InfoGain-RAG: Boosting Retrieval-Augmented Generation through Document Information Gain-based Reranking and Filtering

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to address key limitations of Large Language Models (LLMs), such as hallucination, outdated knowledge, and lacking reference. However, current RAG frameworks often struggle with identifying whether retrieved documents meaningfully contribute to answer generation. This shortcoming makes it difficult to filter out irrelevant or even misleading content, which notably impacts the final performance. In this paper, we propose Document Information Gain (DIG), a novel metric designed to quantify the contribution of retrieved documents to correct answer generation. DIG measures a document's value by computing the difference of LLM's generation confidence with and without the document augmented. Further, we introduce InfoGain-RAG, a framework that leverages DIG scores to train a specialized reranker, which prioritizes each retrieved document from exact distinguishing and accurate sorting perspectives. This approach can effectively filter out irrelevant documents and select the most valuable ones for better answer generation. Extensive experiments across various models and benchmarks demonstrate that InfoGain-RAG can significantly outperform existing approaches, on both single and multiple retrievers paradigm. Specifically on NaturalQA, it achieves the improvements of 17.9%, 4.5%, 12.5% in exact match accuracy against naive RAG, self-reflective RAG and modern ranking-based RAG respectively, and even an average of 15.3% increment on advanced proprietary model GPT-4o across all datasets. These results demonstrate the feasibility of InfoGain-RAG as it can offer a reliable solution for RAG in multiple applications.

1 Introduction

Recent advancements in Natural Language Processing (NLP) have been significantly propelled by the emergence of LLMs (Brown et al., 2020; Achiam et al., 2024), which demonstrates remarkable capabilities across many knowledge-intensive tasks. However, maintaining reliability remains an ongoing challenge for LLMs, as they often struggle with issues such as hallucination, outdated information and lacking reference. RAG has emerged as a promising solution to the aforementioned issues. It can enhance responses by augmenting prompts with external information, especially when the model's inherent knowledge is limited (Ram et al., 2023). However, the generation quality heavily depends on both the selection of relevant documents and their sequential ordering within the LLMs' context window (Liu et al., 2023).

Research addressing RAG document prioritization spans multiple perspectives, of which three pipelines gain significant attention. The first pipeline focuses on retriever optimization, which enhances retrieval performance through task-specific training (Lewis et al., 2020; Shi et al., 2023; Chen et al., 2024a). However, this approach becomes impractical when working with multiple retrievers (Fan et al., 2024). The second pipeline leverages LLMs' self-reflection capabilities to evaluate the utility of documents. It employs LLMs to analyze each document and determine whether it should be used. Although feasible, the multiple LLM calls introduce substantial computational overhead (Asai et al., 2024; Yan et al., 2024; Chang et al., 2024). The third pipeline adds a reranker after the retrieval stage to reorder all retrieved documents (Chen et al., 2024b; Li et al., 2024). While this approach can effectively address multiple retrievers, the only consideration on semantic similarity may fail to select the most useful documents for generation (as shown in the Figure 6 of Appendix A). All these shortcomings limit their further practical application.

To address these limitations, we propose a novel RAG framework, InfoGain-RAG, to filter out irrelevant or even misleading documents, and prioritize the most valuable ones for answer generation. Specifically, we firstly introduce a new metric named Document Information Gain (DIG), which calculates the change in LLM's generation confidence with and without the document augmented. A higher DIG score means the document has higher information value. Then, a multi-task training strategy is designed, enabling one newly added reranking module to predict the DIG score for each document. Only those with a score greater than a certain threshold will be augmented into the LLM for final generation. This reranking module is plug-and-play across diverse models and tasks. Furthermore, it can efficiently handle documents from multiple retrievers by invoking LLM only once for the entire process and the low computational overhead makes it feasible for the real application.

Extensive evaluations on two different types of tasks: open-domain question answering (TriviaQA (Joshi et al., 2017), NaturalQA (Kwiatkowski et al., 2019), and PopQA (Mallen et al., 2023)) and fact verification (FM2 (Eisenschlos et al., 2021)) spanning both proprietary LLMs (GPT, Claude) (Wu et al., 2023; Eisele-Metzger et al., 2024) and open-source models (LLaMA, Qwen, Gemma, DeepSeek) (Touvron et al., 2023; Bai et al., 2023; Team et al., 2024; Liu et al., 2024), demonstrate substantial improvements of InfoGain-RAG over existing methods. Specifically on NaturalQA, it achieves significant gains in Exact Match accuracy: outperforming naive RAG by 17.9%, retriever-optimized RAG by 6.8%, self-reflective RAG by 4.5%, and modern ranking-based RAG by 12.5%. Notably, even compared to the proprietary state-of-the-art reranker GTE-7B (Zhang et al., 2024), our method (335M) still demonstrates a 3.4% improvement. These consistent performance gains extend across TriviaQA, PopQA and FM2, validating our approach's effectiveness across diverse scenarios.

Our main contributions include:

- We introduce a novel metric called **Document Information Gain (DIG)**, to quantify each retrieved document's impact on the LLM's generation confidence. Different from semantic similarity, DIG can more accurately evaluate whether the document is helpful for generating a correct answer;

- We develop a multi-task training strategy, which is used to optimize one reranker added after the retriever, with the aim of fitting the DIG score for each document. This strategy is designed from the exact distinguishing and accurate sorting perspectives, so as to filter out the irrelevant and select the most valuable documents for answer generation.

- Integrating the DIG and the multi-task reranker, we propose **InfoGain-RAG**, a comprehensive framework for enhancing RAG. This framework can improve the quality of generation with both single and multiple retrievers, showing strong adaptability across vairous real-world settings with only an efficient, plug-and-play reranking module.

2 Related Work

RAG has emerged as a promising solution to address fundamental limitations of LLMs. However, a key challenge in RAG systems lies in effectively evaluating and selecting the most valuable documents for answer generation. Existing document selections in RAG broadly follow three approaches:

The first approach optimizes retrievers through training on task-specific datasets. RePlug (Shi et al., 2023) proposed a training pipeline that uses black-box LLM outputs as supervision signals to optimize the retriever, aiming to reduce LLM perplexity. RA-DIT (Lin et al., 2023) proposed a dual instruction tuning framework that jointly optimizes both the LLM and retriever. Though useful, they struggle with multiple retrievers.

The second approach aims to evaluate retrieved documents utility by LLMs's self-reflection capabilities (Asai et al., 2024; Yan et al., 2024). Self-RAG introduces reflection tokens that allow the LLM to adaptively retrieve passages on-demand and critique both the retrieved content and its own generations. While effective in identifying valuable documents, multiple LLM calls introduce substantial computation overhead.

The third approach incorporates a reranker to re-order retrieved documents, typically including the open-source reranker BGE (Chen et al., 2024b) and proprietary GTE-7B (Zhang et al., 2024). BGE is a small encoder initially trained on over 300M text pairs, then supervised fine-tuning on high-quality labeled data, while GTE-7B trains a large long-context LLM to learn the hybrid document representations (both dense and sparse). However, BGE is mainly trained to capture fine-grained semantic relationships, which may fail to select truly helpful documents, and GTE is computationally expensive for practical deployment.

3 Method

In this section, we present InfoGain-RAG to address the key challenges discussed earlier. Our framework consists of two main components: (1) Document Information Gain (DIG), a metric that quantifies a document's contribution to correct answer generation by measuring changes in LLM's generation confidence scores, along with an efficient pipeline for collecting high-quality training data, and (2) a multi-task reranker that combines document relevance classification and ranking objectives to optimize document selection. By incorporating these, our framework enables effective document selection without requiring multiple LLM calls, making it both computationally efficient and practical for real-world applications.

