Lexical Diversity-aware Relevance Assessment for Retrieval-Augmented Generation

Zhange Zhang[1,2,3] **Yuqing Ma**[1,2†] **Yulong Wang**[4] **Shan He**[4]
Tianbo Wang[2,4] **Siqi He**[5] **Jiakai Wang**[6] **Xianglong Liu**[2,4,6]
Institute of Artificial Intelligence, Beihang University[1]
State Key Laboratory of Complex & Critical Software Environment[2]
Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing[3]
Beihang University[4], Peking University[5], Zhongguancun Laboratory[6]
{zhangezhang,mayuqing}@buaa.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) has proven effective in enhancing the factuality of LLMs' generation, making them a focal point of research. However, previous RAG approaches overlook the lexical diversity of queries, hindering their ability to achieve a granular relevance assessment between queries and retrieved documents, resulting in suboptimal performance. In this paper, we introduce a Lexical Diversity-aware RAG (DRAG) method to address the biases in relevant information retrieval and utilization induced by lexical diversity. Specifically, a Diversity-sensitive Relevance Analyzer is proposed to decouple and assess the relevance of different query components (words, phrases) based on their levels of lexical diversity, ensuring precise and comprehensive document retrieval. Moreover, a Risk-guided Sparse Calibration strategy is further introduced to calibrate the generated tokens that is heavily affected by irrelevant content. Through these modules, DRAG is capable of effectively retrieving relevant documents and leverages their pertinent knowledge to refine the original results and generate meaningful outcomes. Extensive experiments on widely used benchmarks demonstrate the efficacy of our approach, yielding a 10.6% accuracy improvement on HotpotQA.[1]

1 Introduction

The rapid development of large language models (LLMs) has led to widespread deployment across various fields, including conversational assistants (Achiam et al., 2023; Touvron et al., 2023), medical diagnosis (Yang et al., 2024b), and code generation (Wei et al., 2023). However, LLMs rely solely on their training parametric knowledge for inference, which frequently results in issues such as factual hallucinations, outdated information, and low interpretability (Mallen et al., 2022; Huang et al., 2023; Ji et al., 2023), particularly in tasks requiring open-domain knowledge or real-time information (Shuster et al., 2021; Li et al., 2023).

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has demonstrated effectiveness in improving factual accuracy by integrating external retrieval knowledge to enhance LLM generation. Typically, RAG methods first utilize a retriever to acquire relevant documents based on an input query in the retrieval stage, and then extract key information from these documents to augment LLMs in the generation stage. For example, Self-RAG (Asai et al., 2023) trains LLM to assess whether the retrieved documents are related to the input query to improve the retrieval validity. Query-decomposed RAG (Press et al., 2022) improves retrieval performance by decomposing complex queries into simpler sub-queries. CAD (Shi et al., 2023a) utilizes contrastive decoding (Li et al., 2022) to refine each tokens and enhance the generation quality.

However, previous RAG methods struggle to establish a granular relevance assessment between queries and retrieval documents, leading to an under-utilization of external relevant knowledge. As shown in Figure 1, in the **retrieval stage**, existing RAG approaches assess documents relevance based on a single criterion, neglecting the lexical diversity of different fine-grained query components (words or phrases): **(1)** Some components, such as proper names, consistently remain in a fixed form and can be assessed for relevance straightforwardly. **(2)** Some components may be expressed in various lexical forms, such as "occupation" being expressed as "profession", a specific job like "actress", or even as achievements like "Academy Award", complicating the relevance assessment. **(3)** Beyond the original query, supplementary information like "American celebrities" in relation to "Hattie McDaniel's occupation" may aid in rele-

[†] Corresponding author

[1] Our code and data are publicly available at https://github.com/Zhange21/DRAG.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27758–27781
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

