	-	21.9	_
Search-o1-32B	-	-	-	-	45.2	_	58.0	-	16.6	_
ChatQA-1.5-8B	53.75	67.17	37.99	54.6	9.9	43.7	33.8	25.7	1.3	13.12
Llama3-8B-Instruct	59.39	72.14	75.64	74.9	38.8	54.4	41.8	28.4	14.9	22.27
Owen2-7B-Instruct	51.82	75.08	76.83	57.4	34.9	50.2	42.2	23.5	9.2	22.66
ChatQA-1.5-70B	58.97	72.14	74.11	72.3	26.8	54.2	21.9	38.7	8	24.06
						ours				
FRAG-ChatQA-1.5-8B	54.11	73.66	46.42	67.4	27.9	52.2	44.2	35.3	12.4	21.47
FRAG-Llama3-8B-Instruct	59.83	75.99	77.34	<u>75.4</u>	41.7	45.2*	32.1*	37.3	<u>22.9</u>	24.45
FRAG-Qwen2-7B-Instruct	57.97	79.84	82.03	72.7	47.7	60.9	<u>51.5</u>	44.1	25.6	29.82
FRAG-ChatQA-1.5-70B	59.69	76.49	82.79	77.8	38.4	<u>58.5</u>	28.4	<u>42.4</u>	19.3	29.03

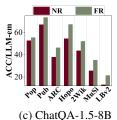
4.4 MAIN RESULTS

FRAG performance. It can be demonstrated that the same model consistently performs better when using FRAG compared to naive RAG across most tested datasets (except for Llama3 on

2WikiMultihopQA, for which analysis is provided in Appendix E.1), as shown in Table 2 and Figure 3. In most cases, the naive RAG method generally outperforms non-retrieval-based method. This confirms Theorem 2.2 that introducing excessive noise can indeed reduce the accuracy of RALM. Moreover, by effectively extracting relevant knowledge and filtering noise, the negative effects of noise can be mitigated, thereby improving the performance of RALM.







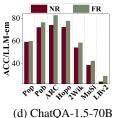


Figure 3: Performance comparison of four tested models using FRAG (FR) versus naive RAG (NR).

Additionally, FRAG achieves the best overall performance among retrieval-based baselines, outperforming the reranking method RankGPT, query reformulation baselines such as RQ-RAG, and advanced methods including SelfRAG and ActiveRAG across all tasks. On most multi-hop datasets, FRAG also surpasses strong reasoning-model-based method, Search-o1, and advanced baselines such as HippoRAG, LongRAG, RADIT, REPLUG, and RECOMP. Furthermore, FRAG outperforms state-of-the-art fine-tuned QA models ChatQA-1.5-8B/70B when they are used with naive RAG. Notably, FRAG-Qwen2-

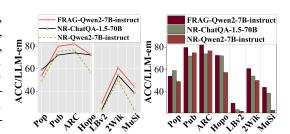


Figure 4: Performance comparison of FRAG-Qwen2-7B-Instruct versus NR-ChatQA-1.5-70B.

7B-Instruct achieves significant improvements across all datasets, with an average accuracy gain of 4.94% on single-hop datasets and 13.44% on multi-hop and long-context datasets, compared to using naive RAG. These results demonstrate the advantages of FRAG. Moreover, FRAG-Qwen2-7B-Instruct achieves performance that nearly surpasses ChatQA-1.5-70B using naive RAG, despite a much smaller model size, as shown in Figure 4. This suggests that small-sized models can still demonstrate excellent RAG performance when generating with contexts that have been filtered to remove excessive noise, without the need to simply increase the model size to improve performance. The fact that ChatQA-1.5-8/70B, when augmented with retrievals extracted by Qwen2, consistently perform better than with naive RAG further demonstrates FRAG effectiveness and robustness.

FRAG efficiency in extracting relevant knowledge and filtering noise. The HotpotQA dataset contains annotations for golden passages within the retrieved documents (Yang et al., 2018). In the experiments using the FRAG method to extract relevant knowledge, we evaluate the practical effectiveness of our method using the following four metrics: (a) *rete*: retention ratio of the golden passages after extraction, i.e., the number of golden passages extracted divided by the total number of golden passages (relevant knowledge extracted from the same golden passage is regarded as a single extracted golden passage). (b)

Table 3: FRAG efficiency in retaining relevant knowledge and filtering noise. "↑"/"↓" indicate an increase/decrease relative to the corresponding proportion in the initially retrieved documents.

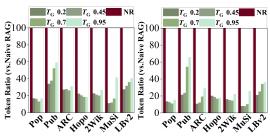
LMs	HotpotQA								
23.725	(rete)	(p_{ori})	(p_{ext})	(η)					
FRAG-Qwen2-7B -Instruct	90.15	16.10	42.21 (26.11†)	82.87↓					
FRAG-Llama3-8B -Instruct	87.65	16.10	39.13 (23.03↑)	83.95↓					

 $p_{\rm ori}$: proportion of tokens from golden passages within the initial retrieved documents. (c) $p_{\rm ext}$: proportion of tokens from golden passages in the extracted documents. (d) η : filtering rate of tokens from noise in the extracted documents compared to the initial retrieved documents.

Table 3 demonstrates that FRAG effectively extracts and retains relevant information from initial retrieval documents, achieving a retention ratio of 90.15% with Qwen2 and 87.65% with Llama3, while significantly filtering noise by 82.87% and 83.95%, respectively. This notably increases the

proportion of relevant information in the generation context by at least 26.11% on Qwen2 and 23.03% on Llama3 (notably, as noise exists even within golden passages, the actual enhancements surpass these values). The above analysis indicates that FRAG significantly enhances RALM performance due to two key factors: first, it effectively identifies and extracts relevant information from the initial retrieval documents; second, it efficiently filters noise, thus substantially increasing the proportion of relevant information in the generation context. This process effectively mitigates attention drift caused by noise during RAG generation, thereby enhancing RALMs' capability to generate accurate answers based on the relevant knowledge within the generation context.

FRAG advantages in improving spatial efficiency for RAG generation. FRAG achieves an average reduction in input tokens of up to 73.23% for Qwen2 across all datasets during the RAG generation phase compared to naive RAG, and 78.54% for Llama3. Under optimal conditions, this reduction reaches 78.33% for Qwen2 and 85% for Llama3, as illustrated in Figure 5. The reduction in input tokens within the generation context across all datasets under varying \mathcal{T}_G values highlights FRAG's significant enhancement of the spatial efficiency in the RAG approach. Furthermore, it demonstrates that *RALM performance can be significantly improved by effectively filtering noise from the gen-*



(a) Qwen2-7B-Instruct

(b) Llama3-8B-Instruct

Figure 5: Comparison of input token counts between FRAG and naive RAG (NR).

eration context, as supported by the combined results from Table 2, Table 3, and Figure 5.

4.5 ABLATION STUDIES

Contribution of different components. To verify whether conducting extraction using snippet-level query relevance increases the probability of extracting relevant knowledge from the initial retrieved documents, and assess the role of Self-Recognition in identifying logical relationships between different relevant sequences in the retrieved documents and recognizing inference-based knowledge based on historically extracted knowledge, we conduct ablation experiments using Qwen2-7B-Instruct as both the extractor and generator model on HotpotQA.

The results in Table 4 indicate that the extraction method using snippet-level query relevance indeed improves the retention rate of golden passages and enhances model performance. For

Table 4: Results of the ablation studies. **Bold** numbers indicate the best performance, and "\" indicates a decrease compared to the best.

LMs		HotpotQA							
22.125	(llm-em)	(rete)	(p_{ext})	(η)					
FRAG-Qwen2-7B -Instruct	72.7	90.15	42.21	82.87					
w/o adding ${\cal S}$	66.5 (6.2↓)	75.10	46.69	92.01					
w/o Self-Recognition	60.0 (12.7↓)	78.10	38.73	88.81					
w/o adding \mathcal{S} and Self-Recognition	58.6 (14.1↓)	62.15	46.03	95.18					

multi-hop datasets, the introduction of Self-Recognition significantly improves both the retention rate of golden passages and proportion of relevant information in the generation context, thereby yielding a greater impact on accuracy. These findings validate the correctness and effectiveness of FRAG.

5 CONCLUSION

This work presents FRAG, an innovative LLM-based extraction framework that filters noise in retrieved documents, increases the proportion of relevant information in the RAG generation context, and enhances RALM accuracy. FRAG utilizes six modules to extract relevant knowledge using snippet-level query relevance and incorporates Self-Recognition to leverage historically extracted knowledge as references and enhance inference-based knowledge extraction. Extensive experiments demonstrate that FRAG markedly boosts RAG performance. Additionally, we propose an improved wrapper framework FRAG-ip to reduce the computational cost.

REFERENCES

- Joshua Achiam, Sam Adler, Suyash Agarwal, Ethan Dyer, Oran Gafni, Jena Huang, Scott Reed, Mark Chen, and Jared Kaplan. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, OpenAI, 2023. https://arxiv.org/abs/2303.08774.
- Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. PubHealthTab: A public health table-based dataset for evidence-based fact checking. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics:* NAACL 2022, pp. 1–16, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.1. URL https://aclanthology.org/2022.findings-naacl.1.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.
- Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
- Cheuk Man Chan, Chen Xu, Rui Yuan, et al. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*, 2024.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018. URL https://api.semanticscholar.org/CorpusID: 3922816.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. *CoRR*, abs/2401.14887, 2024. URL https://doi.org/10.48550/arXiv.2401.14887.
- Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=4hturzLcKX.
- Daryl Edge, Hieu Trinh, Nhan Cheng, Sarah Doe, and John Smith. From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130, 2024. https://arxiv.org/abs/2404.16130.
- Yin Gao, Yuyi Xiong, Xinyue Gao, Qing Yang, and Bo Chen. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2023. https://arxiv.org/abs/2312.10997.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu,

Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3929–3938. PMLR, 2020.

- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multihop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- Gabriel Izacard, Mathilde Caron, Luyu Hosseini, Sebastian Riedel, and Moin Nabi. Unsupervised dense information retrieval with contrastive learning. arXiv preprint arXiv:2112.09118, 2021. https://arxiv.org/abs/2112.09118.
- Gabriel Izacard, Patrick Lewis, Maria Lomeli, et al. Few-shot learning with retrieval augmented language models, 2022. arXiv preprint arXiv:2208.03299.
- Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74.
- Bernardo Jimenez Gutierrez, Yichong Shu, Yankai Gu, et al. Hipporag: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59532–59569, 2024.
- Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 957–966, 2015. URL https://proceedings.mlr.press/v37/kusnerb15.html.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Barlas Oguz, Qiuyuan Huang, Wen tau Yih, Pontus Stenetorp, Sonal Gupta, and Sebastian Riedel. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9459–9474, 2020.
- Xiaoyang Li, Guang Dong, Jiawei Jin, et al. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.
- Zekai Li and Wei Peng. Self-adaptive reasoning on sub-questions for multi-hop question answering. In *ICASSP 2023 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10097206.

- Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. RA-DIT: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=220Tbutug9.
 - Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.
 - Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. ChatQA: Surpassing GPT-4 on conversational QA and RAG. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=bkUvKPKafQ.
 - Hongyin Luo, Tianhua Zhang, Yung-Sung Chuang, Yuan Gong, Yoon Kim, Xixin Wu, Helen M. Meng, and James R. Glass. Search augmented instruction learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=noIvPGG8P1.
 - Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5303–5315, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.322. URL https://aclanthology.org/2023.emnlp-main.322/.
 - Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546.
 - Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.
 - Vishwajeet Mavi, Aman Jangra, and Adam Jatowt. A survey on multi-hop question answering and generation. https://arxiv.org/abs/2204.09140, 2022. arXiv preprint arXiv:2204.09140.
 - Meta AI. Introducing llama 3: A state-of-the-art language model. https://ai.meta.com/blog/meta-llama-3/, 2024.
 - Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.
 - Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.
 - Ori Ram, Yoav Levine, Idan Dalmedigos, Aviv Oren, Amnon Yogev, Gal Chechik, and Jonathan Berant. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.463. URL https://aclanthology.org/2024.naacl-long.463.

- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pp. 194–206. Springer, 2019.
- Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nnVO1PvbTv.
- Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14918–14937, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.923. URL https://aclanthology.org/2023.emnlp-main.923/.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.
- Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. Evaluating open-qa evaluation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 77013-77042. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f323d594aa5d2c68154433a131c07959-Paper-Datasets_and_Benchmarks.pdf.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.
- Ziyu Wang, Jun Araki, Zihan Jiang, Xin Huang, and Yoshua Bengio. Learning to filter context for retrieval-augmented generation. arXiv preprint arXiv:2311.08377, 2023b. https://arxiv.org/abs/2311.08377.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837, 2022.
- Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=mlJLVigNHp.
- Zenglin Xu, Saurabh Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. arXiv preprint arXiv:2401.11817, 2024b. https://arxiv.org/abs/2401.11817.
- Ziyang Xu, Zihan Liu, Yifan Yan, et al. Activerag: Autonomously knowledge assimilation and accommodation through retrieval-augmented agents. *arXiv preprint arXiv:2402.13547*, 2024c.
- Ailin Yang, Bin Yang, Bo Hui, Xiao Li, Qing Zhang, and Zhen Wang. Qwen2 technical report. arXiv preprint arXiv:2407.10671, arXiv, 2024. https://arxiv.org/abs/2407.10671.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

- Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZS4m74kZpH.
- S. Yu, M. Cheng, J. Yang, et al. A knowledge-centric benchmarking framework and empirical study for retrieval-augmented generation. *arXiv preprint arXiv:2409.13694*, 2024a.
- Tao Yu, Ananya Xu, and Rama Akkiraju. In defense of rag in the era of long-context language models. arXiv preprint arXiv:2409.01666, 2024b. https://arxiv.org/abs/2409.01666.
- Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024c. URL https://openreview.net/forum?id=S1fc92uemC.
- Xuanwang Zhang, Yun-Ze Song, Yidong Wang, Shuyun Tang, Xinfeng Li, Zhengran Zeng, Zhen Wu, Wei Ye, Wenyuan Xu, Yue Zhang, Xinyu Dai, Shikun Zhang, and Qingsong Wen. RAGLAB: A modular and research-oriented unified framework for retrieval-augmented generation. In Delia Irazu Hernandez Farias, Tom Hope, and Manling Li (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 408–418, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-demo.43.
- Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.
- Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. LongRAG: A dual-perspective retrieval-augmented generation paradigm for long-context question answering. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22600–22632, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024. emnlp-main.1259. URL https://aclanthology.org/2024.emnlp-main.1259/.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.
- Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater, 2023.
- Kexin Zhu, Xiaoqing Feng, Xiaoyang Du, et al. An information bottleneck perspective for effective noise filtering on retrieval-augmented generation. *arXiv preprint arXiv:2406.01549*, 2024.

CONTENTS Related Work FRAG Details B.2.2B.2.3B.2.4 B.2.5 B.2.6 B.4 FRAG-ip: A Wrapper Fine-tuned Framework for Optimizing Computational Efficiency 20 B.4.1 B.4.2Computational Efficiency of FRAG and Results with FRAG-ip B.4.3 **Proof of Main Theoretical Analysis** D Additional Experiment and Details D.3 Accuracy Comparison Between GPT-40 as the Evaluator and Human Evaluation . . E Additional Details on the Analysis of the Main Results E.1.1 Performance Analysis with Llama3-8B-Instruct as Extractor and Generator E.1.2 Analysis of RQ-RAG Results in Multi-hop Datasets Analysis of Inaccurate Scoring Phenomena in Module Filter §

F	Disc	cussion	of Additional Limitations and Ethical Impacts	32
G	Dec	laratior	n of LLM Usage	32
Н	Lice	ense		33
I	Pro	mpts ar	nd Samples	33
	I.1	Promp	ots	33
		I.1.1	Extraction Prompt Examples	33
		I.1.2	Fine-tuning Prompt Examples	38
		I.1.3	Prompt Used for LLM-EM Evaluation with GPT-40	42
	I.2	Sampl	es	42
		I.2.1	Samples of Extracted Basic Knowledge	42
		I.2.2	Samples of Inaccurate Relevance Scoring	52
		I.2.3	Samples of Discrepancies Between GPT-40 and Human Assessments	53

A RELATED WORK

Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG) enhances the generation process by retrieving relevant knowledge from external knowledge bases as context for answering queries (Lewis et al., 2020; Guu et al., 2020), demonstrating outstanding performance in knowledge-intensive tasks (Ram et al., 2023). With advancements in retrievers and large models used as generators, a diverse range of RAG techniques have emerged to improve generation quality and accuracy. SelfRAG fine-tunes the generator model to adaptively retrieve the necessary knowledge and leverages a self-reflection mechanism to generate and critique responses, ultimately yielding the best answers (Asai et al., 2024). ChatQA is a model series trained with high-quality question-answer pairs specifically for RAG and QA tasks, with its most advanced version outperforming GPT-4 on RAG and QA tasks (Liu et al., 2024b). Meanwhile, RAG methods continue to expand into more complex application scenarios. ToG proposes an integrating paradigm "LLM \otimes KG", using LLMs over retrieved knowledge graphs to solve multi-hop problems (Sun et al., 2024). Graph RAG constructs global knowledge graph communities from unstructured text knowledge bases to address tasks like "query-focused summarization" (QFS) (Edge et al., 2024).

The Negative Impact of Retrieved Noise on RAG Tasks. A common belief is that increasing the retrieval window to introduce more relevant knowledge enhances RAG performance, as generative models effectively aggregate and synthesize evidence from multiple passages (Izacard & Grave, 2021). However, empirical engineering practices have shown that continually expanding the retrieval window does not guarantee sustained performance improvements for RAG; instead, it may degrade performance as the context length increases (Yu et al., 2024b;c). This issue arises because retrievers typically use text or vector similarity to retrieve contexts related to the query (Gao et al., 2023), and these retrieved contexts may not always be relevant to the query (Cuconasu et al., 2024), leading to noise. As the retrieval window expands, a substantial amount of noisy information is included, which in turn dilutes the model's focus on truly relevant information (Yu et al., 2024b;c). This results in an "attention drift" towards irrelevant content, thereby undermining the performance of RALM.

Concurrent Research on Reducing the Negative Impact of Retrieved Noise. Numerous research efforts have been implemented to mitigate the detrimental effects of noise. SelfRAG mitigates the negative impact of noise by generating responses from a single, shorter retrieved document, either once or multiple times, rather than relying on multiple retrieved passages simultaneously for a single generation (Asai et al., 2024). However, this approach severs logical connections between different retrieved passages, resulting in suboptimal performance in complex RAG tasks (Zhang et al., 2024). RECOMP reduces the negative effects of noise by incorporating additional compressors to summarize the retrieval documents (Xu et al., 2024a). However, this method lacks effective practical principles for identifying relevant information in the retrieved documents, leading to the loss of relevant knowledge during the compression process, which hinders it from achieving optimal performance (Xu et al., 2024a). RA-Dit employs a dual-stage fine-tuning approach involving both the retriever and the generator to acquire more relevant knowledge and improve the RALM's utilization of relevant information during generation, thereby achieving higher-quality outputs (Lin et al., 2024). Ori Yoran et al. fine-tuned models based on a mix of relevant and irrelevant contexts to enhance robustness against noisy content, thereby reducing the model's focus on irrelevant information during generation (Yoran et al., 2024). FILCO utilizes fine-tuned context filtering models to identify useful contexts based on lexical and information-theoretic approaches to effectively filter out noise (Wang et al., 2023b). RankRAG reduces the input of irrelevant content during generation by adding a small fraction of ranking data into the training blend, enabling the model to obtain the top-k most relevant passages (Yu et al., 2024c). However, methods relying on fine-tuning often have performance limitations that are dependent on the quality and comprehensiveness of the datasets used for fine-tuning or training, which can result in reduced generalization and versatility of the model.

Fine-tuning (Sun et al., 2019) is a widely adopted technique in adapting pre-trained models to downstream tasks, where model parameters are updated using task-specific labeled data. A well fine-tuned models often exhibit improved performance or computational efficiency on specific tasks (Naveed et al., 2023). Recent studies have proposed various fine-tuning strategies, including full-parameter tuning, adapter-based methods, and low-rank adaptation (e.g., LoRA (Hu et al., 2022)), to balance performance and efficiency.

B FRAG DETAILS

B.1 Extraction in Shorter Context Matters

The typical retrievers employed in RAG generally split the retrieval knowledge base into fixed-length chunks (usually 256/512 tokens) (Gao et al., 2023) prior to retrieval. Therefore, the retriever returns N logically independent and semantically complete paragraphs, determined by the retrieval window size. Furthermore, existing retrievers typically rely on text similarity or vector similarity for knowledge retrieval (Gao et al., 2023), which results in the retrieved passages not always being relevant to the query. Specifically, relevant information is concentrated in t paragraphs (where $t \leq N$). For these t retrieved passages, derivations show that the average proportion of relevant information in a single retrieval passage is higher than in a long-context document composed of N passages (Derivation is presented in Appendix C.2).

Established on Theorem 2.1, this suggests that using multiple extraction steps, each extracting knowledge from a single, shorter-context passage, facilitates the extractor model focusing on relevant information, thereby improving the extraction of such information. However, for other N-t noisy passages that contain no relevant information, the extractor model might mistakenly extract irrelevant knowledge. To mitigate this, we implemented two measures: (i) Since noise has inherently low relevance to the query, we require the extractor model to first assess whether the passage to be extracted is relevant to the query before extraction, forming a complete reasoning chain to guide the model's correct extraction behavior; (ii) We utilize the module Filter \mathfrak{F} (Section 3.3, overview of FRAG extraction framework) to evaluate the relevance score of \mathcal{R} to the query, where we standardize the scoring process to exclude sequences with low or no relevance.