3.1 Document Information Gain

The core of InfoGain-RAG lies in quantifying each document's contribution to correct answer generation through calculating the information gain of each retrieval. This section details our methodology for computing DIG and utilizing it to build high-quality training data. The complete data collection pipeline is presented in Algorithm 1. To compute DIG, we first propose a robust approach for estimating LLM's generation confidence, and then use this estimation to measure the information gain provided by each document.

Algorithm 1 DIG Data Collection Pipeline

Require: Query set $\mathcal{Q}$, Document corpus $\mathcal{D}$, LLM ϕ
Ensure: DIG dataset $\mathcal{T}$

```
𝒯 ← ∅
for each query x ∈ 𝒬 do
    Retrieve candidate documents D_x from 𝒟
    Get confidence p_φ(y|x) (defined in equation 2) without documents
    for each doc d ∈ D_x do
        Get confidence p_φ(y|x, d) with document
        Calculate DIG (defined in equation 3)
        𝒯 ← 𝒯 ∪ {(x, d, DIG(d|x))}
    end for
end for
return 𝒯
```

3.1.1 Answer Generation Probability

A key challenge in computing DIG is estimating the probability of a specific answer. A straightforward way would be to multiply the probabilities of individual tokens as the final confidence score. However, this approach faces two key challenges: First, it suffers from the length bias problem (Shi et al., 2021) where longer sequences tend to receive lower scores as any single low token probability severely impacts the overall score. Second, treating all tokens equally fails to capture the strongest signal for generation quality (Gangi Reddy et al., 2024) which initial tokens often provide. To address these, we propose a two-component approach:

Sliding Window Smoothing: To mitigate the length bias problem, we implement a sliding window smoothing mechanism. For each token t_i in the answer sequence, its smoothed probability is calculated as:

$$p_{\text{smooth}}(t_i) = \frac{1}{W} \sum_{j=i-\lfloor W/2 \rfloor}^{i+\lfloor W/2 \rfloor} p(t_j) \quad (1)$$

where W is the window size and $p(t_j)$ represents the original token probability, obtained by normalizing LLM logits (Yenduri et al., 2024).

Token Importance Weighting: It is reported that initial tokens often carry stronger signals in model generation(Gangi Reddy et al., 2024). Incorporating this observation, we apply higher weights to the first k tokens when computing probability scores, as they typically contain core semantic information for the response. The final formula is as follows:

$$p_\phi(\mathrm{y}|\mathrm{x}) = \prod_{i=1}^{k} (p_{\text{smooth}}(t_i))^{\omega_i \cdot \alpha} \cdot \prod_{j=k+1}^{|\mathrm{y}|} (p_{\text{smooth}}(t_j))^{1-\alpha} \quad (2)$$

where ω_i are the importance weights for the first k tokens, α is a weight hyper-parameter, and $|\mathrm{y}|$ is the answer length.

3.1.2 Calculation of DIG

With a reliable approach to estimate answer generation probability, we now define the calculation of DIG, as shown in Figure 1(NOTE part). Unlike traditional relevance metrics that rely on lexical overlap or semantic similarity, DIG directly measures how much a document improves the LLM's confidence in generating the correct answer.

Formally, given an LLM ϕ, a query x, and its corresponding ground truth answer y, the DIG for a document retrieved $\mathrm{d}_i(\mathrm{d}_i \in \mathcal{D}, \mathcal{D} = \{\mathrm{d}_1, \mathrm{d}_2, \ldots, \mathrm{d}_{|\mathcal{D}|}\})$ is defined as:

$$\mathrm{DIG}(\mathrm{d}_i|\mathrm{x}) \stackrel{def}{=} p_\phi(\mathrm{y}|\mathrm{x}, \mathrm{d}_i) - p_\phi(\mathrm{y}|\mathrm{x}) \quad (3)$$

where $p_\phi(\mathrm{y}|\mathrm{x}, \mathrm{d}_i)$ represents the model's output confidence with both the query and the document, and $p_\phi(\mathrm{y}|\mathrm{x})$ is the query-only confidence.

Based on above, we establish a data collection pipeline that begins by categorizing queries based on the model's baseline performance without retrieved documents, shown in Figure 1 (STEP 1):

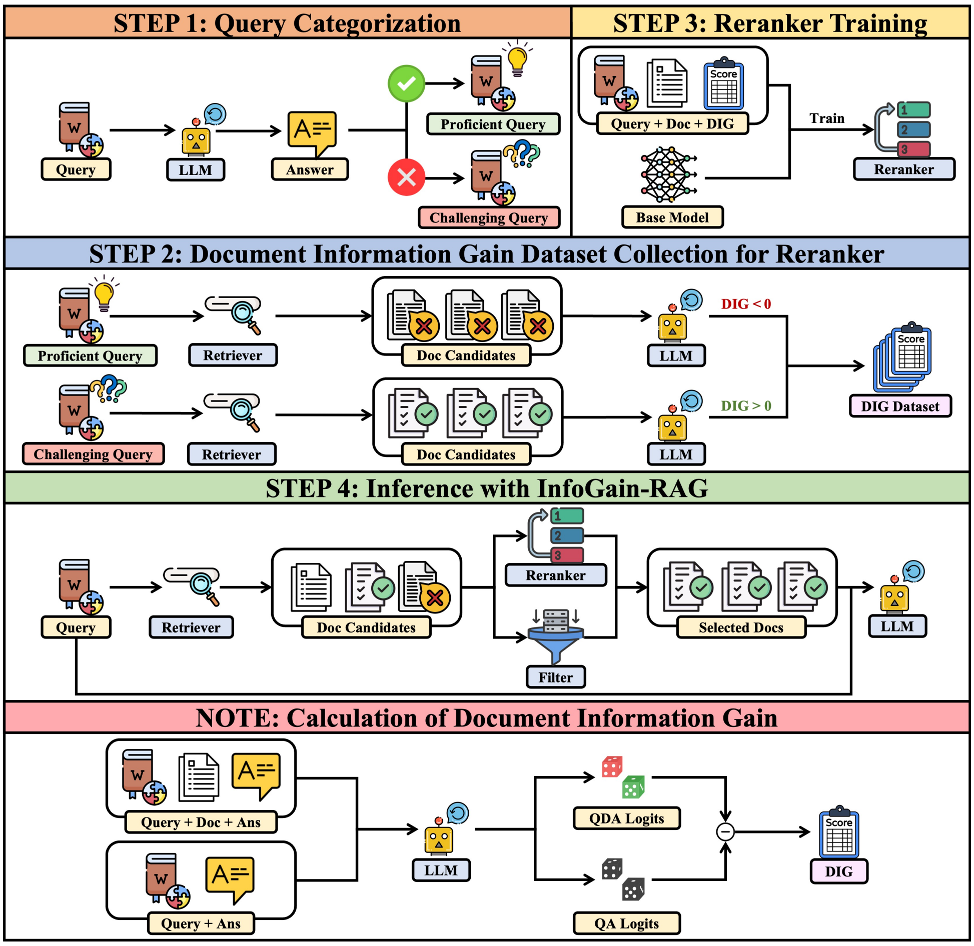

Figure 1: Illustrations of InfoGain-RAG. **STEP 1**: Distinguish proficient queries from challenging ones; **STEP 2**: Retrieve top-k documents for each query and calculate their DIG scores; **STEP 3**: Train the multi-task reranker; **STEP 4**: Inference with InfoGain-RAG; **NOTE**: Calculation of DIG.

- **Model-Proficient Queries**: Queries that the LLM can answer correctly using only its inherent knowledge (i.e., high $p_\phi(\mathrm{y}|\mathrm{x})$). These queries are particularly effective for identifying noisy documents through DIG < 0, while positive DIG samples are naturally rare since external correct information adds little value to already-known answers.

- **Model-Challenging Queries**: Queries that the LLM shows low confidence without external information (i.e., low $p_\phi(\mathrm{y}|\mathrm{x})$). These queries facilitate us to identify helpful documents, as confidence increases (DIG > 0).

Based on DIG, documents are categorized into three groups (see Figure 5):

- **DIG $>$ 0**: Documents that enhance the model's confidence, containing relevant and helpful information that should be prioritized during reranking.

- **DIG $\approx$ 0**: Documents that neither improve nor diminish confidence and occur in two scenarios: (1) the document contains no meaningful information for answering the query, or (2) LLM has already mastered the required knowledge during pre-training, making additional correct information unnecessary.

- **DIG $<$ 0**: Documents that reduce confidence and contain misleading or contradictory information that should be filtered out.

This categorization offers two key advantages: 1) quantitative measurement of document utility through DIG scores, enabling both automatic identification of high-quality documents and precisely filtering noise; and 2) fine-grained document prioritization through continuous DIG scores, which allows optimal document ordering during inference.

By computing DIG across diverse query-document pairs, we create a rich training dataset capturing both absolute relevance and relative importance of documents. This dataset serves as the foundation for training our specialized reranker, as detailed in the following section.

3.2 Multi-task Reranker

Building on DIG-scored training data collected above, we propose a multi-task learning strategy to train our reranker to select the most valuable documents for correct answer generation. The training objective combines Cross-Entropy (CE) loss and Margin loss to filter out noisy content and prioritize highly effective documents based on DIG scores. CE loss enables the model to distinguish between helpful and noisy documents through binary classification, while margin loss optimizes document ordering based on their DIG values. This unified training approach enables our reranker to simultaneously learn discriminative document classification and fine-grained ranking preferences, leading to robust document selection for RAG.

3.2.1 Document Relevance Classification

The first task focuses on the relevance determination of the retrievals through binary classification. Building upon the former collected data, we train the reranker to distinguish documents that have substantial contributions or potential harm to answer generation. Specifically, we employ CE loss to optimize the reranker θ to achieve this objective:

$$\begin{aligned}\min_{\theta} \quad L_{\text{CE}} =& \frac{1}{N}\sum_{i=1}^{N}\Big[-y_i \log(p(\mathrm{x}_i, \mathrm{d}_i)) \\ &- (1-y_i)\log(1-p(\mathrm{x}_i, \mathrm{d}_i))\Big] \\ \text{s.t.} \quad & p(\mathrm{x}_i, \mathrm{d}_i) \in [0,1], y_i \in \{0,1\}, \forall i = 1, \ldots, N\end{aligned} \tag{4}$$

Here, $p(\mathrm{x_i}, \mathrm{d_i})$ represents the predicted probability that document d_i will achieve a positive DIG score for query x_i. The label y_i is determined by our previously computed DIG scores, with $\mathrm{y}_i = 1$ for documents whose score is above upper decision boundary b_1 and $\mathrm{y}_i = 0$ for those below lower decision boundary b_2. These thresholds effectively separate helpful documents from harmful ones. These hyper-parameters selection will be detailed in the experiment section. This classification-based learning not only helps identify useful documents but also facilitates better learning of relative document ordering through the joint training process.

3.2.2 Document Ranking Optimization

The second task focuses on learning relative document importance through pairwise comparison. Inspired by Circle Loss (Sun et al., 2020), we introduce a margin-based learning objective that explicitly models the relative ordering of documents based on their DIG values. Given a query, this objective constrains the maximum score of negative query-document pairs to be lower than the minimum score of positive pairs:

$$\begin{aligned}&\min_{\theta} \quad L_{\text{Margin}} = [\max(s_n) - \min(s_p)]_+ \\ &\text{with} \quad [x]_+ = \max(x, 0)\end{aligned} \tag{5}$$

where s_n and s_p denote scores for pairs with DIG values above b_1 and below b_2 respectively, and θ denotes reranker. To involve all samples in one process, we employ the LogSumExp function to approximate extremal value:

$$\begin{aligned}&\max\{x_1, \ldots, x_n\} = \log(\exp(\max(x_i))) \approx LSE(x_n), \\ &\min\{x_1, \ldots, x_n\} = -\max\{-x_1, \ldots, -x_n\} \approx -LSE(-x_n)\end{aligned} \tag{6}$$

where $LSE(x_n)$ is the LogSumExp function, with detailed derivation provided in Appendix B.1.

Substitute the LogSumExp approximations into equation (5) and yield:

$$\begin{aligned}\min_{\theta} \quad L_{\text{Margin}} &\approx [LSE(\gamma(s_n)) - (-LSE(-\gamma(s_p)))]_+ \\ &\approx \log\left[1 + \sum_{i=1}^{K}\sum_{j=1}^{L}\exp\left(\gamma\left(s_n^j - s_p^i\right)\right)\right]\end{aligned} \tag{7}$$

where γ is a scaling factor controlling the contribution of non-extremal pairs and K and L denote the number of positive and negative document pairs. Detailed derivation is provided in Appendix B.2. Softplus is used to smooth the ReLU function:

$$\text{Softplus}(x) = \log\left(1 + e^x\right) \approx [x]_+ \tag{8}$$

By integrating CE loss and margin loss with weight β, our multi-task training objective enables the reranker to jointly optimize DIG and inter-document relationships:

$$L_{\text{total}} = \beta L_{\text{CE}} + (1-\beta)L_{\text{Margin}} \tag{9}$$

This unified approach produces a robust reranker that considers both absolute document relevance and relative ordering preferences within the retrieved documents, leading to more effective document reranking and filtering for RAG systems (see Figure 2 for empirical study on balancing these two objectives via hyper-parameter β).

During inference, InfoGain-RAG enhances naive RAG pipelines by adding an efficient document reranking step while maintaining low computational overhead, as illustrated in Figure 1 (STEP 4). The process begins with document retrieval, followed by our trained reranker which both reorders

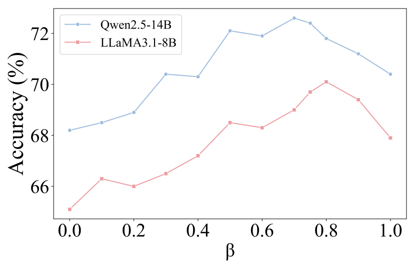

Figure 2: The relationship between the hyper-parameter β and accuracy on TriviaQA, LLaMA3.1-8B achieves optimum at $\beta = 0.8$, while Qwen2.5-14B at 0.7.

documents and filters out those below a quality threshold. The filtered and reranked documents are then passed to LLM for final answer generation while only calling once.

4 Experiment

We evaluate InfoGain-RAG in four experiment series. First, we compare it with modern ranking methods, including the open-source reranker of BGE-Reranker-Large (Chen et al., 2024b) trained on 300M samples, and the state-of-the-art proprietary reranker GTE-7B (Zhang et al., 2024). Second, we compare with retriever optimization approaches like RePlug (Shi et al., 2023) and RADIT (Lin et al., 2023), and self-reflection approaches like Self-RAG (Asai et al., 2024) and CRAG (Yan et al., 2024). Third, we test InfoGain-RAG on combined documents retrieved from Contriever (Lei et al., 2023), BM25 (Robertson and Zaragoza, 2009) and DPR (Karpukhin et al., 2020) to demonstrate its capability to handle multiple retrievers. Last, several ablation studies are conducted to verify the effectiveness from different aspects. The datasets and models we used are publicly accessible.

4.1 Setup

Tasks and Datasets. We experiment on two tasks of four English datasets: (1) **open-domain question answering**, including TriviaQA (Joshi et al., 2017), NaturalQA (Kwiatkowski et al., 2019), and PopQA (Mallen et al., 2023); (2) **fact verification**, FM2 (Eisenschlos et al., 2021). We use the December 2018 Wikipedia dump (Karpukhin et al., 2020) as the retrieval corpus.