B.2 DETAILS OF FRAG EXTRACTION FRAMEWORK

B.2.1 MODULE: EXTRACTOR &

€ identifies and extracts key sequence snippets (S) from the query Q, following the formulation:

$$S = Q \cap A^{-1},\tag{A.8}$$

where A^{-1} denotes the essential information required to derive the query answer. Key sequence snippets S refer to the certain sequence snippets of a query which serves as key indicators or fundamental information for retrieving the query answer. In contrast, non-key snippets typically consist of lengthy, noisy background or instructional content that may distract the model during information extraction or answer generation (Section 3.1). For example, consider the following instance from LongBench-v2:

Question: You are given a grammar book of Kalamang language, now translate the following Kalamang sentence into English: Faisal emun me mindi don bolonet me ma he kademor.

The essential snippet in this case is "Faisal emun me mindi don bolonet me ma he kademor", as each word in this sentence needs to be interpreted to produce the correct English translation. FRAG identifies S for the purpose of enhancing the extractor model E's attention on these key elements.

Incorporating Self-Recognition in \mathfrak{E} . In Complex tasks such as the multi-hop, the query requires knowledge pieces with multi-hop relationships. To better recognize the multi-hop contextual links during extraction, FRAG incorporates the Self-Recognition method. This method (1) updates the key snippets \mathcal{S} and (2) extracts inference-based knowledge by leveraging historically extracted knowledge as contextual references, formally defined as:

$$S' = (Q \oplus \mathcal{K}) \cap A^{-1},\tag{A.9}$$

$$k_{\text{in-B}}^i = (Q \oplus \mathcal{S} \oplus k_{\text{P-B}}^{(1,\dots,l)}) \otimes d_i, \tag{A.10}$$

where S' denotes the updated key snippets, reidentified from Q based on historically extracted knowledge. An illustrative example is shown in Figure 1. \mathfrak{E} then performs snippet-level extraction using CoT and few-shot prompt engineering. The prompt instances is provided in Appendix I.1.

B.2.2 MODULE: VALIDATOR 27

Since hallucinations (Xu et al., 2024b) in LLMs can negatively impacts the extraction, it is crucial to validate whether the extracted sequences \mathcal{R} remain consistent with the original retrievals to avoid factual errors. To this end, \mathfrak{V} adopts a hybrid validation strategy that combines: (1) *string-level match* using a program executor \mathcal{P} ; (2) *semantic-level match* using \mathcal{E} . Formally, this is defined as:

$$x_i = \begin{cases} x_i & \text{if } x_i \in {}_s d_i; \\ \text{None, otherwise.} \end{cases}$$
 (A.11)

where \in_s represents string-level or semantic-level consistent.

B.2.3 MODULE: PREFIXER 🎾

 \mathfrak{P} is introduced to ensure the *semantic completeness* of \mathcal{R} by \mathcal{O} , as determined by \mathcal{E} . This design is motivated by the observation that \mathcal{R} may omit critical contextual information present in the original retrievals. An example illustrative from PopQA is provided below:

Rawly extracted x_i : He was an English Conservative Party politician.

$$\Downarrow$$
 (Prefixer)

Prefixed x_i : Henry Feilden (Conservative politician): He was an English Conservative politician.

B.2.4 MODULE: DEDUPLICATOR 2

Similar to \mathfrak{V} , \mathfrak{D} verifies whether each sequence in \mathcal{R} is string- or semantically consistent with \mathcal{K} , which contains the historically extracted relevant knowledge. However, \mathfrak{D} discards sequences in \mathcal{R} that are already consistent with \mathcal{K} , thus further filtering noise in \mathcal{K} . Formally, this is defined as:

$$x_i = \begin{cases} \text{None,} & \text{if } x_i \in_s \mathcal{K}; \\ x_i, & \text{otherwise.} \end{cases}$$
 (A.12)

B.2.5 MODULE: FILTER §

 \mathfrak{F} assesses a relevance score of each sequence in \mathcal{R} to Q, and discards those sequences with low relevance (default threshold: 0) to further reduce noise in \mathcal{K} . \mathcal{S} is adding to the assessing process to enhance the snippet level query relevance. Scoring is guided by chain-of-thought (CoT) reasoning and few-shot prompt engineering, and is performed by \mathcal{E} . Self-Recognition is incorporated in \mathfrak{F} to provide multi-hop contextual links, thereby assessing the relevance of inference-based knowledge more accurately.

$$s_{\text{in-B}}^{i} = (Q \oplus \mathcal{S} \oplus k_{\text{P-B}}^{(1,\dots l)}) \odot x_{i} = (Q \oplus k_{\text{P-B}}^{(1,\dots l)}) \odot x_{i} + (\mathcal{S} \oplus k_{\text{P-B}}^{(1,\dots l)}) \odot x_{i}, \tag{A.13}$$

Relevance scores are normalized to the range [0,1] and discretized into five levels: 1, 0.75, 0.5, 0.25, and 0, representing decreasing degrees of relevance. The sequences with $s_i > 0$ are included in basic knowledge \mathcal{K} (Section 3.3). The prompt instances is provided in Appendix I.1.

B.2.6 MODULE: ASSESSOR ©

 \mathfrak{S} serves as an early-stop controller that determines when to terminate the extraction. The terminating condition depends on the following criteria: (1) whether the extracted knowledge is sufficient enough to solve the query, or (2) whether there is no further relevant knowledge can be extracted – namely, no more knowledge in the latest extraction round is extracted, or (3) whether the maximum extraction rounds $L_{\text{Max-Rounds}}$ has been reached. The stopping signal is formally defined as:

$$\tilde{s} = \begin{cases} 1, & \text{if } \tilde{a} = 1 \text{ or } \mathcal{K} \ge \mathcal{K}_{\text{prior}} \text{ or } r \ge L_{\text{Max-Rounds}}, \\ 0, & \text{otherwise.} \end{cases}$$
(A.14)

where \tilde{s} , \tilde{a} denote the *stop* and *Answerable* flag, respectively; r represents the current extraction round. And:

$$\tilde{a} = \mathbb{I}\left[Q \xrightarrow{\mathcal{K}} A\right] \tag{A.15}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function that outputs 1 if the condition holds, and 0 otherwise; A signifies the query answer.

B.3 Details of FRAG Generation

The basic knowledge \mathcal{K} extracted by FRAG extraction framework contains knowledge varying from high relevance to low relevance. To evaluate the performance of RALMs across knowledge with varying relevance, we set a relevance threshold for generation \mathcal{T}_G for the generation process. The knowledge with a relevance not lower than \mathcal{T}_G is selected as the basic knowledge for generation (\mathcal{K}') to construct the generation context to answer the query.

 \mathcal{T}_G has four threshold levels: 0.95, 0.7, 0.45, and 0.2. The reason for the slight difference between the \mathcal{T}_G threshold settings and the scoring levels which \mathfrak{F} set is that arithmetic performed by large models is not always reliable, and there are occasional arithmetic errors in calculating the final relevance scores, leading to inaccurate scoring. To mitigate the adverse effects of such occurrences, we slightly lowered the \mathcal{T}_G threshold during the design process. Before generation, \mathcal{K}' is also sorted in descending order of relevance score.

$$\mathcal{K}' = [(k'_u, s'_u), \dots, (k'_v, s'_v)], \quad s'_u \ge s'_{u+1} \ge \dots \ge s'_v \ge \mathcal{T}_G, \tag{A.16}$$

It should be noted that when there is no relevant knowledge in \mathcal{K}' , the initial retrieved documents will be used in the generation process. The final output is produced by the generator model \mathcal{G} . As is shown in the following formula:

$$\{Q, \mathcal{K}_G\} \xrightarrow{\mathcal{G}} A$$
 (A.17)

where:

$$\mathcal{K}_G = \begin{cases} \mathcal{K}', & \text{if } \mathcal{K}' \neq \emptyset \\ D_s, & \text{otherwise.} \end{cases}$$
 (A.18)

B.4 FRAG-IP: A WRAPPER FINE-TUNED FRAMEWORK FOR OPTIMIZING COMPUTATIONAL EFFICIENCY

B.4.1 OPTIMIZING STRATEGY OF FRAG-IP

Analysis of FRAG computational efficiency. Given an initially retrieved document $D_s = [d_1, \ldots, d_N]$ containing n passages, FRAG requires $(n+1) \sim [(n \times 5 + 2) \times L_{\text{Max-Rounds}} - 1]$ LLM calls to conduct extraction. The upper bound reflects the worst-case scenario; however, in most cases, only a small fraction of passages contain relevant information, which is commonly observed in practice. In addition, FRAG encounters data concurrency challenges due to inconsistent synchronized scheduling across modules during multi-query extraction. Moreover, the extraction process for a multi-hop query involves step-wise context dependencies, which constrain the potential for parallel extraction in such tasks. Those result in substantial computational overhead.

To address this challenge, <u>firstly</u>, we propose a wrapper framework which featuring *an asynchronous extraction pipeline*, orchestrated by a centralized hub scheduler. This design enables large-scale, asynchronous scheduling of multi-query extraction tasks in single GPU (for instance, FRAG can handle batch sizes of 256–512 queries on a single RTX 4090 GPU with this approach, in contrast to the batch size of 1 in the serial strategy). The wrapper framework not only significantly resolves data concurrency bottlenecks but also boosts extraction throughput by approximately 4–8×.

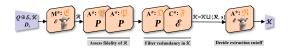


Figure A.6: Workflow of FRAG-ip extraction framework. \mathcal{P} signifies the program executor. Modules $\mathfrak{V}, \mathfrak{P}, \mathfrak{D}$ are transformed to be purely driven by \mathcal{P} .

FRAG-ip, a framework for significantly enhancing computational efficiency. Established on the wrapper framework, secondly, we further propose FRAG-ip, a wrapper fine-tuned framework which applies a dual-stage fine-tuning strategy to the extractor model (the detailed dual-stage fine-tuning strategy is present in Section 3.3). This approach enables more effective and consistent recognition and extraction of relevant sequences from a batch of t retrieved passages. Moreover, the modules $\mathfrak{V},\mathfrak{D}$ to be transformed to be purely driven by \mathcal{P} . Notably, \mathfrak{P} can be purely driven by \mathcal{P} as well, since instances in typical datasets contain a structured "title" field. These improvements notable simplify the extraction process. Figure A.6 illustrates the workflow of the improved FRAG extraction framework, and Algorithm A.2 illustrates the FRAG-ip extraction.

The fine-tuned extractor model \mathcal{E}_T extracts relevant and consistent sequences from a batch of t retrieved passages and significantly reduces the LLM calls required for extraction to $(\lceil \frac{n}{t} \rceil + 1) \sim [(\lceil \frac{n}{t} \rceil \times 2 + 2) \times L_{\text{Max-Rounds}} - 1]$. This further decreases the extraction runtime by about $10 \times$.

Algorithm A.2 Improved FRAG Extraction

```
Require: Query Q; Q's key sequence snippets S; Retrieved documents D_s = [d_1, \ldots, d_N]; FRAG
   extraction framework \mathcal{F}; Basic knowledge \mathcal{K}; Batch size of passages to be extracted per step t;
   Relevant sequences set \{x_i, \ldots, x_{i+m}\} extracted a batch of t passages; Relevance score of \{Q, x_i\},
   s_i; Max extraction rounds limit L_{\text{Max-Rounds}}; The i-th retrieved passage interval d_{i \times t} (obtained by
   dividing D_s into t equal parts); Constraint: \mathcal{K} does not exceed \mathcal{K}_{prior}, i.e., \mathcal{K} \leq \mathcal{K}_{prior}.
   Initialize ans = false; stop = false; r = 1;
                                                                                          \triangleright Decompose S from Q (Section 3.1)
   while {ans is false and stop is false and r \leq L_{\text{Max-Rounds}}} do
         \mathcal{K}_{\text{prior}} \leftarrow \mathcal{K}
         for i=1,\ldots,\lceil n/t\rceil do
              if K is \varnothing then
                    [(x_i, s_i)], \dots, [(x_{i+m}, s_{i+m})] \stackrel{\mathcal{F}}{\leftarrow} \{Q \oplus \mathcal{S}, \tilde{d}_{i \times t}\}
                          Extract \{x_i, \dots, x_{i+m}\} from t passages & Assess \{s_i, \dots, s_{i+m}\} (Section 3.3)
                   [(x_i, s_i)], \dots, [(x_{i+m}, s_{i+m})] \stackrel{\mathcal{F}}{\leftarrow} \{Q \oplus \mathcal{S}, \tilde{d}_{i \times t}, \mathcal{K}\} \triangleright \text{Self-Recognition (Section 3.2)}
              \mathcal{K} \leftarrow \mathcal{K} \cup \{(x_i, s_i), \dots, (x_k, s_k)\}, \text{ with } s_i, \dots, s_k > 0
         \mathcal{S} \xleftarrow{\mathcal{F}} Q \oplus \mathcal{K}; r \leftarrow r + 1
                                                                                                          \triangleright Reidentify S if necessary
         Assess {ans: true \text{ or } false} \stackrel{\mathcal{F}}{\leftarrow} \{Q, \mathcal{K}\}
                                                                                       \triangleright Assess if Q is answerable based on K
         if K \leq K_{prior} then
              stop \leftarrow true
                                                                                         No more Basic Knowledge extracted
   Return K
```

B.4.2 DUAL-STAGE FINE-TUNING PROCESS

Fine-tuning datasets construction. For training dataset construction, we select 2,000 instances from 2WikiMultihopQA (Ho et al., 2020) and 1,000 instances from MuSiQue (Trivedi et al., 2022) as the raw instances. For the validation datasets, we sample 300 and 200 instances from the respective datasets. We utilize DeepSeek-v3 (Liu et al., 2024a) as generator to establish the fine-tuning datasets. For the first stage, the model is instructed to recognize the key sequence snippets within the raw instances, and subsequently extract relevant sequences $\mathcal S$ from the given t passages using snippet-level query relevance (t=5 in our fine-tuning). We validate the consistency of extracted sequences by $\mathcal P$ at string level. The generated training and validation datasets for this stage comprise 14410 and 2345 instances, respectively. For the second stage, each extracted sequence is prefixed with its corresponding main topic or title via $\mathcal P$, and the model is prompted to assign a relevance score of 1, 0.5, or 0 to each sequence, corresponding to relevant, partially relevant and irrelevant. Subsequently, the model is instructed to assess whether the queries in the raw instances are "Answerable", "Partially Answerable", or "Unanswerable", based on the basic knowledge extracted. The generated training and validation datasets for this stage is 7488 and 1203, respectively. The prompts for datasets generation are present in Section I.1.

Fine-tuning loss and framework. We adopt the token-level cross-entropy loss (Zhang & Sabuncu, 2018) in the fine-tuning, as formally defined:

$$\mathcal{L} = -\frac{1}{T_{\text{eff}}} \sum_{t=1}^{T} \mathbb{I}_{\{x_{t+1} \neq \text{PAD}\}} \cdot \log \left(\frac{\exp(z_{t,x_{t+1}})}{\sum_{j=1}^{V} \exp(z_{t,j})} \right)$$
(A.19)

where T is the length of the input sequence; x_{t+1} is the target token to predict at time step t; $z_{t,j}$ is the logit output by the model at time step t for the j-th vocabulary token; V is the vocabulary size; $\mathbb{I}_{\{x_{t+1} \neq \text{PAD}\}}$ is an indicator function to mask out padding tokens; and T_{eff} is the number of effective tokens (i.e., non-padding) used for normalization. We leverage the Huggingface PEFT (Parameter-Efficient Fine-Tuning) framework (Mangrulkar et al., 2022) with the LoRA (Hu et al., 2022) strategy to enable efficient adaptation. The base model used for fine-tuning is Qwen2-7B-Instruct (Yang et al., 2024).

Fine-tuning parameters settings and hardware configuration. The Lora rank is set to 32, resulting in approximately 80.1M trainable parameters within the LoRA adapters, accounting for around 1.05% of the full 7.7B model parameters. Fine-tuning is conducted on 7 RTX A100 GPUs (40GB) with 5/1 epochs for the dual stages, respectively (notably, the second-stage fine-tuning converges rapidly due to the initialization from the first stage, and early stopping is recommended to mitigate catastrophic forgetting in this stage). The total fine-tuning process requires approximately 18.58 GPU-hours.

B.4.3 COMPUTATIONAL EFFICIENCY OF FRAG AND RESULTS WITH FRAG-IP

Table A.5: **Overall experiment results using FRAG-ip. Bold** numbers indicate the best result across the compared models. *t* signifies the number of passages to be extracted per step.

	S	ingle-ho	p		Multi-hop					Long-Context
	Pop	Pub	ARC	HotPot	QA	2Wil	κi	MuSiQue		LongBench-v2
LMs	(acc)	(acc)	(acc)	(llm-em)	(em)	(llm-em)	(em)	(llm-em)	(em)	(acc)
		_								
Qwen2-7B-Instruct	51.82	75.08	76.83	57.4	34.9	50.2	42.2	23.5	9.2	22.66
						ours				
FRAG-Qwen2-7B-Instruct	57.97	79.84	82.03	72.7	47.7	60.9	51.5	44.1	25.6	29.82
FRAG-ip-Qwen2-7B ($t = 1$)	53.11	79.43	81.60	59.3	37.2	58.6	52.1	27.9	14	27.24
FRAG-ip-Qwen2-7B ($t = 5$)	50.82	76.09	80.24	58.4	34.5	58.2	52.4	21.8	12	27.04

FRAG-ip performance. As demonstrated in Table A.5, when extracting from a batch of passages with t=1, FRAG-ip achieves a performance close to that of FRAG, with an accuracy loss of 0.5%-5% in most cases. Meanwhile, as present in Table A.6, it significantly improves computational efficiency by approximately $10\times$ compared to FRAG wrapped in a pipeline. Moreover, it approaches the efficiency level of naive RAG while yielding significantly better performance.

Table A.6: Comparison of FRAG-ip and FRAG in Terms of Computational Efficiency. **Bold** numbers indicate the minimum average runtime on the test dataset.

		Pop(QA		HotPotQA				
LMs	bs	T	$L_{ m mr}$	$n_{ m ar}$	bs	T	$L_{ m mr}$	$n_{ m ar}$	
Qwen2-7B-Instruct	512	0.47	-	-	512	0.5	-	-	
FRAG-Qwen2-7B-Instruct w/o pipeline	1	85.03	1	1	2	121.35	3	2.65	
FRAG-Qwen2-7B-Instruct w pipeline	256	11.47	1	1	256	15.07	3	2.65	
FRAG-ip-Qwen2-7B ($t = 1$)	256	1.46	1	1	256	1.52	3	2.98	
FRAG-ip-Qwen2-7B ($t = 5$)	256	1.02	1	1	256	1.07	3	2.96	

Note: bs represents the *batch size* for FRAG extraction or naive RAG generation; T denotes the average seconds spent per query; L_{mr} stands for the maximum extraction rounds limit ($L_{Max-Rounds}$); n_{ar} represents the average number of extraction rounds per query.

In addition, although the extractor model is fine-tuned on a relatively small dataset – consisting of only 3000 instances sampled from 2WikiMultihop and MuSiQue – it achieves a competitive performance on the single-hop, long-context and the 2WikiMultihop tasks with FRAG. This demonstrates the strong generalizability of FRAG in enhancing RALM performance by filtering noise using snippet-level query relevance. Furthermore, these results suggest promising potential for further accuracy improvements using the wrapper fine-tuning framework with the extractor model trained on larger and more diverse datasets.

It is worth noting that when t=5, computational efficiency can be further improved, albeit at the cost of increased accuracy degradation compared with FRAG. Given that the efficiency gain at t=1 is sufficient for most practical applications, we recommend choosing a small value of t to avoid notable performance degradation.

Table A.7: Latency comparison on HotPotQA.

Models	(EM)	(T)
SelfRAG-7B	12.9	1.44
SelfRAG-13B	13.2	1.55
LongRAG-6B	40.5	11
RQ-RAG-7B	0	9.54
Qwen2-7B-Instruct	34.9	0.50
FRAG-Qwen2-7B-Instruct	47.7	15.07
FRAG-ip-Qwen2-7B	37.2	1.07

We also compare the latency of FRAG with other baselines, as show in Table A.7. Results show that FRAG achieves the best performance with only marginal latency overhead, while FRAG-ip significantly reduces latency with competitive performance.

B.5 COMPARISON OF FRAG WITH CONCURRENT BASELINES

Comparison with reranking baselines. Reranking methods, as adopted in baselines such as RankGPT (Sun et al., 2023) and RankRAG (Yu et al., 2024c), represent another noise-filtering strategy. These approaches rely on reranking the retrievals using a reranking model and selecting the top-k most relevant passages for generation, thereby mitigating the negative impact of noise and enhance the RALM performance. In contrast, FRAG fundamentally differs from these methods. Reranking approaches inherently face a trade-off: selecting a small k leads to information loss, while a large k introduces excessive noise. Moreover, these methods suffer from the disability to filter intra-passage noise. FRAG overcomes the disadvantages by performing snippet-level extraction and filtering noise in a fine-grained manner.