Models and Metrics. All evaluations are conducted across both proprietary LLMs (GPT-4o-20241120, ChatGPT-20240125, and Claude-3.5-Sonnet-20241022) and open-source models (LLaMA3.1, Qwen2, Gemma2, DeepSeek-V3, and DeepSeek-R1). We adpot Exact Match (EM) accuracy (Rajpurkar et al., 2016) as the metric. EM provides a strict evaluation of response accuracy while accommodating multiple correct answer formats, as it compares the model outputs with all valid answers provided.

Implementation Details. We sample 110K queries from TriviaQA dataset (with train-test overlap removed) and calculate DIG scores for all collected <*query, answer, document*> triplets using Qwen2.5-7B. The scoring results in three categories: 70K triplets with high positive gain ($>b_1 = 0.5$), 150K triplets with negative gain ($<b_2 = -0.2$), and 1200K triplets showing negligible information gain ($-0.05 \sim 0.05$). From these scored triplets, we create a unified training dataset of 88K samples through different sampling strategies for each loss: for CE loss, we sample balanced query-document pairs with equal numbers of positive and negative samples (68K), while for margin loss, we sample query-document groups (34K) where each query is paired with 3-5 high-DIG documents and augmented with additional negative and negligible documents.

For experimental settings, we implement our reranker using RoBERTa-large (Liu et al., 2019) to rerank the top 100 documents retrieved by Contriever (Lei et al., 2023). Our reranker is trained on an A800 GPU using Adam optimizer with a learning rate of 5e-6 and β value of 0.75. For DIG calculation, we set importance weights ω_i to 0.8 for the first $k = 3$ tokens and use $\alpha = 0.6$ for balancing token probabilities. During inference, we select the top 4 documents and employ a document filtering threshold of 0.2 while retaining all candidates that exceed this threshold. This threshold is slight different from b_1, as the the addition of margin loss would widen the score distribution of valid samples. Notably, to ensure minimal context for generation, we retain at least 2 documents if fewer exceed the filtering threshold.

4.2 Results

We first present InfoGain-RAG's performance with single retriever across different LLMs and benchmarks, comparing it with naive RAG and reranking approaches. We then show its effectiveness in multiple retriever settings. Finally, we demonstrate our method's advantages over self-reflection and retriever-optimization approaches.

Table 1: Performance Comparison of RAG Reranking approaches with single-retriever (Contriever).

Model	TriviaQA				NaturalQA				PopQA				FM2			
	RAG	BGE(550M)[§]	GTE(7B)[◇]	Ours(355M)	RAG	BGE(550M)[§]	GTE(7B)[◇]	Ours(355M)	RAG	BGE(550M)[§]	GTE(7B)[◇]	Ours(355M)	RAG	BGE(550M)[§]	GTE(7B)[◇]	Ours(355M)
Qwen2.5-0.5B	48.5%	48.6%	49.5%	**55.8%**	22.5%	27.3%	29.5%	**35.3%**	26.5%	35.7%	35.3%	**36.5%**	53.0%	52.3%	55.6%	**58.7%**
Qwen2.5-1.5B	50.4%	59.1%	63.3%	**66.3%**	30.7%	39.5%	45.2%	**47.2%**	31.3%	41.3%	**44.2%**	43.0%	69.1%	69.5%	71.1%	**73.9%**
Qwen2.5-7B	52.9%	67.0%	69.5%	**72.1%**	36.3%	41.8%	49.9%	**53.6%**	32.4%	43.4%	43.7%	**47.6%**	72.5%	74.5%	77.8%	**79.9%**
Qwen2.5-14B	56.1%	68.4%	71.1%	**72.9%**	36.0%	42.7%	52.5%	**53.8%**	31.8%	44.1%	45.9%	**49.4%**	72.6%	75.7%	76.4%	**79.4%**
Qwen2.5-32B	58.7%	70.3%	72.0%	**74.7%**	36.4%	42.1%	53.7%	**55.9%**	32.3%	45.5%	48.1%	**50.5%**	73.7%	75.6%	79.0%	**81.2%**
Qwen2.5-72B	59.9%	70.6%	73.4%	**76.3%**	40.3%	44.9%	53.9%	**58.1%**	34.0%	44.8%	49.5%	**51.4%**	73.6%	75.9%	80.4%	**83.4%**
Qwen3-8B	57.9%	67.6%	71.1%	**72.3%**	34.0%	41.5%	50.9%	**52.6%**	32.1%	43.6%	46.5%	**49.1%**	71.4%	76.1%	**80.9%**	80.0%
LLaMA3.1-8B	55.1%	65.5%	67.5%	**70.4%**	33.6%	39.4%	46.9%	**50.7%**	31.7%	41.3%	44.6%	**47.1%**	74.3%	77.6%	79.5%	**81.2%**
LLaMA3.1-70B	54.5%	67.9%	67.4%	**71.3%**	35.1%	39.9%	48.6%	**51.6%**	30.4%	43.0%	47.2%	**47.6%**	77.0%	79.5%	81.1%	**82.4%**
LLaMA3.1-405B	56.7%	69.2%	73.8%	**74.6%**	35.8%	41.5%	52.3%	**53.3%**	30.5%	43.4%	47.3%	**49.5%**	75.9%	77.6%	80.6%	**83.1%**
Gemma-2-9B	54.3%	64.4%	69.0%	**71.3%**	34.3%	39.6%	44.6%	**56.6%**	31.4%	43.9%	45.5%	**49.3%**	75.4%	78.5%	80.9%	**81.5%**
Gemma-2-27B	59.6%	68.5%	70.9%	**74.3%**	37.6%	42.3%	51.5%	**57.4%**	33.1%	45.4%	49.4%	**50.3%**	76.3%	78.4%	**82.1%**	81.6%
DeepSeek-V3	56.0%	68.0%	72.0%	**73.4%**	37.6%	42.5%	50.7%	**55.1%**	30.8%	43.4%	48.6%	**49.7%**	75.7%	77.5%	78.6%	**80.2%**
DeepSeek-R1	60.4%	71.7%	**75.7%**	75.2%	40.8%	44.8%	56.8%	**58.8%**	31.2%	45.3%	51.1%	**51.6%**	77.1%	78.9%	80.3%	**83.8%**
Claude-Sonnet[†]	54.5%	68.4%	70.7%	**73.9%**	36.7%	41.1%	52.4%	**55.2%**	31.6%	43.1%	48.9%	**50.4%**	76.0%	78.4%	**80.9%**	80.8%
ChatGPT[‡]	62.0%	69.0%	72.1%	**74.1%**	37.1%	42.7%	**55.9%**	54.5%	32.0%	43.5%	48.0%	**48.5%**	71.9%	73.2%	75.0%	**75.3%**
GPT-4o[*]	57.2%	69.2%	74.4%	**75.4%**	37.5%	41.6%	53.1%	**57.2%**	31.4%	43.3%	48.3%	**49.2%**	76.6%	75.1%	78.8%	**82.2%**
GPT-4.1[§]	58.6%	70.8%	76.1%	**76.4%**	35.1%	41.7%	55.6%	**56.2%**	30.9%	45.4%	**51.3%**	50.4%	75.2%	77.1%	76.4%	**80.4%**

[§]BGE-Reranker-Large (550M). [◇] Proprietary GTE-Reranker (7B). [†]241022 version. [‡]240125 version. [*]241120 version. [§] 20250414 version.

Comparison to Reranking approaches with Single Retriever. Table 1 compares InfoGain-RAG (355M) against naive RAG, BGE-Reranker (550M) and GTE-Reranker (7B, SOTA) across different models and datasets. As shown in the results, InfoGain-RAG substantially improves over naive RAG and BGE-Reranker, while surpassing the far larger GTE-Reranker in most cases. On TriviaQA, for instance, DeepSeek-V3 achieves 72.0% with GTE-Reranker and 73.4% with InfoGain-RAG, while Qwen2.5-72B reaches 76.3% with InfoGain-RAG, surpassing naive RAG by 16.4%, BGE-Reranker by 5.7%, and GTE-Reranker by 2.9%. Moveover, these improvements hold across both model scales and families - from smaller models like Qwen2.5-1.5B (+15.9% over naive RAG) to larger ones like LLaMA3.1-405B (+17.9%).

Trained on TriviaQA, InfoGain-RAG demonstrates strong generalization ability across different datasets and tasks. It improves Qwen2.5-72B's accuracy on NaturalQA by 17.8% and PopQA by 17.4% over naive RAG, with particularly notable gains on FM2 from 73.6% to 83.4%.

In particular, our reranker achieves these results with just 88K training samples and merely 335M parameters, compared to BGE-Reranker's 300M samples and GTE-Reranker's 7B parameters(see Appendix C for comparisons with the GTE family).

Comparison to Reranking approaches with Multiple Retrievers. InfoGain-RAG maintains consistent superiority with multiple retrievers. As shown in Figure 3, our reranker achieves the best performance on all four tasks. Specifically, it improves by 9.9% over BGE-Reranker on NaturalQA and by 4.9% over GTE-Reranker on PopQA. Additionally, we observe that all rerankers show improvements in the multi-retriever setting compared to the single-retriever setting. Notably, our method achieves the largest performance gains (when comparing multi-retriever to single-retriever settings) on most tasks, with an average improvement of 3.8%. This clearly demonstrates the superior effectiveness of our reranker in multi-retriever scenarios.

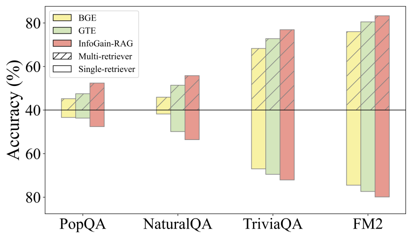

Figure 3: Performance comparison of Qwen2.5-7B across different datasets with single retriever and multiple retrievers.

Comparison with Self-Reflection and Retriever-Optimization approaches. As shown in Figure 4, we evaluate InfoGain-RAG against two types of RAG approaches. For self-reflection, our approach outperforms both Self-RAG(Asai et al., 2024) and CRAG(Yan et al., 2024). With LLaMA2-13B as the base model, InfoGain-RAG achieves 76.2% accuracy on TriviaQA and 51.9% on NaturalQA, surpassing Self-RAG (69.3%, 49.5%) and CRAG (74.5%, 48.2%) while avoiding multiple LLM inference calls. For retriever-optimization, InfoGain-RAG shows substantial improvements using LLaMA-65B, reaching 78.2% on TriviaQA and 54.3% on NaturalQA. This outperforms both RePlug(Shi et al., 2023) (74.9%, 42.3%) and RA-DIT(Lin et al., 2023) (75.1%, 43.9%).

4.3 Ablation Study

In this section, we conduct comprehensive ablation studies to systematically evaluate the critical

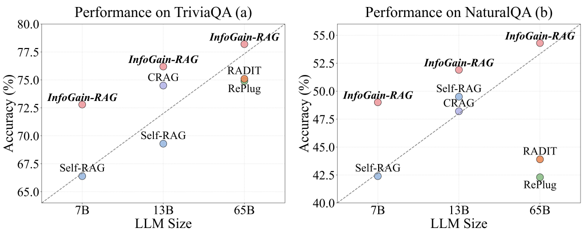

Figure 4: Performance Comparison with self-reflection (7B, 13B) and retriever-optimization (65B) approaches on TriviaQA (a) and NaturalQA (b). We strictly followed the experimental settings of each baseline approach for fair comparison.

components across InfoGain-RAG: 1) examining whether using different base models to generate DIG data will affect the final effect, 2) verifying whether the multi-task learning strategy can bring greater improvement compared to each individual task, and 3) assessing the impact of document filtering during inference.

LLM-agnostic DIG-data Collection. Table 2 demonstrates that InfoGain-RAG's performance remains consistent regardless of which LLM is used for DIG data collection. Despite the changes in the DIG scores of each model due to factors such as structure and size, the trained reranker achieves similar accuracy on TriviaQA. This performance shows that InfoGain-RAG can identify the intrinsic query-document correlations independent of the LLM used for data collection, validating its robustness as a general framework.

Table 2: Compared results of rerankers trained using DIG scores from different base LLMs on TriviaQA.

Model	RAG	Ours (DIG-Qwen)	Ours (DIG-LLaMA)
Qwen2.5-7B	52.9%	**72.1%**	68.8%
Qwen2.5-14B	56.1%	72.9%	**74.2%**
Qwen2.5-72B	59.9%	**76.3%**	75.0%
LLaMA3.1-8B	55.1%	70.4%	**72.1%**
LLaMA3.1-70B	54.5%	**71.3%**	70.2%
LLaMA3.1-405B	56.7%	**74.6%**	73.0%

Single or Multi-task Reranker Training. Table 3 compares the performance differences of single CE or Margin task to the multi-task training. We can see that the combined strategy consistently outperforms individual loss across two types of models. For example, Qwen2.5-72B can get an accuracy of 76.8% with the multi-task training on TriviaQA, but only 73.0% for CE and 71.4% for margin loss. The large improvement demonstrates that the absolute relevance judgments can be combined with the relative rankings to achieve more robust document selection.

Table 3: Performance differences of single CE or Margin task to the multi-task training across models. The testings is conducted on TriviaQA.

Model	Ours (CE loss)	Ours (Margin loss)	Ours (Multi-loss)
Qwen2.5-7B	67.6%	68.2%	**71.8%**
Qwen2.5-14B	70.1%	67.9%	**72.7%**
Qwen2.5-72B	73.0%	71.4%	**76.8%**
LLaMA3.1-8B	68.2%	65.3%	**70.7%**
LLaMA3.1-70B	69.5%	67.1%	**71.4%**
LLaMA3.1-405B	73.6%	70.8%	**74.2%**

Document Filtering during Inference. In table 4 we test the effectiveness of document filtering during inference with the threshold of 0.2. Here, non-filtering means all retrieved documents are ranked without being filtered. It can be observed that peformances are better with filtering than non-filtering. For instance, Qwen2.5-72B improves from 73.6% to 76.8%, and LLaMA3.1-405B gains from 71.2% to 74.6%. These observations jointly confirm that identifying and removing potentially noisy contents is beneficial for final performance.

Table 4: Performance validations of retrieved document filtering operations. All results are tested on TriviaQA.

Model	RAG	Ours (Non-filtering)	Ours (Filtering)
Qwen2.5-7B	52.9%	68.2%	**71.8%**
Qwen2.5-14B	56.1%	71.8%	**72.9%**
Qwen2.5-72B	59.9%	73.6%	**76.3%**
LLaMA3.1-8B	55.1%	67.8%	**70.4%**
LLaMA3.1-70B	54.5%	68.2%	**71.3%**
LLaMA3.1-405B	56.7%	71.2%	**74.6%**

5 Conclusion

In this paper, we present a novel framework InfoGain-RAG to address the critical challenge of RAG about filtering out semantically misaligned and noisy retrieved content. By introducing a principled DIG metric coupled with a multi-task reranker learning strategy, InfoGain-RAG effectively quantifies document utility and optimizes both filtering and reranking processes. Comprehensive experiments across proprietary and open-source LLMs demonstrate substantial improvements across multiple benchmarks while maintaining lower computational overhead compared to existing approaches. The effectiveness and economic applicability of the framework suggest the feasibility of InfoGain-RAG, as it can offer a reliable solution for RAG in partical application.