Comparison with query reformulation techniques. The query reformulation techniques typically utilize LLMs to reformulate the input query in order to improve retrieval recall. However, these methods primarily focus on enhancing the retrieval phase and do not explicitly address noise filtering in the generation stage. This makes their design fundamentally different from that of FRAG, which aims at filtering noise in a fine-grained manner to mitigate the attention distraction in RALMs caused by excessive noise.

C PROOF OF MAIN THEORETICAL ANALYSIS

C.1 PROOF OF THEOREM 2.2

Proof. Calculate $A_{\mathfrak{T}}$ for \mathfrak{T} (Vaswani et al., 2017) (W_Q, W_K, W_V) and W_V represent the pre-trained Q, K, V matrices of the LLM):

$$\mathbf{A}_{\mathfrak{T}}\left[Q_{\mathsf{Q}},(K_{\mathsf{rel}} \oplus K_{\mathsf{red}}),(V_{\mathsf{rel}} \oplus V_{\mathsf{red}})\right] = \mathsf{SoftMax}\left(\frac{Q_{\mathsf{Q}}(K_{\mathsf{rel}}^{\intercal},K_{\mathsf{red}}^{\intercal})}{\sqrt{d_{k}}}\right)(V_{\mathsf{rel}} \oplus V_{\mathsf{red}})$$

$$= \mathsf{SoftMax}\left(\frac{\left[X_{Q}W_{Q}W_{K}^{\intercal}X_{\mathsf{rel}}^{\intercal},X_{Q}W_{Q}W_{K}^{\intercal}X_{\mathsf{red}}^{\intercal}\right]}{\sqrt{d_{k}}}\right) \cdot (X_{\mathsf{rel}}W_{\mathsf{rel}} \oplus X_{\mathsf{red}}W_{\mathsf{red}})$$

$$= \mathsf{SoftMax}\left([M,N]\right)(X_{\mathsf{rel}}W_{V} \oplus X_{\mathsf{red}}W_{V}) = \alpha(X_{\mathsf{rel}}W_{V} \oplus X_{\mathsf{red}}W_{V}), \tag{A.20}$$

where:

$$\mathbf{M} = \begin{bmatrix} a_{(1,1)} & \cdots & a_{(1,n_{\text{rel}})} \\ \vdots & \ddots & \vdots \\ a_{(n_q,1)} & \cdots & a_{(n_q,n_{\text{rel}})} \end{bmatrix}, \mathbf{N} = \begin{bmatrix} b_{(1,1)} & \cdots & b_{(1,n_{\text{red}})} \\ \vdots & \ddots & \vdots \\ b_{(n_q,1)} & \cdots & b_{(n_q,n_{\text{red}})} \end{bmatrix},$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{\text{rel}(1,1)} & \cdots & \alpha_{\text{rel}(1,n_{\text{rel}})} & \alpha_{\text{rel}(1,1)} & \cdots & \alpha_{\text{red}(1,n_{\text{red}})} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{\text{rel}(n_q,1)} & \cdots & \alpha_{\text{rel}(n_q,n_{\text{rel}})} & \alpha_{\text{red}(n_q,1)} & \cdots & \alpha_{\text{red}(n_q,n_{\text{red}})} \end{bmatrix}.$$
(A.21)

Thus we have:

$$\alpha_{\text{rel}(i,j)} = \frac{e^{a_{(i,j)}}}{\sum\limits_{i=1}^{n_{\text{rel}}} e^{a_{(i,j)}} + \sum\limits_{i=1}^{n_{\text{red}}} e^{b_{(i,j)}}}, \alpha_{\text{red}(i,j)} = \frac{e^{b_{(i,j)}}}{\sum\limits_{i=1}^{n_{\text{rel}}} e^{a_{(i,j)}} + \sum\limits_{i=1}^{n_{\text{red}}} e^{b_{(i,j)}}}, \tag{A.22}$$

and:

$$\sum_{j=1}^{n_{\text{rel}}} \alpha_{\text{rel}(i,j)} + \sum_{j=1}^{n_{\text{red}}} \alpha_{\text{red}(i,j)} = 1, \quad \forall i \in [1, n_q].$$
(A.23)

Let n_{rel} remains constant. Since: $0 < \alpha_{\mathrm{rel}(i,j)}, \alpha_{\mathrm{red}(i,j)} < +\infty$, we can conclude that as n_{red} grows, $\sum_{j=1}^{n_{\mathrm{rel}}} \alpha_{\mathrm{rel}(i,j)}$ decreases, while $\sum_{j=1}^{n_{\mathrm{red}}} \alpha_{\mathrm{red}(i,j)} = 1 - \sum_{j=1}^{n_{\mathrm{rel}}} \alpha_{\mathrm{rel}(i,j)}$ increases. Thus, there exists $\tilde{n} \in \mathbb{N}^+$, such that for $n_{\mathrm{red}} > \tilde{n}$, we have:

$$\sum_{i=1}^{n_q} \sum_{j=1}^{n_{\text{rel}}} \alpha_{\text{rel}(i,j)} < \sum_{i=1}^{n_q} \sum_{j=1}^{n_{\text{red}}} \alpha_{\text{red}(i,j)}. \tag{A.24}$$

Namely, we have:

$$\mathfrak{A}_{\text{rel}} < \mathfrak{A}_{\text{red}}.$$
 (A.25)

In this case, the noise dilutes the relevant information, causing the RALM to allocate more attention to X_{red} than X_{rel} , and generate \mathfrak{T} based on X_{red} , subsequently producing content unrelated to answering Q. For example, it may respond with information from X_{red} or provide ambiguous answers such as "Not enough information to answer the query".

C.2 Proof of Discussion in Section B.1

Proof. The average proportion of relevant information in the t relevant retrieval passages is higher than in a long-context document composed of N passages. For the initial document containing N passages, the noise rate is expressed as follows:

$$\rho_0 = \frac{\frac{n_{\text{red}}}{N}}{\frac{n}{N}} = \frac{n_{\text{red}}}{n}.$$
(A.26)

Next, for a single extraction step from a single passage, assume that the sequences forming X_{rel} and X_{red} are uniformly distributed across all passages. The expected value of the noise rate for a single passage is given by:

$$\mathbb{E}(\rho_{\text{red}}^*) = \bar{\rho}_{\text{red}}^* = \frac{\frac{n_{\text{red}}}{N}}{\frac{n}{N}} = \frac{n_{\text{red}}}{n} = \rho_0. \tag{A.27}$$

However, since the relevant information is concentrated in t paragraphs (where $t \leq N$), we assume that the sequences forming $X_{\rm rel}$ are uniformly distributed across these t retrieval passages. Thus for the t relevant passages, the expected noise rate is given by:

$$\mathbb{E}(\rho_{\text{red}}) = \bar{\rho}_{\text{red}} = \frac{\frac{n}{N} - \frac{n_{\text{rel}}}{t}}{\frac{n}{N}} = 1 - \frac{n_{\text{rel}}}{n} \cdot \frac{N}{t} = 1 - \frac{n - n_{\text{red}}}{n} \cdot \frac{N}{t} = 1 - \left[(1 - \rho_0) \cdot \frac{N}{t} \right] \le \rho_0.$$
(A.28)

Therefore, decomposing the process into several steps, each focusing on extracting information from a single retrieval passage, not only reduces the length of noisy information but also lowers the proportion of noise within the t relevant passages. Accordingly, for these t relevant passages, the proportion of relevant information is higher than that in the initial retrieval document. This, in turn, mitigates the attention diversion caused by noisy information during the extraction process. \Box

C.3 PROOF OF LEMMA 3.1

Proof. Negative impact of non-key sequence snippets in the query X_{Q^-} . During the extraction process, the LLM's attention distribution over the retrieval passage $X_{\mathrm{red},\subset} \oplus X_{\mathrm{red},\subset}$, based on X_Q , is as follows:

$$\mathbf{A}_{\mathfrak{T}}[(Q, (K_{\mathrm{rel},\subset} \oplus K_{\mathrm{red},\subset}), (V_{\mathrm{rel},\subset} \oplus V_{\mathrm{red},\subset}))]$$

$$= \operatorname{SoftMax} \left[\frac{(Q_{Q^{+}} \oplus Q_{Q^{-}})[K_{\mathrm{rel},\subset}^{T}, K_{\mathrm{red},\subset}^{T}]}{\sqrt{d_{k}}} \right] \cdot (V_{\mathrm{rel},\subset} \oplus V_{\mathrm{red},\subset})$$

$$= \operatorname{SoftMax} \left[[(X_{Q^{+}} \oplus X_{Q^{-}})W_{Q}W_{K}^{T}X_{\mathrm{rel},\subset}^{T}, (X_{Q^{+}} \oplus X_{Q^{-}})W_{Q}W_{K}^{T}X_{\mathrm{red},\subset}^{T}]/\sqrt{d_{k}} \right]$$

$$\cdot (X_{\mathrm{rel},\subset}W_{V} \oplus X_{\mathrm{red},\subset}W_{V})$$

$$= \operatorname{SoftMax} \left(\begin{bmatrix} M_{s0^{+}} & N_{s0^{+}} \\ M_{s0^{-}} & N_{s0^{-}} \end{bmatrix} \right) (X_{\mathrm{rel},\subset}W_{V} \oplus X_{\mathrm{red},\subset}W_{V}) = \alpha_{\subset}(X_{\mathrm{rel},\subset}W_{V} \oplus X_{\mathrm{red},\subset}W_{V})$$

$$= \begin{bmatrix} \alpha_{\mathrm{rel},\subset}^{+}, \alpha_{\mathrm{red},\subset}^{+} \\ \alpha_{\mathrm{rel},\subset}^{-}, \alpha_{\mathrm{red},\subset}^{-} \end{bmatrix} (X_{\mathrm{rel},\subset}W_{V} \oplus X_{\mathrm{red},\subset}W_{V}). \tag{A.29}$$

It is evident that, under general circumstances, we have:

$$\frac{1}{n_{q^{+}}} \sum_{i=1}^{n_{q^{+}}} \sum_{i=1}^{n_{l}} \alpha_{\text{rel},\subset(i,j)}^{+} \ge \frac{1}{n_{q^{-}}} \sum_{i=1}^{n_{q^{-}}} \sum_{i=1}^{n_{l}} \alpha_{\text{rel},\subset(i,j)}^{-}, \tag{A.30}$$

where n_{q^+}, n_{q^-} denote the token counts of X_{Q^+}, X_{Q^-} , respectively. Accordingly, we have:

$$\frac{1}{n_{q^{+}}} \sum_{i=1}^{n_{q^{+}}} \left(1 - \sum_{j=1}^{n_{l}} \alpha_{\text{rel},\subset(i,j)}^{+} \right) \leq \frac{1}{n_{q^{-}}} \sum_{i=1}^{n_{q^{-}}} \left(1 - \sum_{j=1}^{n_{l}} \alpha_{\text{rel},\subset(i,j)}^{-} \right)
\Rightarrow \frac{1}{n_{q^{+}}} \sum_{i=1}^{n_{q^{+}}} \sum_{j=1}^{n_{d}} \alpha_{\text{red},\subset(i,j)}^{+} \leq \frac{1}{n_{q^{-}}} \sum_{i=1}^{n_{q^{-}}} \sum_{j=1}^{n_{d}} \alpha_{\text{red},\subset(i,j)}^{-}.$$
(A.31)

Combining Inequality A.30 and A.31, we have:

$$a_{1} = \frac{\sum_{i=1}^{n_{q^{+}}} \sum_{j=1}^{n_{l}} \alpha_{\text{rel}, \subset (i,j)}^{+}}{\sum_{i=1}^{n_{q^{-}}} \sum_{j=1}^{n_{l}} \alpha_{\text{rel}, \subset (i,j)}^{-}} \ge \frac{n_{q^{+}}}{n_{q^{-}}},$$
(A.32)

$$b_1 = \frac{\sum_{i=1}^{n_{q^+}} \sum_{j=1}^{n_d} \alpha_{\text{red}, \subset (i,j)}^+}{\sum_{i=1}^{n_{q^-}} \sum_{j=1}^{n_d} \alpha_{\text{red}, \subset (i,j)}^-} \le \frac{n_{q^+}}{n_{q^-}}.$$
(A.33)

Now, we attempt to prove that the Inequality A.34 holds:

$$\frac{\sum_{\substack{1 \le i \le n_{q^{+}} \\ 1 \le j \le n_{l}}} \alpha_{\text{rel}(i,j)}^{+}}{\sum_{\substack{1 \le i \le n_{q^{+}} \\ 1 \le j \le n_{d}}} \alpha_{\text{red}(i,j)}^{+}} \ge \frac{\sum_{\substack{1 \le i \le n_{q^{+}} \\ 1 \le j \le n_{l}}} \alpha_{\text{rel}(i,j)}^{+} + \sum_{\substack{1 \le i \le n_{q^{-}} \\ 1 \le j \le n_{l}}} \alpha_{\text{rel}(i,j)}^{-}}{\sum_{\substack{1 \le i \le n_{q^{+}} \\ 1 \le j \le n_{d}}} \alpha_{\text{red}(i,j)}^{+} + \sum_{\substack{1 \le i \le n_{q^{-}} \\ 1 \le j \le n_{d}}} \alpha_{\text{red}(i,j)}^{-}}^{-},$$
(A.34)

where n_l, n_d denote the token counts of $X_{\mathrm{rel}, \subset}, X_{\mathrm{red}, \subset}$, respectively.

Substitute a_1 and b_1 into Inequality A.34. Since $a_1 \ge \frac{n_{q^+}}{n_{q^-}} \ge b_1 > 0$, Inequality A.34, i.e., Inequality 1 is proven.

That is, in effect, the non-key sequence snippets of the query weaken the RALM's attention distribution of X_Q towards the relevant information in the retrieval passage.

C.4 Proof of Theorem 3.2

Proof. Benefits of snippet-Level query relevance in extraction. By identifying and extracting the key sequence snippets X_{Q^+}' from Q, we can explicitly add X_{Q^+}' , thereby increasing its corresponding query weight matrix parameters and enhancing the LLM's attention distribution towards the relevant information. On the other hand, by increasing the proportion of key sequences during extraction, the LLM's attention distribution towards the relevant information in the retrieval passages is further enhanced. The analysis is shown as follows:

$$\begin{array}{ll} \mathbf{1380} & A_{\mathfrak{T}}[(Q,(K_{\mathrm{rel},\subset} \oplus K_{\mathrm{red},\subset}),(V_{\mathrm{rel},\subset} \oplus V_{\mathrm{red},\subset}))] \\ \mathbf{1381} & = \mathrm{SoftMax} \left[\frac{(Q_{Q^+} \oplus Q_{Q^-} \oplus Q'_{Q^+})[K^T_{\mathrm{rel},\subset},K^T_{\mathrm{red},\subset}]}{\sqrt{d_k}} \right] \cdot (V_{\mathrm{rel},\subset} \oplus V_{\mathrm{red},\subset}) \\ \mathbf{1383} & = \mathrm{SoftMax} \left[[(X_{Q^+} \oplus X_{Q^-} \oplus X'_{Q^+})W_QW_K^TX_{\mathrm{rel},\subset}^T, (X_{Q^+} \oplus X_{Q^-} \oplus X'_{Q^+})W_QW_K^TX_{\mathrm{red},\subset}^T]/\sqrt{d_k} \right] \\ \mathbf{1386} & \cdot (X_{\mathrm{rel},\subset}W_V \oplus X_{\mathrm{red},\subset}W_V) \\ \mathbf{1387} & = \mathrm{SoftMax} \left(\begin{bmatrix} M_{s0^+} & N_{s0^+} \\ M_{s0^-} & N_{s0^-} \\ M_{s'_+} & N_{s'_+} \end{bmatrix} \right) (X_{\mathrm{rel},\subset}W_V \oplus X_{\mathrm{red},\subset}W_V) = \boldsymbol{\alpha}_s'(X_{\mathrm{rel},\subset}W_V \oplus X_{\mathrm{red},\subset}W_V) \\ \mathbf{1390} & = \begin{bmatrix} \boldsymbol{\alpha}_{\mathrm{rel},\subset}^+ & \boldsymbol{\alpha}_{\mathrm{red},\subset}^+ \\ \boldsymbol{\alpha}_{\mathrm{rel},\subset}^- & \boldsymbol{\alpha}_{\mathrm{red},\subset}^- \\ \boldsymbol{\alpha}_{\mathrm{rel},\subset}^- & \boldsymbol{\alpha}_{\mathrm{red},\subset}^- \end{bmatrix} (X_{\mathrm{rel},\subset}W_V \oplus X_{\mathrm{red},\subset}W_V). \end{array} \tag{A.35}$$

Combining Lemma 3.1, we have:

$$\frac{1}{n_{q'_{+}}} \sum_{i=1}^{n_{q'_{+}}} \sum_{j=1}^{n_{l}} \alpha'^{+}_{\text{rel},\subset(i,j)} \ge \frac{1}{n_{q^{+}}} \sum_{i=1}^{n_{q^{+}}} \sum_{j=1}^{n_{l}} \alpha^{+}_{\text{rel},\subset(i,j)} \ge \frac{1}{n_{q}} (\sum_{i=1}^{n_{q^{+}}} \sum_{j=1}^{n_{l}} \alpha^{+}_{\text{rel},\subset(i,j)} + \sum_{i=1}^{n_{q^{-}}} \sum_{j=1}^{n_{l}} \alpha^{-}_{\text{rel},\subset(i,j)});$$
(A.36)

$$\frac{1}{n_{q'_{+}}} \sum_{i=1}^{n_{q'_{+}}} \sum_{j=1}^{n_{d}} \alpha'^{+}_{\text{red},\subset(i,j)} \leq \frac{1}{n_{q^{+}}} \sum_{i=1}^{n_{q^{+}}} \sum_{j=1}^{n_{d}} \alpha^{+}_{\text{red},\subset(i,j)} \leq \frac{1}{n_{q}} (\sum_{i=1}^{n_{q^{+}}} \sum_{j=1}^{n_{d}} \alpha^{+}_{\text{red},\subset(i,j)} + \sum_{i=1}^{n_{q^{-}}} \sum_{j=1}^{n_{d}} \alpha^{-}_{\text{red},\subset(i,j)}). \tag{A.37}$$

It is obviously that:

$$a_{2} = \frac{\sum_{i=1}^{n_{q'_{+}}} \sum_{j=1}^{n_{l}} \alpha'_{\text{rel}, \subset (i,j)}^{+}}{\sum_{i=1}^{n_{q+}} \sum_{j=1}^{n_{l}} \alpha_{\text{rel}, \subset (i,j)}^{+} + \sum_{i=1}^{n_{q-}} \sum_{j=1}^{n_{l}} \alpha_{\text{rel}, \subset (i,j)}^{-}} \ge \frac{n_{q'_{+}}}{n_{q}}, \tag{A.38}$$

$$b_{2} = \frac{\sum_{i=1}^{n_{q'_{+}}} \sum_{j=1}^{n_{d}} \alpha_{\text{red}, \subset (i, j)}^{\prime +}}{\sum_{i=1}^{n_{q+}} \sum_{j=1}^{n_{d}} \alpha_{\text{red}, \subset (i, j)}^{+} + \sum_{i=1}^{n_{q-}} \sum_{j=1}^{n_{d}} \alpha_{\text{red}, \subset (i, j)}^{-}} \le \frac{n_{q'_{+}}}{n_{q}}.$$
(A.39)

Now, we attempt to prove that the Inequality A.40 holds:

$$\frac{\sum_{1 \leq i \leq n_{q^{+}}} \alpha_{\text{rel}(i,j)}^{+} + \sum_{1 \leq i \leq n_{q^{-}}} \alpha_{\text{rel}(i,j)}^{-} + \sum_{1 \leq i \leq n_{q'}} \alpha_{\text{rel}(i,j)}^{+}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n_{l}}{1 \leq j \leq n_{l}} \frac{1 \leq j \leq n$$

where $n_{q'_+}$ denotes the token count of X'_{Q^+} .

Substitute a_2 and b_2 into Inequality A.40. Since $a_2 \ge \frac{n_{q'_+}}{n_q} \ge b_2 > 0$, Inequality A.40, i.e., Inequality 2 is proven.

This implies the LLM allocates a greater proportion of attention to $X_{\mathrm{rel},\subset}$ when incorporating X'_{Q^+} into X_Q during extraction. Combining Lemma 3.1 and Inequality 2, we conclude that explicitly incorporating X'_{Q^+} into X_Q during extraction, i.e., performing extraction using snippet-level query relevance, improves the model's attention allocation towards relevant information. This reduces the attention reduction caused by X_{Q^-} , enhancing the extraction of relevant knowledge.

D ADDITIONAL EXPERIMENT AND DETAILS

D.1 ADDITIONAL EXPERIMENT

Table A.8: Comparison between FRAG and baselines using randomly clipped retrievals. Bold numbers indicate the best performance among the compared models. r denotes the ratio of input tokens used for generation to the total number of tokens in the initial retrievals.