6 Limitation

While InfoGain-RAG demonstrates strong performance improvements, several limitations warrant discussion. The current implementation has only been tested on text modalities, though it is theoretically extensible to other modalities such as visual or code data. Computational constraints limit the reranker to 335M parameters rather than larger models (7B+), which could offer better performance but may significantly increase inference latency in practical applications. Additionally, the DIG metric, while effective, cannot distinguish factual inaccuracies in retrieved documents, which may require an extra module to address this issue. We hope more efforts can be devoted to addressing these limitations collaboratively in the future.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. 2024. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria.

Jinze Bai, Shuai Bai, Yunfei Chu, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems (NeurIPS)*, 33:1877–1901.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, et al. 2024. Main-rag: Multi-agent filtering retrieval-augmented generation. *arXiv preprint arXiv:2501.00332*.

Ben Chen, Huangyu Dai, Xiang Ma, Wen Jiang, and Wei Ning. 2024a. Robust interaction-based relevance modeling for online e-commerce search. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024*, Berlin, Heidelberg.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Angelika Eisele-Metzger, Judith-Lisa Lieberum, Markus Toews, Waldemar Siemens, Felix Heilmeyer, Christian Haverkamp, Daniel Boehringer, and Joerg J Meerpohl. 2024. Exploring the potential of claude 2 for risk of bias assessment: Using a large language model to assess randomized controlled trials with rob 2. *medRxiv*, pages 2024–07.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 352–365.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. *Preprint*, arXiv:2405.06211.

Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. 2024. FIRST: Faster improved listwise reranking with single token decoding. In *The Association for Computational Linguistics: ACL 2024*, pages 8642–8652, Miami, Florida, USA.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1601–1611, Vancouver, Canada.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Virtual.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. Unsupervised dense retrieval with relevance-aware contrastive pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10932–10940, Toronto, Canada.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, Virtual.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, et al. 2023. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics (TACL)*, 12:157–173.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, et al. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9802–9822, Toronto, Canada.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, et al. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

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

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, et al. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Xuewen Shi, Heyan Huang, Ping Jian, and Yi-Kun Tang. 2021. Reducing length bias in scoring neural machine translation via a causal inference method. In *Chinese Computational Linguistics*, pages 3–15, Cham. Springer International Publishing.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, et al. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Shiqi Yan, Jiachen Gu, Zhuyun, and Zhenhua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Gokul Yenduri, M. Ramalingam, G. Chemmalar Selvi, Y. Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G. Deepti Raj, Rutvij H. Jhaveri, B. Prabadevi, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. 2024. Gpt (generative pre-trained transformer)— a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*, 12:54608–54649.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US.

A DIG Cases

(a)

Query
In what city was Abraham Raimbach born?

Document
At his death, he held a gold medal awarded for his "Village Politicians" at the Paris Exhibition of 1814. He was elected corresponding member of the Académie des Beaux-Arts in 1835. He is buried in St Mary's Churchyard, Hendon. Abraham Raimbach Abraham Raimbach (16 February 1776 in London17 January 1843), was an English line-engraver of Swiss descent. He was born in Cecil Court in the West End of London. Educated at Archbishop Tenison's Library School, he was apprenticed to the engraver J. Hall from 1789 to 1796.

Query
What is the capital of Iceland?

Document
Many fjords punctuate Iceland's 4,970-km-long (3,088-mi) coastline, which is also where most settlements are situated. The island's interior, the Highlands of Iceland, is a cold and uninhabitable combination of sand, mountains, and lava fields. The major towns are the capital city of Reykjavík, along with its outlying towns of Kópavogur, Hafnarfjörður, and Garðabær, nearby Reykjanesbær where the international airport is located, and the town of Akureyri in northern Iceland. The island of Grímsey on the Arctic Circle contains the northernmost habitation of Iceland, whereas Kolbeinsey contains the northernmost point of Iceland.

Query
Who is the author of B^2FH paper?

Document
Phases, a shell model that was necessary for Hoyle's 1954 picture to work as simultaneous ejection of the abundances from each burning phase. Understanding this cultural revolution of computing takes one far in understanding why Hoyle (1954) was forgotten and B^2FH appeared to have been the work that founded stellar nucleosynthesis, as many even claimed. B^2FH paper, named after the initials of the authors of the paper, Margaret Burbidge, Geoffrey Burbidge, William A. Fowler, and Fred Hoyle, is a landmark paper on the origin of the chemical elements published in "Reviews of Modern Physics" in 1957.

(b)

Query
In what city was Giulio Bisegni born?

Document
Giulio tries to scamper back down the ledge, but falls and fractures his ankle. The boss gives chase and Giulio limps to his motorbike, barely escaping. Giulio's mother takes him to the doctor who tells him he'll have to be in a cast for several weeks. Arianna comes over to his place to help him mend, but when he starts telling her what happened and his theory of how Federica wants Sasha to kill her boss, Arianna says he's crazier than ever and storms out, saying she never wants to see him again.

Query
In what city was Gloria Porras Valles born?

Document
her father working as a tailor and her mother, a housewife. She was taught values of cleanliness and accountability. Gloria lived with her grandmother after her parents divorced. Gloria's grandmother told her stories about her great-grandfather who was hung in the mountains of Minas Gerais. Daughter of a slave mother, her grandmother was also a beneficiary of the 1871 Law of Free Birth and, thus, born free. Her grandmother taught her that she needed to work to be free and Gloria decided to focus on combatting racial prejudice. Gloria was once married but separated because she didn't want to live.

Query
What genre is *Enter*?

Document
one type of story best. In later periods genres proliferated and developed in response to changes in audiences and creators. Genre became a dynamic tool to help the public make sense out of unpredictable art. Because art is often a response to a social state, in that people write/paint/sing/dance about what they know about, the use of genre as a tool must be able to adapt to changing meanings. Genre suffers from the ills of any classification system. It has been suggested that genres resonate with people because of the familiarity, the shorthand communication.

(c)

Query
In what city was Aki Hata born?

Document
He has recently begun to bring back the "Guitar Zamurai" character sporadically, to play on its nostalgic appeal. In 2008, he formed a clique of other one hit wonder comedians on the quiz/variety show, Quiz! Hexagon II called "Ippatsuya 2008" (一発屋2008). Yoku Hata Yōku Hata (波田陽区, "Hata Yōku", real name: Akira Hada (波田晃, "Hada Akira"), born June 5, 1975 in Shimonoseki, Yamaguchi Prefecture) is a stand up comedian in Japan. He rose to popularity in 2004 with his character "The Guitar Zamurai (Samurai)" (ギター侍) on the program "The God of Entertainment" (エンタの神様).

Query
Who was the producer of *The Mist*?

Document
Because I grew up listening to his [Alan's] music and never thought that one day I would have such a fantastic opportunity of meeting him. Alan and I spoke a lot, especially about the business side and inner relations of the industry, which was incredibly valuable for me at that early stage of my career. Ostinelli's soundtrack for "The Mist" was released in 2017 by BMG Records. The record contains exclusively Ostinelli's score for the show. Based on the Stephen King's novella of the same name, "The Mist" has been reimagined for television by Christian Torpe and stars Frances

Query
2011 lady gaga album that has edge of glory?

Document
Female Video and Best Video with a Social Message awards at the 2011 MTV Video Music Awards. In the following video, "Judas", she portrays Mary Magdalene, and Norman Reedus plays the title role. The video for "The Edge of Glory" consists mostly of interchanging shots of Gaga dancing and singing on the street and was considered the simplest of her career. In the same year, she released "You and I", which focuses on her trying to get her boyfriend back in Nebraska. She also introduces her male alter ego Jo Calderone in the video.

Figure 5: Retrieved documents of which DIG > 0 (a), DIG ≈ 0 (b), and DIG < 0 (c) for the given query.

Query
Who was the producer of The Imitation Game?