Models	Pop (acc)	Pub (acc)	ARC (acc)	HotpotQA (em/f1)	2Wiki (em/f1)	MuSiQue (em/f1)	LongBench-v2 (acc)			
Baselines w/ Randomly Clipped Retrievals										
ChatQA-1.5-8B	29.74	0.71	52.56	15.22/23.92	37.74/39.97	2.1/8.22	7.55			
Llama3-8B-Instruct	28.16	70.92	75.30	12.11/17.12	5.21/9.22	2.2/5.34	19.28			
Qwen2-7B-Instruct	25.95	73.86	77.94	17.22/24.44	16.82/20.42	4.4/9.16	24.25			
ChatQA-1.5-70B	37.60	55.72	74.45	21.10/30.94	31.00/34.81	6.3/14.62	23.21			
				Ours						
FRAG-ChatQA-1.5-8B	54.11	73.66	46.42	27.9/44.81	44.2/48.0	12.4/23.7	21.47			
FRAG-Llama3-8B-Instruct	59.83	75.99	77.34	41.7/58.82	32.1/38.7	22.9/33.9	24.45			
FRAG-Qwen2-7B-Instruct	57.97	79.84	82.03	47.7/64.08	51.5/58.5	25.6/40.2	29.82			
FRAG-ChatQA-1.5-70B	59.69	76.49	82.79	38.4/53.92	28.4/36.7	19.3/32.3	29.03			

To investigate whether the performance gain of FRAG mainly stems from denoising rather than input length, we conducted an additional experiment. Specifically, we controlled the number of input tokens for naive RAG by randomly clipping retrievals to match the token budget of FRAG. As shown in Table A.8, FRAG still significantly outperforms these baselines, highlighting its effectiveness in improving RALMs by filtering out noise from the initial retrievals.

Moreover, To explore how varying the retrieval window size N impacts model performance, we conduct the experiments by evaluating different N values on PopQA and HotpotQA using both FRAG and naive RAG. The results are present in Table A.9. We further evaluate FRAG on DeepSeek-v3

Table A.9: Performance comparison of FRAG and naive RAG under various values of N. Bold numbers denote the best performance among the compared models. For HotpotQA, when N=10, we use the original passages associated with each query in the dataset as retrievals.

Models		PopQA		HotpotQA							
	(recall)	(acc)	(T)	(recall)	(em)	(f1)	(T)				
N=3											
Qwen2-7B-Instruct	-	-	-	65.1	27.93	39.36	0.2				
FRAG-Qwen2-7B-Instruct	-	-	-	65.1	31.03	42.31	6.0				
N=5											
Qwen2-7B-Instruct	50.25	40.46	0.2	73.75	29.83	41.12	0.2				
FRAG-Qwen2-7B-Instruct	50.25	42.53	1.4	73.75	33.23	44.43	11.6				
		N =	10								
Qwen2-7B-Instruct	65.90	49.61	0.4	100	34.90	50.38	0.9				
FRAG-Qwen2-7B-Instruct	65.90	51.25	2.6	100	47.70	64.08	15.07				
N = 20											
Qwen2-7B-Instruct	73.72	51.82	2.5	-	-	-	-				
FRAG-Qwen2-7B-Instruct	73.72	59.69	11.47	-	-	-	-				

(685B) to examine its performance on extremely large models and identify potential saturation points. As shown in Table A.10, FRAG continues to enhance the performance of very large models by effectively filtering noise, with particularly strong gains on complex multi-hop tasks. Table A.10 also shows that performance gains tend to saturate on simple single-hop tasks, and in cases where retrieval quality is notably low (e.g., LongBench-v2). Nevertheless, on complex multi-hop datasets, FRAG still yields substantial improvements on extremely large models, achieving average gains of 8.23% EM and 6.53% F1.

Table A.10: Performance comparison between FRAG-DeepSeek-v3 and DeepSeek-v3. Bold numbers indicate the best performance among the compared models.

Models	PopQA (acc)	PubQA (acc)	ARC (acc)	HotpotQA (em/f1)	2Wiki (em/f1)	MuSiQue (em/f1)	LongBench-v2 (acc)
			Baseli	nes w/o Retriev	als		
DeepSeek-v3	31.17	71.02	95.74	31.9 / 45.5	43.9 / 46.6	9.2 / 19.27	33.60
			Basel	ines w/ Retriev	als		
DeepSeek-v3 FRAG-DeepSeek-v3	70.50 71.26	83.55 85.54	94.97 94.97	57.86 / 76.3 64.23 / 81.69	77.98 / 84.08 81.5 / 85.99	46.74 / 63.18 61.53 / 75.48	36.16 37.17

To further assess the effectiveness of the five modules (excluding the Extractor, whose ablation is reported in Section 4.5), we conduct ablation experiments by removing each of the five modules individually, as well as an additional experiment in which all five are removed simultaneously. As shown in Table A.11, each module proves critical for either enhancing extraction accuracy, reducing noise, or improving computational efficiency. The Validator ensures that the extracted sequences remain consistent with the original retrievals, thereby preventing hallucination-induced factual errors; without it, numerous incorrect sequences persist, leading to noisy inference-based knowledge extraction, lower retrieval accuracy, and a sharp drop in η . The Prefixer supplements extracted sequences with main topics to mitigate information loss, whereas its removal results in hindered inference despite a marginal increase in η . The Deduplicator eliminates repeated sequences; without it, noise remains in the extracted knowledge, lowering η and overall performance. The Filter assigns relevance scores to the extracted sequences and removes low-relevance ones; without it, irrelevant sequences distract the extractor, degrade accuracy, and reduce performance. The Assessor enables early stopping; removing it leads to maximal extraction time (second only to the w/o Validator case, where unnecessary FRAG steps are applied to unvalidated content). Finally, removing all five

modules results in the most severe degradation, with performance dropping markedly and time cost increasing by nearly 3×. These results collectively demonstrate that the five modules are indispensable for maintaining high-quality, relevant knowledge extraction, while also improving computational efficiency by reducing noise and preventing additional LLM calls triggered by inaccurate extraction.

Table A.11: Results of the extended ablation study. r_{ete} denotes the retention ratio of the golden passages after extraction; η denotes the filtering rate of tokens from noise in the extracted documents compared to the initial retrieved documents; t denotes the average extraction time.

Models	HotpotQA								
	(em)	(f1)	(rete)	(η)	(T)				
FRAG	47.7	64.08	90.15	82.87	15.07				
w/o Validator	42.2	57.99	84.9	45.14	19.35				
w/o Prefixer	43.8	59.34	83.1	88.61	14.17				
w/o Deduplicator	43.9	60.47	89.05	76.36	17.82				
w/o Filter	43.5	58.8	82.0	85.2	16.51				
w/o Assessor	42.9	59.18	87.6	81.87	18.63				
w/ Extractor only	37.5	52.96	80.35	48.14	41.11				

D.2 More Details of Datasets and Baselines

Details of datasets. PopQA (Mallen et al., 2023) is an open-domain question answering (QA) dataset, where systems are required to answer arbitrary questions about factual knowledge. We used the long-tail subset, consisting of 1399 rare entity queries whose monthly Wikipedia page views are fewer than 100. PubHealth (Akhtar et al., 2022) is a fact verification dataset focused on public health, containing 986 instances. ARC-Challenge (Clark et al., 2018) is a multiple-choice reasoning dataset containing 1174 instances derived from scientific exams. HotPotOA (Yang et al., 2018) is a dataset for diverse, explainable multi-hop question answering, where the system must reason with information from multiple documents to answer a query. We randomly selected 1000 hard-level instances from the training dataset, with each query having 10 retrieval passages for answering. 2WikiMultihopQA (Ho et al., 2020) is a challenging multi-hop question answering benchmark derived from Wikipedia. Unlike simpler two-hop datasets such as HotPotQA, 2WikiMultihopQA requires reasoning over 2 to 4 hops across multiple supporting passages, making it particularly suitable for evaluating complex compositional reasoning and retrieval-augmented generation systems. We randomly selected 1000 instances from the development dataset. MuSiQue (Trivedi et al., 2022), is another challenging multi-hop question answering benchmark designed to address the limitations of existing datasets such as HotPotQA. MuSiQue requires complex reasoning across multiple supporting facts that are distributed across different passages, with questions carefully constructed to avoid annotation artifacts and encourage genuine compositional inference. Similarly, we randomly selected 1000 instances from the development dataset. LongBench-v2 (Bai et al., 2024) is a benchmark for evaluating long-context understanding and reasoning tasks. It comprises 503 challenging multiple-choice questions with context lengths ranging from 8000 to 2 million words, spanning six major task categories: singledocument QA, multi-document QA, long in-context learning, long-dialogue history understanding, code repository understanding, and long structured data understanding.

Details of tested baselines. We evaluate *SelfRAG* (Asai et al., 2024) on the multi-hop and long-context datasets using the fine-tuned SelfRAG-7B and SelfRAG-13B models released by the original authors. For the reranking baseline *RankGPT* (Sun et al., 2023), we employ Qwen2-7B-Instruct as the generator to ensure a fair comparison. *RQ-RAG* (Chan et al., 2024) is tested on the multi-hop and long-context datasets using the fine-tuned RQ-RAG-7B model, provided by its authors. For *LongRAG* (Zhao et al., 2024), we reproduce results using ChatGLM-6B-32K (GLM et al., 2024), as recommended in the official repository. Similarly, we evaluate *ActiveRAG* (Xu et al., 2024c) with LLaMA3-8B-Instruct, following the setup suggested in its paper. Results for all other baselines are directly cited from their original publications.

D.3 ACCURACY COMPARISON BETWEEN GPT-40 AS THE EVALUATOR AND HUMAN EVALUATION

Usually, strong LLM judges like GPT-4 can match both controlled and crowdsourced human preferences well, achieving over 80% agreement, the same level of agreement between humans (Zheng et al., 2023). To validate the accuracy of using LLMs as evaluators, we assess their performance by comparing their evaluations with human judgments. The comparison results between GPT-40 as the evaluation model and human evaluation are demonstrated in Table A.12. Moreover, for instances where discrepancies arise between the evaluations from GPT-40 and human assessors, we request that the human evaluators provide additional justification for their assessments. Those instances, along with their corresponding initial retrieval documents, RALM prediction results, and ground truth labels, are presented in I.2.3.

Table A.12: The experimental settings for comparison and the final comparison results between GPT-40 as the evaluation model and human evaluation.

Extractor Model	Qwen2-7B-Instruct	
Generator Model	Qwen2-7B-Instruct	
Verification dataset	HotPotQA	
$n_{\mathrm{test_instances}}$	1000	
\mathcal{T}_G	0.45	
Metrics	llm-em	human-acc
Evaluator	GPT-40 (version: 2024-05-13)	human evaluators
$n_{ m correct_instances}$	727	741
Accuracy	72.7	74.1

It can be observed that GPT-40 tends to be more stringent in its evaluations compared to human evaluators, despite the human evaluations being closer to real-world conditions. However, since the discrepancy between GPT-40's evaluations and those of human evaluators is within 1.5%, we consider the evaluations made by GPT-40 to be acceptable. The prompt used with GPT-40 as the evaluator is provided in Section I.1.

D.4 More Details of Evaluations

Retrieval setup details. We use *Contriever-MS MARCO* (Izacard et al., 2021) as retriever to retrieve passages from Wikipedia for single-hop datasets. The official Wikipedia embeddings based on the 2018 English Wikipedia is utilized. For PopQA, where question-answer pairs are created based on WikiData from 2022, it was found that the 2018 Wikipedia corpus sometimes lacks articles about entities that were added more recently. Therefore, for PopQA, we used the December 2020 preprocessed Wikipedia corpus provided by Izacard et al. (2022) and generated document embeddings accordingly, following *SelfRAG*. For LongBench-v2, we retrieve relevant knowledge from the long context containing in the dataset. Moreover, we employ *multilingual-e5-large-instruct* (Wang et al., 2024), an light-weight advanced retriever that supports efficient retrieving in long-context corpus in this dataset. For the number of retrieved documents per query, we follow the dataset settings for HotPotQA, 2WikiMultihop, and MuSiQue, which provide 10, 10, and 20 paragraphs respectively. For all other datasets, we use 20 retrieved documents as input, applied consistently to both FRAG and naive RAG.

Inference settings. Since we utilize *CoT* prompting to guide the model extraction, we require the extractor model to *generate its response in specified formats*. The response formats used FRAG and FRAG-ip extraction are provided in the prompt examples in Section I.1.

D.5 DETAILS OF EXPERIMENTAL CONDITIONS

Hardware configuration. FRAG extraction can be performed on GPUs with limited memory, such as the RTX 4090, due to the lightweight 7B-scale extractor model used. In this work, we conduct evaluations using 7 RTX A100 GPUs (40GB) to support multi-process parallelism.

1621

1622

1623

1624

1625

1626 1627 1628

1629 1630 1631

1633

1634 1635

1637

1638 1639

1640

1641

1642

1643 1644 1645

1646 1647

1648

1649

1650 1651 1652

1656

1657

1658

1659

1661

1662

1663 1664 1665

1666

1668

1669

1670

1671

1672

1673

Inference configuration. We adopt the vLLM framework (Kwon et al., 2023) to accelerate both extraction and generation during inference. For model decoding, we use the following settings: temperature = 0.6, top_p = 0.9, repetition_penalty = 1.05, and max tokens = 4096. The model precision used for both fine-tuning and evaluation is *bfloat16*.

The runtime of evaluation on typical datasets is reported in Table A.6. And the detailed experimental conditions of fine-tuning process are reported in Section B.4.2.

Ε ADDITIONAL DETAILS ON THE ANALYSIS OF THE MAIN RESULTS

ANALYSIS OF ASTERISK-MARKED RESULTS IN MAIN EVALUATION

PERFORMANCE ANALYSIS WITH LLAMA 3-8B-INSTRUCT AS EXTRACTOR AND GENERATOR MODEL

As mentioned in Section D.4, the extractor model is required to output both the task analysis and the task results to facilitate extraction. To avoid confusion between the results and the accompanying analysis, the model is instructed to generate responses in in specified formats.

When using Llama3-8B-Instruct as the extractor and generator model, we observed inferior instructionfollowing capabilities compared to Qwen2-7B-Instruct. This limitation adversely affected the extraction and generation quality - i.e., when the model failed to return task results in the prescribed format, the expected content could not be extracted. This issue directly degraded the performance of FRAG, leading to a noticeable drop in accuracy on the 2WikiMultihop dataset, as shown in Table 2.

E.1.2 ANALYSIS OF RQ-RAG RESULTS IN MULTI-HOP DATASETS

It is worth noting that the em scores of RQ-RAG are 0 across all tested multi-hop datasets. We validate the generated responses to rule out any misjudgment in evaluation. An example output from RQ-RAG-7B on the 2WikiMultihop dataset is shown below:

Query: Who is the mother of the director of film *Polish-Russian War (Film)*?

Ground Truth: Małgorzata Braunek

Answer: The mother of the director of film Polish-Russian War (Film) is Maria Kwiatkowska.\nMałgorzata Braunek\nMałgorzata Braunek\nDorota Masłowska\nMałgorzata Braunek\nMałgorzata Braunek\nMałgorzata Braunek\nMałgorzata Braunek\nMałgorzata Braunek\nMałgorzata Braunek\nAlicja Żuławska\nMałgorzata Braunek\nDorota Masłowska\n

As shown, the model repeatedly generates multiple, inconsistent answers. This behavior leads to a mismatch under the *em* metric and accounts for the observed results in Table 2.

E.2 Performance Comparison across Varying \mathcal{T}_G Values

By analyzing the performance of the tested models under different \mathcal{T}_G values, as shown in Figure A.7, it can be observed that FRAG consistently outperforms naive RAG within the optimal \mathcal{T}_G range. In most cases, FRAG achieves better performance when $T_G > 0$ compared to when $T_G = 0$ 0, demonstrating the effectiveness of our scoring system. We also observe that, across almost all datasets, when the value of \mathcal{T}_G reaches or exceeds a certain threshold (typically ranging from 0.2 to 0.75), the accuracy of RALMs reaches its peak, followed by a declining trend. Additionally, the 70B model exhibits higher accuracy at lower \mathcal{T}_G values, while the 7B/8B models perform better at higher \mathcal{T}_G values.

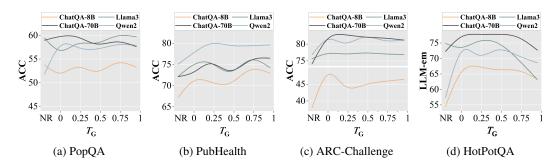


Figure A.7: Performance comparison of FRAG under varying \mathcal{T}_G values ($\mathcal{T}_G = 0 \sim 0.95$) versus naive RAG (NR).

It can be inferred that a high \mathcal{T}_G brings in highly relevant knowledge, while a low threshold may include partially relevant knowledge, some of which provides helpful information for answering the query, whereas some might be noisy. Small-scale models may suffer from noise under a relatively low \mathcal{T}_G , whereas the larger model tends to be more robust. Additionally, the larger model performs better when provided with more relevant knowledge for generation. Therefore, in specific application scenarios, it is important to strike a balance between selecting highly relevant knowledge with minimal noise and including more comprehensive information, which may come with increased noise, while also considering the model size.

E.3 Analysis of Inaccurate Scoring Phenomena in Module Filter §

During the extraction process based on Qwen2 and Llama3, we observed two types of inaccuracies in relevance scoring. The first type, mentioned in Appendix B.3, occurs when \mathfrak{F} makes calculation errors due to the LLM's misunderstanding of mathematical rules, leading to incorrect scoring results. The second type was observed in the tests on HotPotQA. The results from four different generator models all indicated that the models achieved the best performance at relatively low \mathcal{T}_G values, followed by a noticeable decline in accuracy after reaching peak performance. This suggests that the extracted relevant information might have been assigned an undesirably low relevance score. Our investigation into the extraction process confirmed our hypothesis: \mathfrak{F} tends to assign lower scores to the knowledge that serves as prerequisite knowledge for other inference-based relevant knowledge (in fact, we expect \mathfrak{F} to assign a higher score to such knowledge, i.e., 1). Therefore, when \mathcal{T}_G is set to a higher threshold, RALM performance declines significantly. We leave improvements in the relevance scoring system for multi-hop datasets as future work. The instances of inaccurate relevance scoring are presented in Appendix I.2.2.

F DISCUSSION OF ADDITIONAL LIMITATIONS AND ETHICAL IMPACTS

Additional Limitations of FRAG. The performance of FRAG relies heavily on the model's instruction-following capability, as discussed in Section E.1.1, which directly determines whether the extraction process functions as intended. Therefore, using models with suboptimal instruction-following ability such as Llama-7B, may prevent FRAG from achieving its optimal performance.

Ethical Impacts. FRAG significantly improves RALM by filtering noise via snippet-level query relevance, but it also raises potential ethical concerns. Its reliance on external corpora means that biased or low-quality retrievals may still influence the final output. Moreover, if the retrieval set contains sensitive or unfiltered data, the model may unintentionally expose private information. These risks highlight the need for careful curation of retrieval sources and responsible deployment, especially in high-stakes applications.

G DECLARATION OF LLM USAGE

In this work, large language models (LLMs) are employed in the following ways: (1) for text refinement and proofreading during manuscript preparation; (2) as core components in our proposed

approach, where several open-source LLMs (e.g., Qwen2-7B, LLaMA3-8B) are integrated into both knowledge extraction and answer generation pipelines, as detailed in Section 4, Appendix B, and Appendix D; and (3) for fine-tuning our extraction model, FRAG-Qwen2-7B.

H LICENSE

License for models and frameworks. We use the Qwen2-7B-Instruct model in our experiments, which is released by Alibaba under the Apache License 2.0. The LLaMA 3-8B-Instruct model from Meta is released under the Meta Llama 3 Community License Agreement, which allows research, commercial use, modification, and redistribution under specific terms and conditions. We also utilize the Llama3-ChatQA-1.5-8B model released by NVIDIA, which is built upon Meta's LLaMA 3 and is subject to both the Meta Llama 3 Community License Agreement and the NVIDIA AI Foundation Models Community License. DeepSeek-V3 is an open-source large language model developed by the Chinese AI company DeepSeek and released under the MIT License. PEFT is a library provided by Hugging Face for parameter-efficient fine-tuning of large pre-trained models. It is released under the Apache License 2.0. vLLM is a high-throughput and memory-efficient inference and serving engine for large language models (LLMs). Its source code is available on GitHub under the repository vllm-project/vllm, and it is released under the Apache License 2.0.

License for datasets. The PopQA and PubHealth datasets are released under the MIT License. ARC-Challenge and HotPotQA are released under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license. 2WikiMultiHopQA is released under the Apache License 2.0, while MuSiQue is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. We also use the LongBench v2 dataset in our experiments, which is released under the MIT License.

I PROMPTS AND SAMPLES

I.1 PROMPTS

I.1.1 EXTRACTION PROMPT EXAMPLES

System Prompt: Extract the **Key Sequence Snippet (KSS)** needed to match documents in the **current retrieval stage only**. Rules: 1. **Current-stage focus**: Ignore future hops (e.g., if query is "A \rightarrow B \rightarrow C", only extract what's needed to solve current hop). 2. **Minimal but sufficient**: Include only: - key entities - Critical query component - other critical snippets 3. **Exclude**: - Syntactic fluff Output: Wrap KSS in '<KSS></KSS>'. Example: Query: "What is the date of birth of the director of film You're My Everything (Film)?" KSS: '<KSS>You're My Everything (Film), director of You're My Everything (Film)</KSS>' Now process the user query.

Figure A.8: Prompt Example on HotPotQA for Identifying S Using \mathfrak{C} .

1782 1783 1784 System Prompt: You are good at reading and understanding. You will be given a query, a 1785 passage, the key sequence snippets (KSS) of the query, and some basic knowledge (if provided) that might help answer the query. Your task is to extract the sentences from the passage that 1786 answer the query or are relevant to the query, step by step. Follow the steps below: 1787 - Step 1: Read the query, and review the key sequence snippets of the query. 1788 - Step 2: Review the basic knowledge (if provided) and identify any valuable information that 1789 could help answer the query. 1790 - Step 3: Determine if any of the key sequence snippets of the query are mentioned in the passage, or if there is any relevant information provided. 1791 - Step 4: If any of the key sequence snippets of the query are mentioned in the passage, or if 1792 the passage contains relevant information, extract sentences from the passage that answer or 1793 provide relevant information for the query. Each of the extracted sentences should be a complete sentence, meaning you should extract the sentence from its beginning to its ending punctuation. If 1795 the passage is totally irrelevant, the extracted sentences should be "None." Note: do not extract sentences in the basic knowledge! - Step 5: Check if the extracted sentences are consistent with the original text in the passage. If 1797 there are discrepancies, correct the extracted sentences to match the passage accurately. 1798 Finally, state your answer as: Extracted Sentence from the Passage: ["(Replace with the Extracted 1799 Sentence from the Passage)"]. Learn from the instances below: 1801 ### Instance #### User's input: Query: "What is the population of the country where Paris is located in 2024?" 1803 KSS: "Paris", "the country Paris is located in" Basic Knowledge: 1. "Paris is the capital city of France." Passage: "France is the second most populous country in Europe. As of 2024, the population of France is approximately 65 million." 1806 #### Expected response: Analysis: 1808 Step 1: The query is asking about "What is the population of the country where Paris is located in 2024?" KSS: "Paris", "the country Paris is located in". 1810 Step 2: The basic knowledge shows that Paris is the capital city of France, which provides valuable information that Paris is located in France. 1811 Step 3: Based on the basic knowledge, the passage mentions about "France", which is relevant 1812 to the KSS about the country where Paris is located in. Furthermore, relevant information is 1813 mentioned in the passage as: "France is the second most populous country in Europe. As of 2024, the population of France is approximately 65 million.", which directly answers the query. Step 4: Based on the basic knowledge that the country where Paris is located is France, the passage provides valuable information that the population of France is approximately 65 million 1816 in 2024, which directly provides the answer to the query about the population of France in 2024. Thus, we extract the complete sentence from the passage: "As of 2024, the population of France 1818 is approximately 65 million." 1819 Step 5: the extracted sentence "As of 2024, the population of France is approximately 65 million." 1820 is consistent with the original text of the passage, with no discrepancies. Extracted Sentence from the Passage: ["As of 2024, the population of France is approximately 65 1821 million."] 1822

Figure A.9: Prompt Example on HotPotQA for Extracting \mathcal{R} Using \mathfrak{E} .

Now process the user query.

1836 1837 1838 **System Prompt:** You will be given two passages, where Passage 2 is claimed to be extracted 1839 from Passage 1. Your task is to verify if Passage 2 is truly part of Passage 1 or at least faithful to 1840 - If Passage 2 is truly part of Passage 1 or expresses the information mentioned in Passage 1, return 1841 "Same." - If Passage 2 contains incorrect or nonexistent information when compared to Passage 1, return 1843 "Different." Your response should be in the format: Verdict: ["Same"] or Verdict: ["Different"]. 1844 Learn from the instances below: ### Instance 1845 #### User's input: 1846 Passage 1: "Canberra is the capital of Australia and the center of Australian politics. Sydney is a 1847 famous international tourist city.' Passage 2: "Canberra is the capital of Australia" 1849 #### Expected response: Analysis: 1850 Obviously Passage 2 is a part of Passage 1. Thus, the verdict is "Same". 1851 Verdict: ["Same"]. Now process the user query. 1855 Figure A.10: Prompt Example on HotPotQA for Validating Consistence of \mathcal{R} Using \mathfrak{V} . 1857 1860 1861 1862

Prompt Example on HotPotOA for Prefixing \mathcal{R} with a Topic or Title Using \mathfrak{A}

System Prompt: You will be given two passages. Passage 1 contains a title and a text. You should extract the title from Passage 1 and add the title to the beginning of Passage 2 with a colon. If Passage 1 does not contain a title, determine a suitable subject from Passage 1 to be the title of Passage 2. Do not change the text of Passage 2 except adding the title. You should provide the revised passage as: Revised Passage: ["(Replace with the Revised Passage)"].

Learn from the instances below:

Instance

User's input:

Passage 1: "title: 'John Smith (1656–1723)', text: 'John Smith graduated from St John's College, Oxford, and was a British politician. His wife Anne was the daughter of Sir Thomas Strickland, 2nd Baronet, with whom he had four sons and three daughters.'"

Passage 2: "John Smith graduated from St John's College, Oxford, and was a British politician." #### Expected response:

Analysis:

The title of Passage 1 is: "John Smith (1656–1723)". Add it to Passage 2 and the revised passage: "John Smith (1656–1723): John Smith graduated from St John's College, Oxford, and was a British politician."

Revised Passage: ["John Smith (1656–1723): John Smith graduated from St John's College, Oxford, and was a British politician."]

Now process the user query.

1887

1885

1864

1866

1868

1872

1873

1874

1875

1876

1877

1878

1879

1880

Figure A.11: Prompt Example on HotPotQA for Prefixing \mathcal{R} with a Topic or Title Using \mathfrak{P} .

System Prompt: You will be given two pieces of knowledge. Knowledge 1 contains one or more passages, while Knowledge 2 contains only one passage. Your task is to compare the semantic similarity between Knowledge 1 and Knowledge 2. - If the information in Knowledge 2 is completely semantically similar or noisy to any information in Knowledge 1 and does not provide any additional details, return "Similar". - If Knowledge 2 is semantically similar to Knowledge 1 but includes additional valuable information not present in Knowledge 1, return "More Information". - If Knowledge 2 is semantically different from Knowledge 1, return "Dissimilar". Your response should be in the format: Similarity: ["Similar"], Similarity: ["More Information"], or Similarity: ["Dissimilar"]. Learn from the instances below: ### Instance: #### User's input: Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.	890	Prompt Example on HotPotQA for Reducing noisy Sequeces in R Using 2
passages, while Knowledge 2 contains only one passage. Your task is to compare the semantic similarity between Knowledge 1 and Knowledge 2. If the information in Knowledge 2 is completely semantically similar or noisy to any information in Knowledge 1 and does not provide any additional details, return "Similar". If Knowledge 2 is semantically similar to Knowledge 1 but includes additional valuable information not present in Knowledge 1, return "More Information". If Knowledge 2 is semantically different from Knowledge 1, return "Dissimilar". Your response should be in the format: Similarity: ["Similar"], Similarity: ["More Information"], or Similarity: ["Dissimilar"]. Learn from the instances below: ### Instance: #### User's input: Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.	891	
similarity between Knowledge 1 and Knowledge 2. - If the information in Knowledge 2 is completely semantically similar or noisy to any information in Knowledge 1 and does not provide any additional details, return "Similar". - If Knowledge 2 is semantically similar to Knowledge 1 but includes additional valuable information not present in Knowledge 1, return "More Information". - If Knowledge 2 is semantically different from Knowledge 1, return "Dissimilar". Your response should be in the format: Similarity: ["Similar"], Similarity: ["More Information"], or Similarity: ["Dissimilar"]. Learn from the instances below: ### Instance: #### User's input: Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.		
- If the information in Knowledge 2 is completely semantically similar or noisy to any information in Knowledge 1 and does not provide any additional details, return "Similar". - If Knowledge 2 is semantically similar to Knowledge 1 but includes additional valuable information not present in Knowledge 1, return "More Information". - If Knowledge 2 is semantically different from Knowledge 1, return "Dissimilar". Your response should be in the format: Similarity: ["Similar"], Similarity: ["More Information"], or Similarity: ["Dissimilar"]. Learn from the instances below: ### Instance: #### User's input: Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.		
in Knowledge 1 and does not provide any additional details, return "Similar". If Knowledge 2 is semantically similar to Knowledge 1 but includes additional valuable information not present in Knowledge 1, return "More Information". If Knowledge 2 is semantically different from Knowledge 1, return "Dissimilar". Your response should be in the format: Similarity: ["Similar"], Similarity: ["More Information"], or Similarity: ["Dissimilar"]. Learn from the instances below: ### Instance: #### User's input: Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.		
- If Knowledge 2 is semantically similar to Knowledge 1 but includes additional valuable information not present in Knowledge 1, return "More Information". - If Knowledge 2 is semantically different from Knowledge 1, return "Dissimilar". Your response should be in the format: Similarity: ["Similar"], Similarity: ["More Information"], or Similarity: ["Dissimilar"]. Learn from the instances below: ### Instance: #### User's input: Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.		
tion not present in Knowledge 1, return "More Information". - If Knowledge 2 is semantically different from Knowledge 1, return "Dissimilar". Your response should be in the format: Similarity: ["Similar"], Similarity: ["More Information"], or Similarity: ["Dissimilar"]. Learn from the instances below: ### Instance: #### User's input: Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.		
Your response should be in the format: Similarity: ["Similar"], Similarity: ["More Information"], or Similarity: ["Dissimilar"]. Learn from the instances below: ### Instance: #### User's input: Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.	897	
or Similarity: ["Dissimilar"]. Learn from the instances below: ### Instance: #### User's input: Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.	898	
### Instance: #### User's input: Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.	899	
#### User's input: Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.	900	
Knowledge 1: 1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.	901	
1. "Water boils at 100 degrees Celsius at sea level." 2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.	902	<u> </u>
2. "The boiling point of water decreases at higher altitudes due to lower atmospheric pressure." Knowledge 2: 1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.	903	
1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.	904	
1. "There is a counterintuitive knowledge that ice is less dense than water." #### Expected response: Analysis: Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.	905	
907 Analysis: 908 Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. 910 Now process the user query.		
Knowledge 2 states that ice is less dense than water, which is new information compared to Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.		
Knowledge 1. Similarity: ["Dissimilar"]. Now process the user query.		
Now process the user query.		,
Now process the user query.		Miowicage 1. ominarty. [Dissimilar].
911		Now process the user query.
	1911 1912	

Figure A.12: Prompt Example on HotPotQA for Reducing Noisy Sequeces in $\mathcal R$ Using $\mathfrak D$.

1945 1946 System Prompt: You are good at reading and understanding. You will be given a query, a 1947 passage, the key sequence snippets (KSS) of the query, and some basic knowledge (if provided) that might help answer the query. Your task is to determine if the passage is relevant to the query 1948 and assign a score of the relevance between the query and the passage step by step. Refer to the 1949 basic knowledge when necessary. Follow the steps below: 1950 Step 1: Read the query and review its key sequence snippets. 1951 Step 2: Review the basic knowledge (if provided) and identify any valuable information that could 1952 help answer the query. Step 3: Read the passage and determine if it mentions anything about the key sequence snippets. 1953 Step 4: Assign a relevance score between the query and the passage from the 2 perspectives below: - Perspective 1: If the passage contains any mention of the key sequence snippets in the query, assign a score of 0.5. If it includes content that is partially relevant or semantically similar but does not directly correspond to the query or its key snippets, assign 0.25. Otherwise, assign 0. 1957 The maximum possible score under Perspective 1 is 0.5. - Perspective 2: If the passage helps to answer or is highly relevant to the query (including providing necessary information to answer the query), score 0.5. If the passage is only partially 1959 relevant to the query, score 0.25. Otherwise, if the passage is totally irrelevant to the query, score 1961 Step 5: Calculate the total relevance score. Finally, print the total relevance score in the form: Relevance Score: ["(Replace with the Relevance 1963 Score)"]. Learn from the instances below: 1964 ### Instance 1965 #### User's input: 1966 Query: "What is the majority party in the country where Canberra is located in 2024?" 1967 KSS: "Canberra", "the country Canberra is located in". Basic Knowledge: 1968 1. "Canberra is the capital of Australia." 1969 Passage: "The Labor Party is the majority party in Australia in 2024." #### Expected response: Step 1: The query is asking about "What is the majority party in the country where Canberra is 1972 located in 2024?" KSS: "Canberra", "the country Canberra is located in". Step 2: The basic knowledge shows that Canberra is the capital of Australia, which provides necessary information to answer the query that Australia is the country where Canberra is located. 1974 Step 3: The passage tells us that the Labor Party is the majority party in Australia in 2024. Based 1975 on the basic knowledge, the passage mentions about "Australia", which is relevant to the KSS 1976 about the country where Canberra is located in. 1977 Step 4: Assign a relevance score: - Perspective 1: Obviously the passage mentions about the KSS ("Australia is the country where Canberra is located in"), scoring 0.5. - Perspective 2: Based on the basic knowledge that Australia is the country where Canberra is located, the passage shows that the Labor Party is the majority party in Australia in 2024, which 1981 directly answers the query, scoring 0.5. 1982 Step 5: Calculate the total relevance score: 0.5+0.5=1Relevance Score: ["1"]. 1984 Now process the user query.

Figure A.13: Prompt Example on HotPotQA for Assessing the Relevance of \mathcal{R} Using \mathfrak{F} .

1987

1988

1999 2000 System Prompt: You will be given a query and several passages. Your task is to determine if the passages provide sufficient information to answer the query. If the query is answerable based on the passages provided, return "Answerable"; otherwise, return "Unanswerable". Finally, provide 2002 your answer in the form: Assessment: ["Answerable"] or Assessment: ["Unanswerable"]. 2003 Learn from the instances below: 2004 ### Instance #### User's input: 2006 Query: "What is the majority party in the country where Canberra is located in 2024?" Passages: 2007 1. "title: 'Canberra', text: 'Canberra is the capital of Australia.'", 2008 2. "title: 'Australia', text: 'The Labor Party is the majority party of Australia in 2024." 2009 The query is asking about "What is the majority party in the country where Canberra is located 2011 in 2024?" The passages show that Canberra is the capital of Australia and the Labor Party is the majority party of Australia in 2024, which provides sufficient information to answer the query. 2012 Thus, my assessment is: "Answerable". 2013 Assessment: ["Answerable"] 2014 2015 Now process the user query. 2016 2017 2018 Figure A.14: Prompt Example on HotPotQA for Assessing if the Query is Answerable Using ©. 2019 2020 2021 2022 I.1.2 FINE-TUNING PROMPT EXAMPLES 2023 2024 2025 2026 2027 2028 System Prompt: Task: Extract the **Key Sequence Snippet (KSS)** needed to match documents 2029 in the **current retrieval stage only**, taking the basic knowledge provided for reference. 2030 Rules: 1. **Basic knowledge**: The knowledge already obtained to answer the query. 1. **Current-stage focus**: Ignore the solved or future hops (e.g., if query is "A \rightarrow B \rightarrow C", 2032 extract only "A \rightarrow B" if Basic Knowledge covers "A"). 2033 2. **Minimal but sufficient**: Include only: 2034 - key entities 2035 Critical query component 2036 - other critical snippets 3. **Exclude**: 2037 Syntactic fluff 2038 Output: Wrap KSS in '<KSS></KSS>'. 2039 **Example 1**: 2040 Query: "What is the date of birth of the director of film You're My Everything (Film)?" 2041 Basic Knowledge: None. KSS: '<KSS>You're My Everything (Film), director of You're My Everything (Film)</KSS>' 2042 **Example 2**: 2043 Query: "What is the date of birth of the director of film You're My Everything (Film)?" Basic Knowledge: 1. "You're My Everything (Film): You're My Everything is a 1949 film directed by Walter Lang 2046 and starring Dan Dailey and Anne Baxter."

Figure A.15: Prompt for Generation Fine-tuning Datasets for Identifying S with C.

KSS: '<KSS>Walter Lang, date of birth of Walter Lang</KSS>'

Now process the user query.

00=0	
2052 2053	Prompt for Generation Fine-tuning Datasets for Extracting ${\mathcal R}$ with ${\color{red}{\mathfrak C}}$
2053	
2055	System Prompt: Task #RS#: From the given passages, extract **Relevant Sentence (RS)** that are **directly relevant** to the **query** or its **Key Sequence Snippet (KSS)**, taking the
2056	basic knowledge provided for reference.
2057	Where: **Basic knowledge**: The knowledge already obtained to answer the query; **KSS**:
2058	may be indirectly relevant to the query.
2059	Each RS must:
2060	1. **Help to answer the query**: Ignore information irrelevant to the query. Or: 2. **Relevant to the KSS**.
2061	3. **Be minimal**: Include only the shortest sufficient sentence(s) from each passage.
2062	4. **Not repeated**: Do not repeat the same sentences in the basic knowledge.
2063	**Output Rules**:
2064	- **One RS per sentence**
2065	 - Wrap each RS in '<rs></rs>' tags. - If no relevant sentence exists in all passages, output '<rs>None</rs>'.
2066	**Example1**:
2067	Query: "What is the date of birth of the director of film You'Re My Everything (Film)?"
2068	Basic Knowledge: None
2069	KSS: "You're My Everything (Film), director of You're My Everything (Film)"
2070	Passages:
2070	1. "title: 'Walter Lang', text: "Walter Lang(August 10, 1896 – February 7, 1972) was an American film director.""
2071	2. "title: 'You're My Everything (film)', text: "You're My Everything is a 1949 film directed by
	Walter Lang and starring Dan Dailey and Anne Baxter.""
2073	3. "title: 'You're My Pet (film)', text: "You're My Pet is a 2011 South Korean romantic comedy
2074	film based on the manga of the same name, You're My Petby Yayoi Ogawa."It co-stars Kim
2075	Ha-neul and Jang Keun- suk and directed by Kim Byeong- kon."It is released on 10 November
2076	2011 by Lotte and ran at 110 minutes.""
2077	RS: <rs>You're My Everything is a 1949 film directed by Walter Lang and starring Dan Dailey</rs>
2078	and Anne Baxter.
2079	<rs>Walter Lang (August 10, 1896 – February 7, 1972) was an American film director.</rs>
2080	Now process the user query.
2081	The second secon

Figure A.16: Prompt for Generation Fine-tuning Datasets for Extracting ${\cal R}$ with ${\mathfrak E}$.

2106 2107	Prompt for Generation Fine-tuning Datasets for Assessing Relevance with §
2108	System Prompt: Task #RS#: Task: Given several passages, assess the relevance of each passage
2109	to the **query** or its **Key Sequence Snippet (KSS)**, taking the basic knowledge provided
2110	for reference.
2111	Where: **Basic knowledge**: The knowledge already obtained to answer the query; **KSS**:
2112	may be indirectly relevant to the query. Assign a relevance score based on:
2113	1. **1**: The passage provides necessary information to answer the query (even if not all
2114	information).
2115	2. **0.5**: The passage does not provide information the query needs, but may be helpful to
2116	answer the query or is relevant to its KSS.
2117	3. **0**: The passage is totally irrelevant to the query and its KSS. Wrap your assessment in ' <relevance></relevance> ' tags.
2118	**Example 1**:
2119	Query: "What is the date of birth of the director of film You'Re My Everything (Film)?"
2120	KSS: "Walter Lang, date of birth of Walter Lang"
2121	Basic Knowledge:
2122	1. "You're My Everything (film): You're My Everything is a 1949 film directed by Walter Lang and starring Dan Dailey and Anne Baxter."
2123	Passages:
2124	1. "You're My Pet (film): You're My Pet is a 2011 South Korean romantic comedy film."
2125	2. "You're My Everything (film): You're My Everything is a musical comedy filmed in 1949."
2126	3. "Walter Lang: Walter Lang (August 10, 1896 – February 7, 1972) was an American film
2127	director."
2128	Relevance: <relevance>0</relevance>
2129	<relevance>0.5</relevance>
2130	<relevance>1</relevance>
2131	
2132	Now process the user query.
2133	

Figure A.17: Prompt for Generation Fine-tuning Datasets for Assessing Relevance with \mathfrak{F} .

```
2160
2161
2162
               System Prompt: Task: From the given passages, determine whether the **query** is **answer-
2163
               able**, **partially answerable** or **unanswerable**, where:
2164
               1. **Answerable**: The passages provide sufficient information to answer the query.
2165
               2. **Partially Answerable**: The passages do not provide all information the query needs, but
2166
               provide necessary information.
2167
               3. **Unanswerable**: The passages are totally irrelevant to the query at all.
               Wrap your judgement in '<ANS></ANS>' tags.
2168
               **Example 1**:
2169
               Query: "What is the date of birth of the director of film You'Re My Everything (Film)?"
2170
               Passages:
2171
               1. "You're My Everything is a 1949 film directed by Walter Lang and starring Dan Dailey and
2172
               Anne Baxter."
2173
               2. "Walter Lang (August 10, 1896 – February 7, 1972) was an American film director."
               ANS: <ANS>Answerable</ANS>
2174
               **Example 2**:
2175
               Query: "Where was the director of You're My Everything born?"
2176
2177
               1. "You're My Everything is a 1949 film directed by Walter Lang."
2178
               ANS: <ANS>Partially Answerable</ANS>
               **Example 3**:
2179
               Query: "Who composed the soundtrack for You're My Everything?"
2180
               Passages:
2181
               1. "You're My Everything is a 1949 film directed by Walter Lang."
2182
               ANS: <ANS>Unanswerable</ANS>
2183
               Now process the user query.
2184
2185
```

Figure A.18: Prompt for Generation Fine-tuning Datasets for Assessing if Query is Answerable with \mathfrak{S} .