Correct Answer
Teddy Schwarzman

LLM Answer with Retrieved Documents
Teddy Schwarzman

Reranked Documents by InfoGain-RAG Reranker

[Document 1]
Teddy Schwarzman produced The Imitation Game through his production company Black Bear Pictures, who won a competitive bid against Nora Grossman and Ido Ostrowsky who wanted to acquire the script. "The Imitation Game" received many accolades, including Academy Award and BAFTA Award nominations for Best Picture and Best British Film, respectively. Schwarzman, Grossman and Ostrowsky were also nominated for a Producers Guild of America Award. Schwarzman married Ellen Marie Zajac, a New York lawyer whom he met at Duke University, in November 2007 in Montego Bay, Jamaica. They have three children. Teddy Schwarzman Edward Frank "Teddy" Schwarzman (born May 29, 1979) is an American film producer and former corporate lawyer.

[Document 2]
Teddy Schwarzman who wanted to acquire the script. "The Imitation Game" received many accolades, including Academy Award and BAFTA Award nominations for Best Picture and Best British Film, respectively. Schwarzman, Grossman and Ostrowsky were also nominated for a Producers Guild of America Award. Schwarzman married Ellen Marie Zajac, a New York lawyer whom he met at Duke University, in November 2007 in Montego Bay, Jamaica. They have three children. Teddy Schwarzman Edward Frank "Teddy" Schwarzman (born May 29, 1979) is an American film producer and former corporate lawyer. He is the founder, president and chief executive of Black Bear Pictures.

[Document 3]
The US distributor TWC stated that the film would initially debut in four cinemas in Los Angeles and New York, expanding to six new markets on 12 December before being released nationwide on Christmas Day. "The Imitation Game" was released on 31 March 2015 in the United States in two formats: a one-disc standard DVD and a Blu-ray with a digital copy of the film. "The Imitation Game" grossed $91.1 million in North America and $142.4 million in other territories for a worldwide total of $233.5 million, against a budget of $14 million. It was the top-grossing independent film distributed by Schwarzman.

[Document 4] The Imitation Game (play) The Imitation Game is a television play written by Ian McEwan and directed by Richard Eyre, a BBC "Play for Today", first broadcast on 26 April 1980. It is 1940 in Frinton and 19-year-old Cathy Raine turns down a job at the local munitions factory and, much to the consternation of her parents and boyfriend Tony, joins the ATS. She is assigned to a wireless listening station, transcribing Enigma coded morse transmissions from Nazi Germany and makes friends with Mary.

Query
Who was the producer of The Imitation Game?

Correct Answer
Teddy Schwarzman

LLM Answer with Retrieved Documents
The Weinstein Company

Reranked Documents by BGE Reranker

[Document 1]
Adjusted for inflation, the Imitation Game outperformed the Weinstein Company's own Oscar-winning films "The King's Speech" ($88,863 in 2010) and "The Artist" ($51,220 in 2011), which were also released on Thanksgiving weekend. The film expanded into additional markets on 12 December and was released nationwide on Christmas Day. On Rotten Tomatoes, the film has an approval rating of 91% based on 258 reviews, with an average rating of 7.7/10. The site's critical consensus reads, "With an outstanding starring performance from Benedict Cumberbatch illuminating its fact-based story, "The Imitation Game" serves as an eminently well-made entry in the prestige biopic genre.

[Document 2]
his colleagues worked during the war, and Central Saint Martins campus on Southampton Row in London. Other locations included towns in England such as Nettlebed (Joyce Grove in Oxfordshire) and Chesham (Buckinghamshire). Scenes were also filmed at Bicester Airfield and outside the Law Society building in Chancery Lane, and at West London Film Studios. Principal photography finished on 11 November 2013. The bombe seen in the film is based on a replica of Turing's original machine, which is housed in the museum at Bletchley Park. However, production designer Maria Djurkovic admitted that her team made the machine more cinematic.

[Document 3]
The Imitation Game The Imitation Game is a 2014 American historical drama film directed by Morten Tyldum and written by Graham Moore, based on the biography "" by Andrew Hodges. It stars Benedict Cumberbatch as British cryptanalyst Alan Turing, who decrypted German intelligence codes for the British government during the Second World War. Keira Knightley, Matthew Goode, Rory Kinnear, Charles Dance, and Mark Strong also star. The screenplay topped the annual Black List for best unproduced Hollywood scripts in 2011. The Weinstein Company acquired the film for $7 million in February 2014, the highest amount ever paid for U.S. distribution.

[Document 4]
The Imitation Game was announced that Alexandre Desplat would provide the original score of the film. It was recorded by the London Symphony Orchestra at Abbey Road Studios in London. Desplat uses continuous piano arpeggios to represent both Turing's thinking mind and the workings of a mechanical machine. He said of the complexity of the continuity and structure of the score: [W]hen the camera at the end of the film has those beautiful shots of the young boy, the young Alan, and he's meeting with the professor who's telling him his friend Christopher is dead, and the camera is pushing in on him.

Query
What genre is *Inside*?

Correct Answer
horror film

LLM Answer with Retrieved Documents
Horror film

Reranked Documents by InfoGain-RAG Reranker

[Document 1]
title: "Inside (2007 film)"
text: "guy chasing after young girls; it's one of the clichés of the genre. So the first main idea was changing the identity of the bad guy. We wondered what was the motivation for a woman to hunt another woman?" The film was given a budget of 1.7 million Euros. "Inside" received mostly positive reviews. Rotten Tomatoes reported that 83% of critics gave it a positive review. Bloody Disgusting ranked the film twelfth in their list of the 'Top 20 Horror Films of the Decade', with the article saying "One of the most audacious, brutal, unrelenting horror films ever made, "Inside""

[Document 2]
title: "The Inside (film)"
text: "she hits him over the head with a stone, escapes from the warehouse and is hit by a car. "The Inside" was shot on a prosumer HD camcorder in six days. The assault scene was performed in a single 14-minute take. Eion Macken states in the making-of that there are a lot of things in Irish history to inspire horror films; and he eschews banshees, leprechauns, and sídhe in favor of something more abstractly-formed (partly inspired by the worst in man, as well as possibly a paucity of imagination). "The Inside" premiered at Empire, Leicester Square as part of the"

[Document 3]
title: "Inside In / Inside Out"
text: "was the first album he bought, at the age of nine. He credited the record with inspiring him to become a musician. Inside In / Inside Out Inside In / Inside Out is the debut studio album by British indie rock band The Kooks. It was released on 23 January 2006 on Virgin Records. It contains the singles, "Eddie's Gun", "Sofa Song", "You Don't Love Me", "Naïve", "She Moves in Her Own Way", and "Ooh La". The album was produced by Tony Hoffer of record label Virgin Records. Reaching No. 2 in the UK Albums Chart, the album has sold"

[Document 4]
title: "Blood Inside"
text: "took form it did not fit anymore. Then there was the "Heart" Album. Even we thought that was a bit pretentious. So Jørn and I are walking/talking outside one night trying to figure out what it is all about. Thinking only in key words: heart, blood, red, rose, beauty, violence, body, life, death, ambulance, hospital and so forth. Then it struck us: "Blood Inside"." "In terms of genre, it is more rock than electronica. Even a bit psychedelic and/or progressive at times. The mood is kind of sanctified and sad. It has a few frivolous moments as well, but as"

Query
What genre is *Inside*?

Correct Answer
horror film

LLM Answer with Retrieved Documents
Drama

Reranked Documents by BGE Reranker

[Document 1]
title: "The Woman Inside"
text: "Inside. The Woman Inside The Woman Inside is a 1981 (but shot in 1978) drama film made by 20th Century Fox, and directed by Joseph Van Winkle who co-wrote screenplay with Steve Fisher (uncredited). This drama film portrays the actions of a tough Vietnam vet who wants to have a sex-change operation. Her aunt (Joan Blondell) struggles to understand why she would want do such a thing. The film was released after Blondell's death, ending a career spanning more than half a century. The son of Eddy Lawrence Manson now has released a new Project named after the film The"

[Document 2]
title: "Film genre"
text: "A film's genre will influence the use of filmmaking styles and techniques, such as the use of flashbacks and low-key lighting in film noir, tight framing in horror films, fonts that look like rough-hewn logs for the titles of Western films, or the "scrawled" title-font and credits of "Se7en" (1995), a film about a serial killer. As well, genres have associated film-scoring conventions, such as lush string orchestras for romantic melodramas or electronic music for science-fiction films. The basic genres include fiction and documentary, from which subgenres have emerged"

[Document 3]
title: "Pretty on the Inside"
text: "The first 3,000 pressings of the LP featured blue vinyl, while the following pressings were in standard black. "Pretty on the Inside" was received with positive acclaim by many British and American alternative press. In a review by "NME", the album was positively compared to Patti Smith's "Horses", as well as the debut albums of the Ramones, Television, and New York Dolls, and was branded as being in "a class of its own", while Elizabeth Wurtzel wrote in "The New Yorker" that ""Pretty on the Inside" is such a cacophony ... very few people are likely to get through it"

[Document 4]
title: "Interstitial art"
text: "working in a system that clearly labels one shelf for romances, a second shelf for fantasies, and a third shelf for tales of horror? There's no single, obvious answer, because such a novel is interstitial fiction, its essence residing somewhere in between the boundaries of these genres. Or consider the performance artist Laurie Anderson: She might go onstage and sing, tell a spoken-word story, project shadow puppets on a screen, and play a hacked violin whose bow is strung with audio tape. Is she a singer, a monologist, a puppeteer, or some kind of tinkering instrumentalist? Classifying such an act"

Figure 6: Comparison of documents retrieved by InfoGain-RAG reranker and BGE reranker.

B Mathematical Derivations

In this section, we provide detailed mathematical derivations for two key components of margin loss: (1) how LogSumExp (LSE) function approximates the maximum function, and (2) the complete derivation steps of our margin loss formulation based on LSE.

B.1 LSE Approximation of Maximum Function

The LogSumExp function is defined as:

$$LSE(x_1, \ldots, x_n) = \log(\sum_{i=1}^{n} \exp(x_i)) \quad (10)$$

First, we prove that LSE provides an upper bound for the maximum function. For any i:

$$\begin{aligned} LSE(x_1, \ldots, x_n) &= \log\left(\sum_{j=1}^{n} \exp(x_j)\right) \\ &\geq \log(\exp(x_i)) \\ &= x_i \end{aligned} \quad (11)$$

Since this holds for all i, we have:

$$LSE(x_1, \ldots, x_n) \geq \max(x_1, \ldots, x_n) \quad (12)$$

Let $x^* = \max(x_1, \ldots, x_n)$. We can rewrite LSE as:

$$\begin{aligned} LSE(x_1, \ldots, x_n) &= \log\left(\sum_{i=1}^{n} \exp(x_i)\right) \\ &= \log\left(\exp(x^*) \sum_{i=1}^{n} \exp(x_i - x^*)\right) \\ &= x^* + \log\left(1 + \sum_{i:x_i \neq x^*} \exp(x_i - x^*)\right) \end{aligned} \quad (13)$$

Since $x_i - x^* \leq 0$ for all i (with equality only when $x_i = x^*$), and typically $x_i - x^* \ll 0$ for $x_i \neq x^*$, we have:

$$\exp(x_i - x^*) \to 0 \text{ when } x_i - x^* \ll 0 \quad (14)$$

Therefore:

$$\log(1 + \sum_{i:x_i \neq x^*} \exp(x_i - x^*)) \to 0 \quad (15)$$

This yields our final approximation:

$$LSE(x_1, \ldots, x_n) \approx x^* = \max(x_1, \ldots, x_n) \quad (16)$$

The approximation becomes more accurate as the differences between the maximum value and other values increase.

B.2 Derivation of Margin Loss

Starting from the initial margin loss formulation:

$$L_{\text{Margin}} \approx [LSE(\gamma(s_n)) - (-LSE(-\gamma(s_p)))]_+ \quad (17)$$

We can expand this expression:

$$\begin{aligned} &[LSE(\gamma(s_n)) - (-NLSE(\gamma(s_p)))]_+ \\ &= \left[\log \sum_{j=1}^{L} \exp\left(\gamma\left(s_n^j\right)\right) + \log \sum_{i=1}^{K} \exp\left(\gamma\left(-s_p^i\right)\right)\right]_+ \\ &= \left[\log\left(\sum_{j=1}^{L} \exp\left(\gamma\left(s_n^j\right)\right) \sum_{i=1}^{K} \exp\left(\gamma\left(-s_p^i\right)\right)\right)\right]_+ \\ &= \left[\log \sum_{i=1}^{K} \sum_{j=1}^{L} \exp\left(\gamma\left(s_n^j - s_p^i\right)\right)\right]_+ \end{aligned} \quad (18)$$

Finally, using the softplus function to smooth the ReLU operation:

$$L_{\text{Margin}} \approx \log\left[1 + \sum_{i=1}^{K} \sum_{j=1}^{L} \exp\left(\gamma\left(s_n^j - s_p^i\right)\right)\right] \quad (19)$$

This completes the derivation of our margin loss formulation.

C Comparisons with GTE Family

Table 5: Comparative analysis of InfoGain-RAG and various GTE models as rerankers, with Qwen2.5 as the answer generation model on TriviaQA. Results demonstrate InfoGain-RAG's superior performance across all tested configurations.

Method	GTE-1.5B	GTE-7B	GTE-Proprietary	InfoGain-RAG
Qwen2.5-0.5B	45.3%	46.5%	49.5%	**55.8%**
Qwen2.5-1.5B	59.7%	61.7%	63.3%	**66.3%**
Qwen2.5-3B	63.3%	65.6%	65.8%	**68.2%**
Qwen2.5-7B	67.4%	69.2%	69.5%	**72.1%**
Qwen2.5-14B	67.5%	70.5%	71.1%	**72.9%**
Qwen2.5-32B	70.1%	71.9%	72.0%	**74.7%**
Qwen2.5-72B	69.2%	72.2%	73.4%	**76.3%**

D Information of Datasets

TriviaQA[1] consists of 174,000 questions based on Wikipedia pages, with answers and their justifications also determined from Wikipedia, including 138,000 for the training set, 17,900 for the validation set, and 17,200 for the test set. NaturalQA[2]

[1] https://huggingface.co/datasets/mandarjoshi/trivia_qa

[2] https://huggingface.co/datasets/sentence-transformers/natural-questions

is a dataset consists of 307,373 training questions, 7,830 validation questions, and 7,842 test questions. where all questions originate from Google's search records, with answers derived from Wikipedia. PopQA[3] contains approximately 14,000 questions all sourced from the Wikidata database. PopQA focuses on long-tail entities and can effectively assess how well a LLM can grasp infrequent factual knowledge.

FM2[4] is a dataset that contains 10,400 training questions, 1,170 validation qustions and 1,380 test questions, which are designed to test the ability of LLMs to answer simple, factual questions. These questions cover a wide range of topics and are collected from various online sources. The answers to these questions are also provided, making it a valuable resource for training and evaluating question-answering systems.

The December 2018 Wikipedia dump is a comprehensive collection of the content available on Wikipedia up to December 2018. This dump includes nearly 23 millions articles, discussions, and metadata, providing a vast amount of information on a diverse range of topics. It is a valuable resource for natural language processing tasks, such as information extraction, text summarization, and question answering. Researchers and developers can use this dump to train and test their models on a large and diverse corpus of text, helping to improve the performance and accuracy of their systems.

[3] https://huggingface.co/datasets/akariasai/PopQA

[4] https://huggingface.co/datasets/tasksource/fool-me-twice/viewer