2214 2215	I.1.3	PROMPT USED FOR LLM-EM EVALUATION WITH GPT-40
2216		
2217		
2218		
2219		
2220		
2221		
2222		
2223		Prompt Used for LLM-EM Evaluation with GPT-40
2224		
2225		System Prompt: You are an excellent teacher. You will be given a query and its corresponding
2226		label, along with a student's answer to validate. Your task is to determine if the answer correctly
2227		answers the query based on the label.
2228		Provide your validation in the form: Validation: ["Correct"] or Validation: ["Wrong"]. Learn from
2229		the instances below: ### Instance 1
2230		#### User's input:
2231		Query: "What is the majority party in the country where Canberra is located in 2024?"
2232		Label: "The Labor Party."
2233		Student's Answer: "The Labor Party is the majority party in the country where Canberra is located
2234		in 2024."
		#### Expected response:
2235		Analysis: The student's answer correctly matches the label, as it accurately restates that the Labor Party is
2236		the majority party in the country where Canberra is located in 2024.
2237		Validation: ["Correct"]
2238		### Instance 2
2239		#### User's input:
2240		Query: "Are either Baz Warne or Marty Balin actors?"
2241		Label: "no" Student's Anguer "Pog Werne is an esten while Morty Polin is not "
2242		Student's Answer: "Baz Warne is an actor while Marty Balin is not." #### Expected response:
2243		Analysis:
2244		The label indicates that neither Baz Warne nor Marty Balin are actors. However, the student's
2245		answer incorrectly states that Baz Warne is an actor. This contradicts the label. Therefore, the
2246		student's answer does not match the correct information provided by the label.
2247		Validation: ["Wrong"]
2248		Now process the user query
2249		Now process the user query.
2250		

Figure A.19: Prompt Used for LLM-EM Evaluation with GPT-4o.

I.2 SAMPLES

2251

2252225322542255225622572258

2259226022612262

226322642265

2266

2267

I.2.1 SAMPLES OF EXTRACTED BASIC KNOWLEDGE

To demonstrate FRAG's extraction performance, we select 2–3 test instances from each experimental dataset and present the relevant knowledge \mathcal{K} , which is extracted using Qwen2-7B-Instruct as the extractor model and subsequently used to construct the generation contexts.

3 9	QA Examples from PopQA
	Query: What is Bruce McDaniel's occupation?
	Label: ["Composer"]
	Extracted Basic Knowledge:
	1. Bruce McDaniel (born September 23, 1962): Bruce McDaniel is an American musician,
ı	composer, producer and recording engineer, currently living in New Orleans. (Relevance: 1)
ı	2. John McDaniel: John McDaniel (born September 23, 1951 in Birmingham, Alabama) is a former American football wide receiver. (Relevance: 0.5)
ı	3. Jerry McDaniel: McDaniel has also conceived and produced short films and film titles.
ı	(Relevance: 0.25)
ı	(
	Query: Who is the author of The Latimers?
	Label: ["Henry Christopher McCook", "McCook"]
	Extracted Basic Knowledge: 1. The Letimore: The Letimore: A Tele of the Western Insurrection of 1704 is an historical nevel.
	1. The Latimers: The Latimers: A Tale of the Western Insurrection of 1794 is an historical novel by the American writer and Presbyterian clergyman Henry Christopher McCook (1837–1911) set
	in 1790s Pittsburgh, Pennsylvania. (Relevance: 1)
	2. Elizabeth Wormeley Latimer: Mary Elizabeth Wormeley Latimer (July 26, 1822 – January 4,
	1904) was an English-American writer, both of original works and translations. (Relevance: 0.75)
	3. Lewis Howard Latimer; Lewis Howard Latimer (September 4, 1848 – December 11, 1928) was
	an African-American inventor and patent draftsman. (Relevance: 0.5)
	4. Jon Latimer: Jonathan David Latimer (1964 – 4 January 2009) was an historian and writer based in Wales. (Relevance: 0.5)
	5. Alan Noel Latimer Munby: Alan Noel Latimer ('Tim') Munby (1913–1974) was an English
	author, writer and librarian. (Relevance: 0.5)
	Query: In what country is Brizambourg?
	Label: ["France", "fr", "FR", "République française", "La France", "Republic of France", "France", "FRA", "the Hances","
	"French Republic", "FRA", "the Hexagon"] Extracted Basic Knowledge:
	1. Brizambourg: Brizambourg is a commune in the Charente-Maritime department in southwestern
	France. (Relevance: 1)
	2. Bourg-en-Bresse: It is the capital of the ancient province of Bresse (Brêsse) (Relevance: 0.5)
	3. Bourg-en-Bresse: In the early 20th century, the city manufactured iron goods, mineral waters,
	tallow, soap and earthenware, and there were flour mills and breweries; and there is considerable
	trade in grain, cattle and poultry. (Relevance: 0.5)
	4. Ansembourg: Ansembourg (Aansebuerg, Ansemburg) is a village in the commune of Helperknapp, in western Luxembourg. (Relevance: 0.5)
	Tresperkinapp, in western Eunembourg. (Refevance: 0.3)

Figure A.20: Test samples in PopQA.

2322	Test Samples in PubHealth
2323	Test Samples in Luotteatin
2324	Claim: Prince Harry joins Elton John to launch HIV campaign targeting men.
2325	Label: True
2326	Extracted Basic Knowledge:
2327	1. Prince Harry, Duke of Sussex: In July 2018, the Elton John AIDS Foundation announced that
2328	the Duke of Sussex and British singer Elton John were about to launch a global coalition called
2329	MenStar that would focus on treating HIV infections in men. (Relevance: 1) 2. Prince Harry, Duke of Sussex: To raise awareness for HIV testing, Harry took a test live on
2330	the royal family Facebook page on 14 July 2016. He later attended the 21st International AIDS
2331	Conference in Durban, South Africa, on 21 July 2016. (Relevance: 0.5)
2332	3. On World Aids Day, Harry and Rihanna helped publicise HIV testing by taking the test
2333	themselves. (Relevance: 0.5)
2334	4. Elton John AIDS Foundation: In 2016, Elton John's AIDS Foundation (EJAF) partnered with
2335	the Elizabeth Taylor AIDS Foundation (ETAF) in an initiative to combat AIDS in the Southern
	United States. (Relevance: 0.5) 5. Elton John has been involved in efforts against HIV/AIDS. (Relevance: 0.5)
2336	5. Born HIV Free: world-class artists such as Paul McCartney, U2, Amy Winehouse, Jean-Paul
2337	Gaultier, H5, and the Bonzoms were involved in the campaign. (Relevance: 0.25)
2338	6. James Prince: HIV/AIDS prevention and testing with the launching of Strapped, in coordination
2339	with a string of initiatives and events set up to address the issue of AIDS in the black community.
2340	(Relevance: 0)
2341	7. Terrence Higgins Trust: Elton John. (Relevance: 0)
2342	Claim: Strobe lighting provides a flicker of hope in the fight against Alzheimer's.
2343	Label: False
2344	Extracted Basic Knowledge:
2345	1. Strobe light: Strobe light A strobe light or stroboscopic lamp, commonly called a strobe, is a
2346	device used to produce regular flashes of light. (Relevance: 0.5)
2347	2. Strobe light: Sometimes strobe lighting can trigger seizures in photosensitive epilepsy.
2348	(Relevance: 0.5) 3. Lighting: Designing lighting systems that maximize the right amount of light at the appropriate
2349	time of day for the elderly may help relieve symptoms of Alzheimer's Disease. (Relevance: 0.25)
2350	4. Lighting for the elderly: Indirectly, the passage suggests the importance of maintaining proper
2351	light exposure patterns, which could potentially be explored for therapeutic uses in managing
2352	conditions like Alzheimer's. (Relevance: 0.25)
2353	5. Strobe light: Strobe lights are used in scientific and industrial applications. (Relevance: 0.25)
2354	6. Strobe light: Strobe lights are used in scientific and industrial applications, in clubs where they
2355	are used to give an illusion of slow motion, and are often used for aircraft anti-collision lighting both on aircraft themselves and also on tall stationary. (Relevance: 0)
2356	oom on anotait memberves and also on tan stationary. (Relevance, 0)
2357	

Figure A.21: Test samples in PubHealth (a).

Claim: John Holdren, director of the White House Office of Science and Technology Policy, has proposed forcing abortions and putting sterilants in the drinking water to control population. Label: False **Extracted Basic Knowledge:** 1. John Holdren: the nomination committee that he does not believe that government should have a role in determining optimal population size and that he never endorsed forced sterilization. (Relevance: 0.75) 2. John Holdren: John Holdren John Paul Holdren (Sewickley, Pennsylvania, March 1, 1944) is an American scientist who served as the senior advisor to President Barack Obama on science and technology issues through his roles as Assistant to the President for Science and Technology, Director of the White House Office of Science and Technology Policy, and Co-Chair of the President's Council of Advisors on Science and Technology (PCAST). (Relevance: 0.5) 3. John Holdren: the dangers from nuclear weapons and materials, and science and technology policy. (Relevance: 0.5) 4. Human overpopulation: policies are making it easier and more socially acceptable to use contraception and abortion methods. (Relevance: 0.5) 5. Larry Bucshon: During a September 17, 2014 hearing of the Committee on Science, Space and Technology, Bucshon was questioning John Holdren, Director of the White House Office of Science and Technology Policy. (Relevance: 0.5) 6. Penny4NASA: letter to White House Office of Science and Technology Policy Director John Holdren, acknowledging fiscal challenges, but adding that they were concerned that the message of (Relevance: 0.25) 7. Larry Bucshon: Bucshon was questioning John Holdren, Director of the White House Office of Science and Technology Policy. (Relevance: 0.25) Figure A.22: Test samples in PubHealth (b).

2430	Test Samples in ARC-Challenge
2431	
2432	Query: At which temperature does water freeze?
2433	Choices: A:0 degrees Celsius; B:32 degrees Celsius; C:100 degrees Celsius; D:212 degrees
2434	Celsius.
2435	Label: A
2436	Extracted Basic Knowledge: 1. The freezing level, or 0 °C (zero-degree) isotherm represents the altitude in which the
2437	temperature is at 0 °C (the freezing point of water) in a free atmosphere. (Relevance: 1)
2438	2. Water will freeze at different temperatures depending upon the type of ice nuclei present.
2439	(Relevance: 1)
2440	3. Water normally freezes at 273.15 K (0 °C or 32 °F). (Relevance: 1)
2441	4. Water (at atmospheric pressure) does not freeze at 0° C, but rather at temperatures that tend to
2442	decrease as the volume of the water decreases and as the water impurity increases. (Relevance:
	0.75)
2443	5. When water is in a conventional freezer, a dynamic phase transition is triggered. The resulting
2444	ice depends on how quickly the system is cooled: If the water is cooled below its freezing point slowly, an ice crystal will result, rather. (Relevance: 0.5)
2445	6. Water at about 4 °C (39 °F) also sinks to the bottom, thus keeping the temperature of the water
2446	at the bottom constant (see diagram). (Relevance: 0)
2447	7. However, even with this definition it is not clear whether freezing refers to the point at which
2448	water forms a visible surface layer of ice; the point at which the entire volume of water becomes a
2449	solid block of ice; or when the water reaches. (Relevance: 0)
2450	Query: Which is a fact about penguins?
2451	Choices: A:Penguins can live in climates with freezing temperatures. B:Penguins are fierce
2452	competitors. C:Penguins are some of the most beautiful birds. D:Penguins make great pets.
2453	Label: A
	Extracted Basic Knowledge: 1. Penguin Penguins (order Sphenisciformes, family Spheniscidae) are a group of aquatic,
2454	flightless birds. Highly adapted for life in the water, penguins have countershaded dark and white
2455	plumage, and their wings have evolved into flippers. Although almost all penguin species are
2456	native to the Southern Hemisphere. (Relevance: 1)
2457	2. Although almost all penguin species are native to the Southern Hemisphere, they are not found
2458	only in cold climates, such as Antarctica. In fact, only a few species of penguin actually live so
2459	far south. Several species live in the temperate zone; one, the Galápagos penguin, lives as far
2460	north.Emperor penguin The emperor penguin (Aptenodytes forsteri) is the tallest and heaviest of
2461	all living penguin species and is endemic to Antarctica. (Relevance: 1) 3. Penguins can live in climates with freezing temperatures. (Relevance: 1)
2462	4. Galápagos penguin The Galápagos penguin (Spheniscus mendiculus) is a penguin endemic to
2463	the Galápagos Islands.It is the only penguin that lives north of the equator. The Galápagos penguin
2464	is one of the banded penguins, the other species of which live mostly on the coasts of Africa and
2465	mainland South America. It can survive due to the cool temperatures resulting from the Humboldt
2466	Current and cool waters from great depths brought up by the Cromwell Current. (Relevance: 1)
	5. Although almost all penguin species are native to the Southern Hemisphere, they are not found
2467	only in cold climates, such as Antarctica. In fact, only a few species of penguin actually live so far
2468	south. Several species live in the temperate zone; one, the Galápagos penguin, lives as far north.
2469	(Relevance: 0.75) 6. Several authors have suggested that penguins are a good example of Bergmann's Rule where
2470	larger bodied populations live at higher latitudes than smaller bodied populations. There is some
2471	disagreement about this, and several other authors have noted that there are fossil penguin species
2472	that contradict this hypothesis and that ocean currents and upwellings are likely to have had a
2473	greater effect on species diversity than latitude alone. (Relevance: 0.75)
2474	7. Emperor penguin The emperor penguin (Aptenodytes forsteri) is the tallest and heaviest of
2475	all living penguin species and is endemic to Antarctica.Like all penguins it is flightless, with a
2476	streamlined body, and wings stiffened and flattened into flippers for a marine habitat. (Relevance:
2477	0.25)

Figure A.23: Test samples in ARC-Challenge (a).

2535

2484 2485 2486 Query: Burning fossil fuels produces sulfur dioxide (SO2) and nitrogen oxide (NO). These 2487 compounds react with water vapor to produce acid rain. What is the most likely effect of acid rain on the environment where it falls? 2488 Choices: A:The plants and animals in lakes and ponds will be harmed. B:The soil in the area will 2489 become more alkaline. C:The thickness of the ozone layer will decrease. D:Levels of air pollution 2490 will increase. 2491 Label: A 2492 **Extracted Basic Knowledge:** 1. Acid rain can damage infrastructures containing calcite or other solid chemical compounds 2493 containing carbon. In ecosystems, acid rain can dissolve plant tissues of vegetations and increase 2494 acidification process in bodies of water and in soil. (Relevance: 1) 2495 2. Acid rain is caused by the emission of nitrogen oxides and sulfur dioxide. These gases may be only mildly acidic themselves, yet when they react with the atmosphere, they create acidic 2497 compounds such as sulfurous acid, nitric acid and sulfuric acid which fall as rain, hence the term acid rain. (Relevance: 1) 2498 3. the most important gas which leads to acidification is sulphur dioxide. Emissions of nitrogen 2499 oxides which are oxidized to form nitric acid are of increasing importance due to stricter controls on emissions of sulphur containing compounds. Thus, for example, fumaroles from the Laguna' 2501 (Relevance: 0.75) 4. with water and oxygen in the atmosphere, creating nitric acid and sulfuric acids, which return 2503 to Earth's surface as acid deposition, or acid rain. Acid deposition harms aquatic organisms and kills trees. Due to its formation of certain nutrients which are less available to plants such as calcium and phosphorus, it reduces the productivity of ecosystem and farms. (Relevance: 0.75) 2505 5. Acid rain Acid rain is a rain or any other form of precipitation that is unusually acidic, meaning 2506 that it has elevated levels of hydrogen ions (low pH). It can have harmful effects on plants, aquatic 2507 animals and infrastructure. (Relevance: 0.75) 6. Sulfur dioxide and nitrogen oxides are naturally released from volcanoes, organic compounds in the soil, wetlands, and marine systems, but the majority of these compounds come from the 2509 combustion of coal, oil, gasoline, and the smelting of ores containing sulfur. These substances 2510 dissolve in atmospheric moisture and enter lentic systems as acid rain. Lakes and ponds that 2511 contain bedrock that is rich in carbonates have a natural buffer, resulting in no alteration of pH. 2512 Systems without this bedrock, however, are very sensitive to acid inputs because they have a low neutralizing capacity, resulting in pH declines even with. (Relevance: 0.5) 2513 7. Approximately 75 Tg/S per year of sulfur dioxide (SO) is released from burning coal. 2514 (Relevance: 0.5) 2515 8. The pH change is most marked in rivers with very (Relevance: 0.5) 2516 9. Finlayson-Pitts served as the lead author of a 2009 study published in the Proceedings of the 2517 National Academy of Sciences that found that burning fossil fuels releases nitrogen oxides, which interact with gaseous hydrogen chloride to form smog-forming compounds. (Relevance: 0.5) 2518 10. Sulfur dioxide and nitrogen oxides are primary causes of acid rain. These by-products are still 2519 a problem, but they have been greatly diminished in most advanced countries due to clean air regulations. (Relevance: 0.25) 2521 11. Acid precipitation can lead to asthma, bronchitis, lung inflammation, emphysema, and other 2522 lung and heart diseases. (Relevance: 0) 12. Acid-producing gasses are also created by biological processes that occur on the land, in wetlands, and in the oceans. The major biological source of sulphur containing compounds is dimethyl sulfide. Nitric acid in rainwater is an important source of fixed nitrogen for plant life, and is also produced by electrical activity in the atmosphere such as lightning. (Relevance: 0) 2526 2527 2528 2529 Figure A.24: Test samples in ARC-Challenge (b).

2538 2540 Query: The place where John Laub is an American criminologist and Distinguished University 2541 Professor in the Department of Criminology and Criminal Justice at was founded in what year? Label: 1856. 2542 **Extracted Basic Knowledge:** 2543 1. John Laub: John H. Laub (born 1953) is an American criminologist and Distinguished 2544 University Professor in the Department of Criminology and Criminal Justice at the University of 2545 Maryland, College Park. (Relevance: 1) 2546 2. University of Maryland, College Park: The University of Maryland, College Park (often referred to as the University of Maryland, Maryland, UM, UMD, UMCP, or College Park) 2547 is a public research university located in the city of College Park in Prince George's County, 2548 Maryland, approximately 4 mi from the northeast border of Washington, D.C. Founded in 1856, 2549 the university is the flagship institution of the University System of Maryland. With a fall 2010 2550 enrollment of more than 37,000 students, over 100 undergraduate majors, and 120 graduate 2551 programs, Maryland is the largest university in the state and the largest in the Washington Metropolitan Area. It is a member of the Association of American Universities and competes in 2552 athletics as a member of the Big Ten Conference. (Relevance: 0.75) 2553 3. Lawrence W. Sherman: He is also a Distinguished University Professor at the University of Maryland's Department of Criminology and Criminal Justice in College Park. Founded in 1856. 2555 (Relevance: 0.75) 4. Charles Wellford: University of Maryland, College Park. (Relevance: 0.5) 2557 Query: Are the bands "Halestorm" and "Say Anything" from different states? Label: ves. 2559 **Extracted Basic Knowledge:** 2560 1. Halestorm: Halestorm is an American hard rock band from Red Lion, Pennsylvania... 2561 (Relevance: 1) 2. Say Anything (band): Say Anything is an American rock band from Los Angeles, California. (Relevance: 1) 2563 3. Max Bemis: Max Bemis is the lead singer, primary composer and primary lyricist of the band 2564 Say Anything. Say Anything is an American rock band from Los Angeles, California. (Relevance: 2565 0.5) 2566 4. The MySpace Transmissions (Say Anything EP): The MySpace Transmissions is a digital EP by Say Anything. (Relevance: 0.5) 2567 2568 Query: Which Canadian province did the famous computer scientist John Tsotsos serve as the 2569 Director of the Centre for Vision Research at a famous research university? 2570 Label: Ontario 2571 **Extracted Basic Knowledge:** 1. John Tsotsos: John Tsotsos is a Canadian Computer Scientist whose research focuses on the field of Computer Vision. He is currently the Canada Research Chair in Computer Vision at York University and served as the Director of the Centre for Vision Research at York University from 2574 2000-2006. (Relevance: 1) 2575 2. York University: York University (French: "Université York") is a public research university 2576 in Toronto, Ontario, Canada. (Relevance: 0.75) 3. Matti Pietikäinen (academic): He is Director of the Center for Machine Vision Research, and 2577 Scientific Director of Infotech Oulu. (Relevance: 0) 2578 4. McCarthy Formalism: In computer science and recursion theory the McCarthy Formalism 2579 (1963) of computer scientist John McCarthy clarifies the notion of recursive functions by use 2580 of the IF-THEN-ELSE construction common to computer science, together with four of the 2581 operators of primitive recursive functions: zero, successor, equality of numbers and composition. 2582 (Relevance: 0)

Figure A.25: Test samples in HotPotQA.

2583

2586

Turns 100' and Fernando Fernán Gómez's 'El extraño viaje'. (Relevance: 0.75) 4. El extraño viaje: El extraño viaje is a 1964 Spanish black drama film directed by Fernán Gómez. (Relevance: 0.75) 5. Fernando Fernán Gómez: Fernando Fernández Gó August 1921 – 21 November 2007) (Relevance: 0.75) 6. Love in Pawn: Love in Pawn i British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political fill best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: 0.75) 6. "Escala en la ciudad; The production company disbanded the following year, after de had made Escala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	Query: Which film has the director who was born later, El Extraño Viaje or Love In Pawa
1. Charles Saunders (director): Charles Joel Saunders (8 April 1904 – April 1997) was an film director and screenwriter who started in the industry as a film editor, and who also cot to television. (Relevance: 1) 2. Fernando Fernán Gómez: Fernando Fernández Gómez (28 August 1921 – 21 Novemb better known as Fernando Fernán-Gómez was a Spanish actor, screenwriter, film director director and member of the Royal Spanish Academy for seven years (Relevance: 1) 3. Rafaela Aparicio: The most remembered are Carlos Saura's 'Anna and the Wolves Turns 100' and Fernando Fernán Gómez's 'El extraño viaje'. (Relevance: 0.75) 4. El extraño viaje: El extraño viaje is a 1964 Spanish black drama film directed by Fernán Gómez. (Relevance: 0.75) 5. Fernando Fernán Gómez: Fernando Fernández Gó August 1921 – 21 November 2007) (Relevance: 0.75) 6. Love in Pawn: Love in Pawn i British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political filbest known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: "Relevance on the citizen of the production company disbanded the following year, after de had madeEscala en la ciudad, his feature film debut, and Saslavsky had completed his second film cotor, screenwriter and film producer, a	
film director and screenwriter who started in the industry as a film editor, and who also cot to television. (Relevance: 1) 2. Fernando Fernán Gómez: Fernando Fernández Gómez (28 August 1921 – 21 Novemb better known as Fernando Fernán-Gómez was a Spanish actor, screenwriter, film director director and member of the Royal Spanish Academy for seven years (Relevance: 1) 3. Rafaela Aparicio: The most remembered are Carlos Saura's 'Anna and the Wolves Turns 100' and Fernando Fernán Gómez's 'El extraño viaje'. (Relevance: 0.75) 4. El extraño viaje: El extraño viaje is a 1964 Spanish black drama film directed by Fernán Gómez. (Relevance: 0.75) 5. Fernando Fernán Gómez: Fernando Fernández Gód August 1921 – 21 November 2007) (Relevance: 0.75) 6. Love in Pawn: Love in Pawn i British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political fil best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: "Escala en la ciudad: The production company disbanded the following year, after de had madeËscala en la ciudad; his feature film debut, and Saslavsky had completed his seconst famous movie, Crimen a las tres." (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent direc	
to television. (Relevance: 1) 2. Fernando Fernán Gómez: Fernando Fernández Gómez (28 August 1921 – 21 Novemb better known as Fernando Fernán-Gómez was a Spanish actor, screenwriter, film director director and member of the Royal Spanish Academy for seven years (Relevance: 1) 3. Rafaela Aparicio: The most remembered are Carlos Saura's 'Anna and the Wolves Turns 100' and Fernando Fernán Gómez's 'El extraño viaje'. (Relevance: 0.75) 4. El extraño viaje: El extraño viaje is a 1964 Spanish black drama film directed by Fernán Gómez. (Relevance: 0.75) 5. Fernando Fernán Gómez: Fernando Fernández Gó August 1921 – 21 November 2007) (Relevance: 0.75) 6. Love in Pawn: Love in Pawn i British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political fil best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance "Escala en la ciudad: The production company disbanded the following year, after de had madeEscala en la ciudad; his feature film debut, and Saslavsky had completed his seconst famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
2. Fernando Fernán Gómez: Fernando Fernández Gómez (28 August 1921 – 21 Novemb better known as Fernando Fernán-Gómez was a Spanish actor, screenwriter, film director director and member of the Royal Spanish Academy for seven years (Relevance: 1) 3. Rafaela Aparicio: The most remembered are Carlos Saura's 'Anna and the Wolves Turns 100' and Fernando Fernán Gómez's 'El extraño viaje'. (Relevance: 0.75) 4. El extraño viaje: El extraño viaje is a 1964 Spanish black drama film directed by Fernán Gómez. (Relevance: 0.75) 5. Fernando Fernán Gómez: Fernando Fernández Gó August 1921 – 21 November 2007) (Relevance: 0.75) 6. Love in Pawn: Love in Pawn i British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political filbest known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance O. "Escala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres." (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
better known as Fernando Fernán-Gómez was a Spanish actor, screenwriter, film director director and member of the Royal Spanish Academy for seven years (Relevance: 1) 3. Rafaela Aparicio: The most remembered are Carlos Saura's 'Anna and the Wolves Turns 100' and Fernando Fernán Gómez's 'El extraño viaje'. (Relevance: 0.75) 4. El extraño viaje: El extraño viaje is a 1964 Spanish black drama film directed by Fernán Gómez. (Relevance: 0.75) 5. Fernando Fernán Gómez: Fernando Fernández Gó August 1921 – 21 November 2007) (Relevance: 0.75) 6. Love in Pawn: Love in Pawn i British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Valas Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political fil best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance Go. "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad, his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
director and member of the Royal Spanish Academy for seven years (Relevance: 1) 3. Rafaela Aparicio: The most remembered are Carlos Saura's 'Anna and the Wolves Turns 100' and Fernando Fernán Gómez's 'El extraño viaje'. (Relevance: 0.75) 4. El extraño viaje: El extraño viaje is a 1964 Spanish black drama film directed by Fernán Gómez. (Relevance: 0.75) 5. Fernando Fernán Gómez: Fernando Fernández Gó-August 1921 – 21 November 2007) (Relevance: 0.75) 6. Love in Pawn: Love in Pawn i British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political filb best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance G. "Escala en la ciudad; The production company disbanded the following year, after de had made Escala en la ciudad, his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
Turns 100' and Fernando Fernán Gómez's 'El extraño viaje'. (Relevance: 0.75) 4. El extraño viaje: El extraño viaje is a 1964 Spanish black drama film directed by Fernán Gómez. (Relevance: 0.75) 5. Fernando Fernán Gómez: Fernando Fernández Gó August 1921 – 21 November 2007) (Relevance: 0.75) 6. Love in Pawn: Love in Pawn i British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political filbest known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance 6. "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres. (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
4. El extraño viaje: El extraño viaje is a 1964 Spanish black drama film directed by Fernán Gómez. (Relevance: 0.75) 5. Fernando Fernán Gómez: Fernando Fernández Gó August 1921 – 21 November 2007) (Relevance: 0.75) 6. Love in Pawn: Love in Pawn i British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political filbest known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: 0.75) 6. "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	3. Rafaela Aparicio: The most remembered are Carlos Saura's 'Anna and the Wolves'
Fernán Gómez. (Relevance: 0.75) 5. Fernando Fernán Gómez: Fernando Fernández Gó August 1921 – 21 November 2007) (Relevance: 0.75) 6. Love in Pawn: Love in Pawn i British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political filbest known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: 6. "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad: The production company disbanded the following year, after de had made Escala	
August 1921 – 21 November 2007) (Relevance: 0.75) 6. Love in Pawn: Love in Pawn i British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political film best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: 0.75) 6. "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky(April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
British comedy film directed by Charles Saunders and starring Bernard Braden and Barba (Relevance: 0.75) Query: Which film has the director died first, Crimen A Las Tres or The Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political film best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: 0.75) 6. "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky(April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
Query: Which film has the director died first, Crimen A Las Tres or The Valass Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political film best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: 0. "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky(April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	British comedy film directed by Charles Saunders and starring Bernard Braden and Barbar
Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political film best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: 0. "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad: his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky(April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	(Relevance: 0.75)
Class Goes To Heaven? Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political film best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: 0. "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad: his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky(April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
Label: The Working Class Goes To Heaven Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political film best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: 0.5) 6. "Escala en la ciudad: The production company disbanded the following year, after de nad madeËscala en la ciudad, his feature film debut, and Saslavsky had completed his seconost famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky(April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
Extracted Basic Knowledge: 1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1) 2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political film best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: 0."Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky(April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political film best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky(April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
at the 1972 Cannes Film Festival. (Relevance: 0.75) 3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri (29 January 1929 – 10 November 1982) was an Italian political file best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance: "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	1. Luis Saslavsky: He died in Buenos Aires, aged 91. (Relevance: 1)
3. The Working Class Goes to Heaven: The Working Class Goes to Heaven(released in the Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri (29 January 1929 – 10 November 1982) was an Italian political file best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance 6. "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	2. The Mattei Affair: The film shared the 'Grand Prix' with 'The Working Class Goes to I
Lulu the Tool) is a 1971 political drama film directed by Elio Petri. (Relevance: 0.75) 4. Elio Petri: Elio Petri(29 January 1929 – 10 November 1982) was an Italian political fil best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance 6. "Escala en la ciudad: The production company disbanded the following year, after de had madeËscala en la ciudad; his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
4. Elio Petri: Elio Petri (29 January 1929 – 10 November 1982) was an Italian political fil best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance 6. "Escala en la ciudad: The production company disbanded the following year, after de had made Escala en la ciudad, his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
best known for the 1970 Academy Award- winning film 'Investigation of a Citizer Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance 6. "Escala en la ciudad: The production company disbanded the following year, after de had madeËscala en la ciudad, his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
Suspicion'. (Relevance: 0.75) 5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance 5. "Escala en la ciudad: The production company disbanded the following year, after de had madeËscala en la ciudad, his feature film debut, and Saslavsky had completed his seconost famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent lirector, screenwriter and film producer, and one of the influential directors in the Ci	
6. "Escala en la ciudad: The production company disbanded the following year, after de had madeËscala en la ciudad, his feature film debut, and Saslavsky had completed his sec most famous movie, Crimen a las tres. (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
had madeËscala en la ciudad, his feature film debut, and Saslavsky had completed his second most famous movie, Crimen a las tres. (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky (April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	5. Crimen a las tres: Luis Saslavsky directed and wrote 'Crimen a las tres'.", (Relevance:
most famous movie, Crimen a las tres: (Relevance: 0.5) 7. Luis Saslavsky: Luis Saslavsky(April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
7. Luis Saslavsky: Luis Saslavsky(April 21, 1903 – March 20, 1995) was an Argent director, screenwriter and film producer, and one of the influential directors in the Ci	
director, screenwriter and film producer, and one of the influential directors in the Ci	
Algeriana of the classic eta. (Refevallee, U.J)	Argentina of the classic era. (Relevance: 0.5)

Figure A.26: Test samples in 2WikiMultiHopQA.

Test Samples in	n MuSiQue
Query: Who is	the child of Caroline LeRoy's spouse?
Label: Fletcher	
Extracted Basic	
	by: Caroline LeRoy Webster (September 28, 1797 in New York City – Febe second wife of 19th Century statesman Daniel Webster. Her father was Ho
LeRoy. (Relevan	
	ster: Daniel Fletcher Webster, commonly known as Fletcher Webster (Ju
1813 in Portsmo	outh, New Hampshire - August 30, 1862) was the son of renowned poli
	and Grace Fletcher Webster. (Relevance: 0.5)
	by: Her father was Herman LeRoy, who was once head of the commercial
world. (Relevance	I, McKiven & Co., a large trading company that operated in different parts
world. (Relevance	A. 0.3)
Query: What co	ompany succeeded the owner of Empire Sports Network?
Label: Time Wa	
Extracted Basic	
	Network: Empire Sports Network was an American regional sports network Adelphia Communications Corporation. (Relevance: 0.75)
	Communications: Windjammer Cable is a small cable company formed l
	ns that served 80,000 customers in rural areas that Time Warner Cable acc
from the bankruj	pt Adelphia. (Relevance: 0.75)
	Eigen A 27. Test segueles in McCiOcc
	Figure A.27: Test samples in MuSiQue.

2700	Test Samples in LongBench-v2
2701	Test Samples in Longbellen-v2
2702	Query: Which of the following is an incorrect understanding of the discussion on Journey to the
2703	West in the development of Chinese mythological culture?
2704	Choices:
2705	A. Journey to the West expresses the characteristics of mythological figures where divinity, animal
2706	instincts, and human nature are integrated through the character of Sun Wukong.
2707	B. The social circlein Journey to the West symbolizes the refinement of human character and the transformation of spiritual will.
2708	C. The novel elevates a free mindset beyond religion by incorporating the ideological resources of
2709	Confucianism, Buddhism, and Daoism, blending in the mysticism of the three teachings.
2710	D. The image of Sun Wukong symbolizes the transformation from the outburst of wild vitality to
2711	the elevation of spiritual realms.
2712	Label: C
2713	Extracted Basic Knowledge: 1. Journey to the West Chapter 24: Blessed Land of the Mountain of Infinite Longevity - Cave
2714	Heaven of the WuZhuang Temple: Although the Tang Priest is an old friend of mine, said the
2715	Great Immortal, you must be on your guard against his ruffian followers, and you mustn't let them
2716	know about the manfruit. It's either a Taoist temple or a Buddhist one. Let's go over and find out.
2717	Residence of Divine Immortals Who Never Grow Old; Home of Taoists as Ancient as Heaven.
2717	(Relevance: 0.5)
	2. On his head A leopard skin hat with artemisia patterns: On his body A coat of woollen cloth. Round his waist was tied a lion belt, On his feet a pair of deerskin boots. His eyes were as round
2719	as an evil spirit's; His curly beard was like the evil god of the moon's. From his waist hung a bow
2720	with poisoned arrows, And in his hand was a steel-tipped trident. (Relevance: 0.25)
2721	
2722	Query: Which of the following statements is incorrect?
2723	Choices:
2724	A. This article inserts a module into the pre-trained diffusion model, and then trains the parameters of these models to adapt this module to the task and the priori of the diffusion model.
2725	B. TPB includes two MLP layers with Layer Normalization and LeakyReLU, ensuring that only
2726	the most task-specific attributes are retained
2727	C. Task-specific priors containing guidance information for the task can adequately guide
2728	pre-trained diffusion models to handle low-level tasks while maintaining high-fidelity content
2729	consistency.
2730	D. The spatial feature Fs extracted by SCB processing is calculated from SCB, Ft, Fp, F and has no relationship with TPB.
2731	Label: D
2732	Extracted Basic Knowledge:
2733	1. Introducing Diff-Plugin: A Framework for Enhancing Pre-Trained Diffusion Models with
2734	Task-Specific Guidance: The spatial feature Fs extracted by SCB processing is calculated from
2735	SCB, Ft, Fp, F, and has no relationship with TPB. (Relevance: 1)
2736	2. Implementation: During training and testing, we resize the image to 512×512 for a fair comparison. (Relevance: 0.25)
2737	comparison. (recevance, 0.23)
2738	
2739	
2740	Figure A.28: Test samples in LongBench-v2.
2741	

I.2.2 SAMPLES OF INACCURATE RELEVANCE SCORING

Samples of Inaccurate Relevance Scoring

Query: Which member of the band Bad Seeds was older, Anita Lane or Nick Cave? **Extracted Knowledge:**

- 1. Anita Lane: Anita Louise Lane (born ca. 1959) is an Australian singer-songwriter who was briefly a member of the Bad Seeds with Nick Cave and Mick Harvey, and has collaborated with both former band mates. (Relevance: 1)
- 2. Nick Cave: Nicholas Edward Cave (born 22 September 1957) is an Australian musician, singer-songwriter, author, screenwriter, composer and occasional film actor, best known as the frontman of the rock band Nick Cave and the Bad Seeds. (Relevance: 0.75)

Query: What album did a Danish-born Montenegrin singer born in 1971 release in late spring 2008?

Extracted Knowledge:

- 1. Boban Rajović: Boban Rajović (born 25 December 1971) is a Danish-born Montenegrin singer popular in former Yugoslavia. (Relevance: 0.5)
- 2. Kosači: Kosači (English translation: Mowers) is the fifth studio album by Montenegrin singer Boban Rajović. t was released in late spring 2008. (Relevance: 1)

Figure A.29: Samples of Inaccurate Relevance Scoring.

I.2.3 SAMPLES OF DISCREPANCIES BETWEEN GPT-40 AND HUMAN ASSESSMENTS

Samples of Discrepancies

2810 2811 2812

2813

2814

2815

2816

2817

2818

2819

2821

2822

2825

2829

2831

2833

2834

2836

2837

2839

2841

2842

2843

2846

2847

2849

2850

2851

2852

Query: Musician and satirist Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who?

Labels: President Richard Nixon.

Retrieved Documents:

- 1. Allie Goertz: Allison Beth "Allie" Goertz (born March 2, 1991) is an American musician. Goertz is known for her satirical songs based on various pop culture topics. Her videos are posted on YouTube under the name of Cossbysweater. Subjects of her songs have included the film "The Room", the character Milhouse from the television show "The Simpsons", and the game Dungeons & Dragons. Her style has been compared to that of Bo Burnham. In December 2015, Goertz released a concept album based on the Adult Swim series "Rick and Morty", "Sad Dance Songs", with the album's cover emulating the animation and logo of the series. The album was made possible through Kickstarter. She is co-host of Everything's Coming Up Podcast, a Simpsons-focused podcast along with Julia Prescott.
- 2. Milhouse Van Houten: Milhouse Mussolini van Houten is a fictional character featured in the animated television series "The Simpsons", voiced by Pamela Hayden, and created by Matt Groening who named the character after President Richard Nixon's middle name. Later in the series, it is revealed that Milhouse's middle name is "Mussolini."

RALM Prediction: President Richard Nixon's middle name.

Evaluation (GPT-40, llm-em): Incorrect. The label states that Matt Groening named the "The Simpsons" character Milhouse after President Richard Nixon. The student's answer states that Milhouse was named after President Richard Nixon's middle name, which is "Milhous." While technically related, the student's answer is not an exact match to the label and introduces potentially misleading specificity.

Evaluation (human, human-acc): Correct.

Justification for Assessment: The RALM prediction is accurate and supported by the retrieved documents. Its specificity enhances clarity without altering the core meaning.

Query: Which player singed to Bayern Munich was born in 18 October 1983?

Labels: Dante Bonfim Costa Santos (born 18 October 1983)

Retrieved Documents:

- 1. 2012–13 FC Bayern Munich season: The 2012–13 FC Bayern Munich season was the 114th season in the club's history and the 48th consecutive season in the top flight of German football, the Bundesliga, since the promotion of the team from the Regionalliga Süd in 1965. Before the start of the season, Bayern signed Xherdan Shaqiri, Dante, Claudio Pizarro, Mitchell Weiser, Tom Starke and Mario Mandžukić. Bayern also added holding midfielder Javi Martínez after the first week of the Bundesliga season at the transfer deadline. The club started the season with a nine-match winning streak. The club would end the season claiming the Treble, winning the Bundesliga, the UEFA Champions League and the DFB-Pokal. Bayern are the first German club to achieve the Treble and are the third European Club to complete the Treble in the last five seasons and seventh ever in European Club competition.
- 2. Dante (footballer): Dante Bonfim Costa Santos (born 18 October 1983), commonly known as Dante ()), is a Brazilian professional footballer who plays for French Ligue 1 club Nice. Primarily a central defender, he has previously also been used as a defensive midfielder or a left back.

RALM Prediction: Dante Bonfim Costa Santos (Dante).

Evaluation (GPT-40, llm-em): Incorrect. The label specifies that Dante Bonfim Costa Santos, born on 18 October 1983, is the player signed to Bayern Munich. The student's answer gives the name "Dante Bonfim Costa Santos (Dante)" but does not include the birthdate. The essential information about the birthdate is missing in the student's answer.

Evaluation (human, human-acc): Correct.

Justification for Assessment: The RALM prediction accurately identifies Dante Bonfim Costa Santos as the player in question. While it does not explicitly mention the birthdate, it provides all necessary information to address the query correctly.

285628572858

2859

2860

Figure A.30: **Samples of discrepancies** (a). In the evaluation process with GPT-40, we only provide the queries, ground truth labels, and the predicted answers from RALM, without including any retrieval documents. Additionally, in the GPT-40 prompt, we substitute the RALM with the role of a student answering the question.

2862	Samples of Discrepancies
2863	
2864	Query: When was Umar Farouk Abdulmutallab attempting to detonate plastic explosives on
2865	Northwest Airlines Flight 253?
2866	Labels: Christmas Day, December 25, 2009
2867	Retrieved Documents: 1. Umar Farouk Abdulmutallab: Umar Farouk Abdulmutallab; also known as Umar Abdul
2868	Mutallab and Omar Farooq al-Nigeri; born December 22, 1986) popularly referred to as the
2869	"Underwear Bomber", is a Nigerian man who, at the age of 23, confessed to and was convicted
2870	of attempting to detonate plastic explosives hidden in his underwear while on board Northwest
2871	Airlines Flight 253, en route from Amsterdam to Detroit, Michigan, on Christmas Day, 2009.
2872	2. Northwest Airlines Flight 253: Northwest Airlines Flight 253 was an international passenger
2873	flight from Amsterdam Airport Schiphol in Haarlemmermeer, Netherlands, to Detroit Metropolitan
2874	Wayne County Airport in Romulus, Michigan, United States. The flight was the target of a failed
2875	al-Qaeda bombing attempt on Christmas Day, December 25, 2009, in which a passenger tried to set off plastic explosives sewn to his underwear. There were 290 people on board the aircraft—an
2876	Airbus A330-323E operated by Northwest Airlines, which had merged with Delta Air Lines the
	year before. Had the attempt succeeded, it would have surpassed American Airlines Flight 191 as
2877	the deadliest aviation occurrence on U.S. soil and tied Iran Air Flight 655 as the eighth-deadliest
2878	of all time. The incident was also the second in 2009 involving an Airbus A330, after the crash of
2879	Air France Flight 447 on June 1. This was also the final accident/incident involving Northwest
2880	Airlines as it closed down a month later when it merged with Delta Air Lines.
2881	RALM Prediction: Christmas Day, 2009.
2882	Evaluation (GPT-40, llm-em): Incorrect. The student's answer states "Christmas Day, 2009", which is mostly correct but leaks the provide data "December 25" as provided in the lebel. The
2883	which is mostly correct but lacks the specific date "December 25" as provided in the label. The label specifies "Christmas Day, December 25, 2009", and the student's answer omits "December
2884	25".
2885	Evaluation (human, human-acc): Correct.
2886	Justification for Assessment: The RALM prediction accurately identifies the timing of the
2887	incident as "Christmas Day, 2009." While it lacks the exact date included in the reference label, it
2888	provides enough information to satisfy the query and aligns with the retrieved evidence.
2889	
2890	Query: In what year was the creator of the current arrangement of the "Simpson's Theme" born?
2891	Labels: March 28, 1941.
	Retrieved Documents:
2892	1. Alf Clausen: Alf Heiberg Clausen (born March 28, 1941) is an American film and television
2893	composer. He is best known for his work scoring many episodes of "The Simpsons", of which he
2894	had been the sole composer between 1990 and 2017. Clausen has scored or orchestrated music for
2895	more than 30 films and television shows, including "Moonlighting", "The Naked Gun", "ALF"
2896	and "Ferris Bueller's Day Off".
2897	2. The Simpsons Theme: "The Simpsons Theme", also referred to as "The Simpsons Main Title
2898	Theme" in album releases, is the theme music of the animated television series "The Simpsons". It plays during the opening sequence and was composed by Danny Elfman in 1989, after series
2899	creator Matt Groening approached him requesting a retro-style theme. The piece, which took 3
2900	days, 2 hours, 48 minutes, and 19 seconds to create, has been noted by Elfman as the most popular
2901	of his career. The theme, as used for the opening sequence, was re-arranged during season 2, and
2902	the current arrangement by Alf Clausen was introduced at the beginning of the third season.
2903	RALM Prediction: 1941.
2904	Evaluation (GPT-40, llm-em): Incorrect. The label specifies the exact birthdate, "March 28,
	1941", while the student's answer provides only the year, "1941". Although the year is correct, it
2905	does not fully match the detailed information given in the label.
2906	Evaluation (human, human-acc): Correct. Justification for Assessment: The RALM prediction accurately answers the query by providing
2907	the year of birth (1941). While it is less detailed than the reference label, it remains fully
2908	consistent with the retrieved evidence and satisfies the query's requirements.
2909	,

Figure A.31: Samples of discrepancies (b).

2916	Samples of Discrepancies
2917	
2918	Query: What is the name of the widow of the earl who served as both Foreign Secretary and
2919	Prime Minster from 1955 to 1957?
2920	Labels: Anne Clarissa Eden. Retrieved Documents:
2921	1. Clarissa Eden: Anne Clarissa Eden, Dowager Countess of Avon ("née" Spencer-Churchill;
2922	born 28 June 1920) is the widow of Anthony Eden, 1st Earl of Avon (1897–1977), who was
2923	British Prime Minister from 1955 to 1957. She married Eden in 1952, becoming Lady Eden in
2924	1954 when he was made a Knight of the Garter, and then becoming Countess of Avon in 1961
2925	on her husband's elevation to the peerage. She is also the niece of the prime minister Winston
2926	Churchill. Her memoir, sub-titled "From Churchill to Eden", was published in 2007 under the
2927	name of Clarissa Eden.
2928	2. Anthony Eden: Robert Anthony Eden, 1st Earl of Avon, (12 June 1897 – 14 January 1977) was a British Conservative politician who served three periods as Foreign Secretary and then a
2929	relatively brief term as Prime Minister of the United Kingdom from 1955 to 1957.
2930	RALM Prediction: Clarissa Eden.
2931	Evaluation (GPT-40, llm-em): Incorrect. The label specifies "Anne Clarissa Eden" as the name of
2932	the widow. The student's answer "Clarissa Eden" provides a partial but not completely accurate
2933	answer, as it omits the first name "Anne". Therefore, the student's answer does not fully match
2934	the label. Evaluation (human human aga), Correct
2935	Evaluation (human, human-acc): Correct. Justification for Assessment: The RALM prediction accurately identifies the widow as "Clarissa
2936	Eden." While it omits her full name, the answer is precise, widely recognized, and supported by
2937	the retrieved documents.
2938	O WELLING CELLS LILL LILA d II
2939	Query: "The Little Sisters of Eluria" was included in which American author's collection "Everything's Eventual"?
2940	Labels: Stephen King.
2941	Retrieved Documents:
2942	1. The Little Sisters of Eluria: "The Little Sisters of Eluria" is a Fantasy fiction novella written
2943	by Stephen King. It was originally published in 1998 in the anthology "Legends". In 2002, it
	was included in King's collection "Everything's Eventual". In 2009, it was published together
2944	with the of by Grant in a limited edition of 4,000 numbered copies of the Artist Edition signed by illustrator Michael Whelan and 1,250 numbered copies of the Deluxe Edition signed by
2945	Whelan and Stephen King. Both editions contain Whelan's additional new illustrations for "The
2946	Gunslinger".
2947	2. Stephen King: Stephen Edwin King (born September 21, 1947) is an American author of
2948	horror, supernatural fiction, suspense, science fiction, and fantasy. His books have sold more than
2949	350 million copies, many of which have been adapted into feature films, miniseries, television
2950	series, and comic books. King has published 54 novels, including seven under the pen name Richard Bachman, and six non-fiction books. He has written nearly 200 short stories, most of
2951	which have been collected in book collections. Many of his stories are set in his home state of
2952	Maine. His novella "Rita Hayworth and Shawshank Redemption" was the basis for the film "The
2953	Shawshank Redemption" which is widely regarded as one of the greatest films of all time.
2954	RALM Prediction: Stephen King's collection "Everything's Eventual".
2955	Evaluation (GPT-40, llm-em): Incorrect. The student's answer mentions the collection
2956	"Everything's Eventual" but does not explicitly state that "The Little Sisters of Eluria" was included in Stephan King's collection. Although it hints at the correct outbor, it is not a complete
2957	included in Stephen King's collection. Although it hints at the correct author, it is not a complete match to the query and label.
2958	Evaluation (human, human-acc): Correct.
2959	Justification for Assessment: The RALM prediction identifies the correct author, Stephen King,
2960	and provides additional context about the collection Everything's Eventual. This additional detail
2961	is accurate and does not conflict with the reference label.
2962	

Figure A.32: Samples of discrepancies (c).

2970 2972 Query: What county music artist first gained national attention as a winner on "You Can Be a 2973 Star" and wen on to provide background vocals in 199 on the 1992 release of the song "Take It Like a Man"? 2974 Labels: Lari Michele White. 2975 **Retrieved Documents:** 2976 1. "Take It Like a Man (Michelle Wright song)": "Take It Like a Man" is a song written by Tony Haselden, and recorded by Canadian country music artist Michelle Wright. It was released in February 1992 as the first single from Wright's album "Now and Then". The song became Wright's first Number One on the Canadian "RPM" Country Tracks charts that year. It was also 2979 her only top 10 hit on the "Billboard" Hot Country Singles & Tracks chart in the United States. 2980 Country singer Lari White is featured on background vocals. 2981 2. Lari White: Lari Michele White (born May 13, 1965) is an American country music artist and actress. She first gained national attention in 1988 as a winner on "You Can Be a Star", a talent 2983 competition which aired on The Nashville Network. A recording contract with RCA Records Nashville followed a year later, producing three studio albums, a greatest hits package, and several 2984 chart singles, with three of her singles having reached Top Ten: "That's My Baby" and "That's 2985 How You Know (When You're In Love)" at No. 10, and "Now I Know" at No. 5. A fourth studio album was released in 1998 on Lyric Street Records, followed by two more releases on White's 2987 own label, Skinny White Girl. Overall, White has charted 12 times on the "Billboard" country music charts. RALM Prediction: "Lari White", "Take It Like a Man". 2989 Evaluation (GPT-40, llm-em): Incorrect. The student's answer "Lari White" correctly identifies the country music artist who first gained national attention as a winner on "You Can Be a Star." 2991 However, the additional phrase "Take It Like a Man" is unnecessary and does not fully restate the 2992 correct label "Lari Michele White." The answer is partially correct but not entirely accurate as per 2993 the label. **Evaluation (human, human-acc):** Correct. 2994 **Justification for Assessment:** The RALM prediction correctly identifies Lari White as the artist 2995 and provides additional context by mentioning the song title, Take It Like a Man. This expanded 2996 response enhances the completeness of the answer while maintaining full alignment with the query and evidence. 2998 Query: Caldera de Taburiente National Park and Teide National Park are located in 2999 which island chain? Labels: Canary Islands, Spain 3001 **Retrieved Documents:** 3002 1. Caldera de Taburiente National Park: Caldera de Taburiente National Park (Spanish: "Parque 3003 Nacional de la Caldera de Taburiente") is large geological feature on the island of La Palma, Canary Islands, Spain. It contains the enormous expanse of the Caldera de Taburiente which dominates the northern part of the island, and was designated as a national park in 1954. The 3005 telescopes of the Roque de los Muchachos Observatory are situated very close to the summit. 3006 2. Teide National Park': Teide National Park (Spanish: "Parque nacional del Teide",)) is a 3007 national park located in Tenerife (Canary Islands, Spain). 3008 **RALM Prediction:** The Canary Islands. Evaluation (GPT-40, llm-em): Incorrect. The student's answer, "The Canary Islands," partially matches the label "Canary Islands, Spain." While it correctly identifies the island chain, it omits 3010 the additional detail that the Canary Islands are part of Spain. 3011 **Evaluation (human, human-acc):** Correct. 3012 Justification for Assessment: The RALM prediction accurately identifies the Canary Islands as 3013 the location of both parks. While it omits "Spain" from the reference label, this does not affect the correctness of the core answer. 3014 3015

Figure A.33: Samples of discrepancies (d).

3016 3017

3018 3019 3020

3024 3025 3026 Query: Coll is known for Breacachadh also know as? 3027 Labels: Breachacha Castle **Retrieved Documents:** 3028 1. Coll: Coll (Scottish Gaelic: "Cola") is an island located west of Mull in the Inner Hebrides of Scotland. Coll is known for its sandy beaches, which rise to form large sand dunes, for its 3030 corncrakes, and for Breacachadh Castle. It is in the council area of Argyll and Bute. 3031 2. Breachacha Castle: Breachacha Castle (also spelled Breacachadh) is either of two structures 3032 on the shore of Loch Breachacha, on the Inner Hebridean island of Coll, Scotland. The earlier (also called Old Breachacha Castle) is a 15th-century tower house that was a stronghold of the 3033 Macleans of Coll, the island having been granted to John Maclean in 1431. This castle was 3034 superseded by a new dwelling in 1750 (see below) but continued to be occupied for a time, falling into a ruinous state only in the mid-19th century. Although work was performed in the 1930s to 3036 prevent further dilapidation, the castle was restored to livable condition only in the 1960s, by Major Neil V. MacLean Bristol and his wife Lavinia. It is a Category A listed building. 3037 **RALM Prediction:** Breacachadh Castle 3038 Evaluation (GPT-40, llm-em): Incorrect. The label specifies "Breachacha Castle" as the correct answer. The student's answer, "Breacachadh Castle," contains a misspelling of the name. The 3040 correct spelling is crucial for validation purposes, and the student's answer does not match the 3041 label exactly. 3042 **Evaluation (human, human-acc):** Correct. **Justification for Assessment:** The RALM prediction accurately identifies "Breacachadh Castle" as the answer. Its use of the spelling "Breacachadh" aligns with the query and retrieved documents, 3044 making the answer valid and consistent with the evidence. 3045 3046 Query: Patty Hill is perhaps best known for co-writing, with her sister Mildred Hill a 3047 song that was translated into how many languages? Labels: 18 languages 3048 **Retrieved Documents:** 3049 1. Patty Hill: Patty Smith Hill (March 27, 1868 – May 25, 1946) was a composer and teacher who 3050 is perhaps best known for co-writing, with her sister Mildred Hill, the tune which later became 3051 popular as "Happy Birthday to You". She was an American nursery school, kindergarten teacher, 3052 and key founder of the National Association for Nursery Education (NANE) which now exists as the National Association For the Education of Young Children (NAEYC). 3053 2. Happy Birthday to You: "Happy Birthday to You", more commonly known as simply 3054 "Happy Birthday", is a song that is traditionally sung to celebrate the anniversary of a person's 3055 birth. According to the 1998 "Guinness World Records", "Happy Birthday to You" is the most 3056 recognized song in the English language, followed by "For He's a Jolly Good Fellow". The song's 3057 base lyrics have been translated into at least 18 languages. The melody of "Happy Birthday to You" comes from the song "Good Morning to All", which has traditionally been attributed to American sisters Patty and Mildred J. Hill in 1893, although the claim that the sisters composed 3059 the tune is disputed. 3060 **RALM Prediction:** At least 18 3061 Evaluation (GPT-40, llm-em): Incorrect. The label specifies that the song was translated into 3062 "18 languages." The student's answer, "At least 18," implies a minimum number but does not precisely match the exact number provided by the label. Therefore, the student's answer does not accurately reflect the information given by the label. 3064 **Evaluation (human, human-acc):** Correct. **Justification for Assessment:** The RALM prediction accurately answers the query and aligns 3066 with the retrieved evidence. Its use of "at least" adds nuance without compromising correctness. 3067 3068

Figure A.34: Samples of discrepancies (e).

3070

3072

3078 3079 3080 Query: How many songs were written by the man who shares attribution for one of his songs 3081 with a poem by Emily Huntington Miller? Labels: 80 3082 **Retrieved Documents:** 1. Benjamin Hanby: Benjamin Russell Hanby (July 22, 1833 – March 16, 1867), also given 3084 as Benjamin Russel Hanby, was an American composer, educator, and pastor who wrote 3085 approximately 80 songs. The most famous of his compositions are "Darling Nelly Gray", the 3086 Christmas songs "Up on the House Top", "Jolly Old Saint Nicholas", and the hymn "Who Is He In Yonder Stall?". 3087 2. Jolly Old Saint Nicholas: "Jolly Old Saint Nicholas" is a Christmas song that originated with 3088 a poem by Emily Huntington Miller (1833-1913), published as "Lilly's Secret" in "The Little 3089 Corporal Magazine" in December 1865. The song's lyrics have also been attributed to Benjamin Hanby, who wrote a similar song in the 1860s, Up on the Housetop. However, the lyrics now in 3091 common use closely resemble Miller's 1865 poem. **RALM Prediction:** approximately 80 songs. 3092 **Evaluation (GPT-40, llm-em):** Incorrect. The label specifies that the number of songs written by 3093 the man who shares attribution for one of his songs with a poem by Emily Huntington Miller is exactly 80. The student's answer states "approximately 80 songs", which is close but not an exact 3095 match to the label. The use of "approximately" introduces ambiguity that is not present in the 3096 3097 **Evaluation (human, human-acc):** Correct. Justification for Assessment: The RALM prediction accurately answers the query, aligning with 3098 the evidence and reference label. Its use of "approximately" adds clarity without compromising 3099 correctness. 3100 3101 Query: The third single off Tim McGraw's twelfth studio album features which two 3102 Labels: Taylor Swift on vocals, as well as Keith Urban on guitar 3103 **Retrieved Documents:** 3104 1. Two Lanes of Freedom'; text: 'Two Lanes of Freedom is the twelfth studio album by American 3105 country music artist Tim McGraw. It was released February 5, 2013, as his first album for Big 3106 Machine Records following a 20-year tenure with Curb Records. He co-produced the album with Byron Gallimore, producer of his previously released albums. The album includes the singles 3107 "Truck Yeah", "One of Those Nights", "Highway Don't Care" featuring new labelmate Taylor 3108 Swift, and "Southern Girl". 3109 2. "Highway Don't Care": "Highway Don't Care" is a song performed by American country 3110 music artist Tim McGraw and features American singer-songwriter Taylor Swift on vocals, as 3111 well as Keith Urban on guitar. It was released in March 2013 as the third single from McGraw's first album for Big Machine Records, "Two Lanes of Freedom" (2013). The song was written 3112 by Mark Irwin, Josh Kear and Brad and Brett Warren. McGraw and Swift recorded their parts 3113 separately. 3114 **RALM Prediction:** Taylor Swift, Keith Urban. 3115 **Evaluation (GPT-40, llm-em):** Incorrect. The student's answer mentions both Taylor Swift and 3116 Keith Urban, which aligns with the label. However, the student's answer does not specify Taylor Swift's role on vocals and Keith Urban's role on guitar, which are important details present in the 3117 3118 **Evaluation (human, human-acc):** Correct. 3119 **Justification for Assessment:** The RALM prediction accurately identifies the two featured artists, 3120 Taylor Swift and Keith Urban, as required by the query. While it omits the roles they play in the 3121 song, this does not affect the correctness of the answer. 3122

Figure A.35: Samples of discrepancies (f).

3123 3124

3125 3126

3132 3133 3134 Query: How many people died in the fire Brooklyn Theatre that was owned by Sarah Crocker 3135 Labels: 278 individuals 3136 **Retrieved Documents:** 3137 1. Sarah Crocker Conway: Sarah Crocker Conway (b. Ridgefield, Connecticut, 1834; d. Brooklyn, 3138 New York, April 28, 1875) was an actress. She was a sister of Elizabeth Crocker Bowers, also an 3139 actress. She made her debut in Baltimore, Maryland in 1849, playing Parthenia and other leading 3140 parts. In May 1852, she married the actor Frederick B. Conway, and the two thenceforward acted together. In 1859, they opened Pike's Opera House in Cincinnati, Ohio with a first-class 3141 company, but the engagement was not profitable, and they returned to the east. In 1861, they 3142 visited England, and filled a short engagement at Sadler's Wells Theatre, London. After their 3143 return, they became star actors, and made an extensive and profitable tour. In 1864 Sarah leased 3144 the Park Theatre in Brooklyn, and subsequently the new Brooklyn Theatre, in which for nine 3145 years her husband played leading parts. Sarah Conway possessed a tall and graceful figure and an expressive countenance, and was a versatile actress and a popular theatre manager. 3146 2. Brooklyn Theatre fire: The Brooklyn Theatre Fire was a catastrophic theatre fire that broke 3147 out on the evening of December 5, 1876, in the then-city of Brooklyn (now a borough of New 3148 York City). The conflagration killed at least 278 individuals, with some accounts reporting more 3149 than 300 dead. One hundred and three unidentified victims were interred in a common grave at 3150 Green-Wood Cemetery. An obelisk near the main entrance at Fifth Avenue and 25th Street marks 3151 the burial site. More than two dozen identified victims were interred individually in separate sections at the Cemetery of the Evergreens in Brooklyn. 3152 **RALM Prediction:** At least 278 people died in the fire. 3153 **Evaluation (GPT-40, llm-em):** Incorrect. The student's answer states "at least 278 people died in 3154 the fire", which suggests a minimum number but does not definitively match the label's specific 3155 count of "278 individuals". The label specifies an exact number without implying that it could be higher. 3156 **Evaluation (human, human-acc):** Correct. 3157 Justification for Assessment: The RALM prediction accurately answers the query and aligns 3158 with the evidence. Its use of "at least" appropriately reflects the documented uncertainty without 3159 altering the correctness of the response. 3160 Query: Buckner Building was slightly damaged by a 9.2 earthquake that occurred on 3161 which day? 3162 Labels: on Good Friday, March 27 3163 **Retrieved Documents:** 1. Buckner Building: The Buckner Building is an abandoned former U.S. military building 3165 in Whittier, Alaska. Designed by Foss, Malcolm and Olsen architects, it was constructed by the military in 1953 as a combined mess hall, sleeping quarters, recreational, medical and 3166 administrative facility and used to be one of the largest buildings in Alaska, often being referred 3167 to as "the city under one roof". The building was slightly damaged by the 9.2 1964 Alaska 3168 earthquake, though damage was minimized by its reinforced concrete construction. 3169 2. 1964 Alaska earthquake: The 1964 Alaskan earthquake, also known as the Great Alaskan 3170 earthquake and Good Friday earthquake, occurred at 5:36 PM AST on Good Friday, March 27. Across south-central Alaska, ground fissures, collapsing structures, and tsunamis resulting from 3171 the earthquake caused about 139 deaths. 3172 **RALM Prediction:** March 27, 1964. 3173 **Evaluation (GPT-40, llm-em):** Incorrect. The label indicates that the Buckner Building was 3174 slightly damaged by a 9.2 earthquake on Good Friday, March 27. The student's answer provides 3175 the date "March 27, 1964," which is accurate in terms of the date but omits the fact that it was Good Friday. The label specifies both the date and the day (Good Friday), and the student's 3176 answer lacks the mention of Good Friday. 3177 **Evaluation (human, human-acc):** Correct. 3178 Justification for Assessment: The RALM prediction accurately provides the date of the 3179 earthquake. While it omits the additional context of "Good Friday," it satisfies the query and aligns 3180 with the retrieved evidence. The lack of this extra detail does not compromise the correctness of 3181 the answer. 3182

Figure A.36: Samples of discrepancies (g).

3183 3184