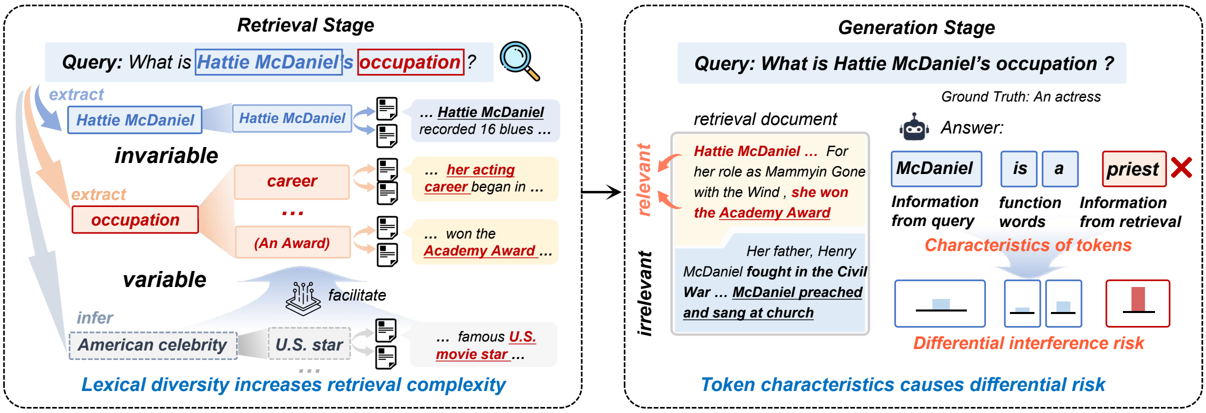
Retrieval Stage
Query: What is Hattie McDaniel's occupation ?
extract
Hattie McDaniel
Hattie McDaniel
... Hattie McDaniel recorded 16 blues ...
invariable
career
... her acting career began in ...
extract
occupation
...
(An Award)
... won the Academy Award ...
variable
facilitate
infer
American celebrity
U.S. star
... famous U.S. movie star ...
...
Lexical diversity increases retrieval complexity
Generation Stage
Query: What is Hattie McDaniel's occupation ?
Ground Truth: An actress
retrieval document
relevant
Hattie McDaniel ... For her role as Mammyin Gone with the Wind , she won the Academy Award
irrelevant
Her father, Henry McDaniel fought in the Civil War ... McDaniel preached and sang at church
Answer:
McDaniel
is
a
priest
Information from query
function words
Information from retrieval
Characteristics of tokens
Differential interference risk
Token characteristics causes differential risk

Figure 1: The challenge of previous methods. In retrieval, the lexical diversity results in differing retrieval complexities. In generation, reductive summarization induces information omission.

vance assessment, adding another layer of lexical diversity. This variation across different query components causes documents with partially similar phrases to be incorrectly considered highly relevant, while documents containing related content but expressed in different ways are overlooked. In the **generation stage**, predicted tokens with varying characteristics might be subject to differential interference from extraneous noise in the retrieved documents. Specifically, tokens representing response entities, which are typically extracted from the retrieved content, are particularly vulnerable to such interference. In contrast, other tokens, such as conjunctions or pronouns, are either minimally affected or semantically insignificant. Therefore, granularly identifying and calibrating these high-risk tokens is necessary and effective for improving generation performance.

In this paper, we propose a Lexical Diversity-aware RAG (DRAG) comprising a Diversity-sensitive Relevance Analyzer (DRA) and a Risk-guided Sparse Calibration strategy (RSC), effectively harnessing the relevant external knowledge. Specifically, to handle the lexical diversity, DRA introduces distinct relevance assessment criteria for different query components, enhancing granular query-document matching. We investigate the varied lexical diversity attributes and assessment mechanism, and prompt the DRA to achieve query decoupling and relevance evaluation. As the degree of lexical diversity increases, the evaluation criteria become more flexible and detailed, ensuring the accuracy and adequacy of the returned documents.

To granularly calibrate the disturbed generation, RSC applies sparse decoding adjustments to the high-risk predicted tokens, thereby minimizing the impact of irrelevant information present in retrieved documents. It introduces the Irrelevance Risk to quantify the irrelevance noise impact on each predicted token and calibrates the decoding process of the high-risk tokens by contrasting with strong-interference noise reference. This strategy enables effective and precise calibration of the high-risk generation while avoiding the high computational cost of extensive calibration. Thus, our method facilitates the retrieval of comprehensive relevant documents through granular assessment and promotes the effective calibration of tokens disturbed by irrelevant noise during generation.

Experiments on commonly used datasets demonstrate that our method achieves significant improvements on open-domain question-answering tasks. In summary, our contributions are:

- We propose a Diversity-sensitive Relevance Analyzer which addresses lexical diversity for the first time to enable fine-grained relevance assessment of retrieved documents, significantly improving the ability of RAG to retrieve semantically relevant information.
- We propose a Risk-guided Sparse Calibration strategy to quantify the impact of irrelevant noise on each token and calibrate the decoding distribution of high-risk tokens, thereby enabling effective and precise enhancement of generation quality.
- Extensive experiments conducted on a widely used benchmark demonstrate the effectiveness of our method. In particular, our approach achieves a 10.6% improvement in accuracy over the second-best method on HotpotQA.

2 Related Work

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) enhances the performance of large language models (LLMs) by incorporating relevant information retrieved from external document repositories during the generation process. Research on RAG has primarily focused on enhancing the model's ability to retrieve and utilize external knowledge through training (Izacard et al., 2023; Borgeaud et al., 2022; Wei et al., 2024).

For instance, Self-RAG (Asai et al., 2023) focuses on improving retrieval accuracy by training LLMs to reflect the relevance of retrieved documents to input queries. RA-DIT (Lin et al., 2023) fine-tunes both the LLM and the retriever, improving performance in both retrieval and generation stages. However, these approaches commonly depend on single-standard relevance assessments, failing to account for the lexical diversity of granular query components, which can hinder their effectiveness in fine-grained retrieval and generation. While query-decomposed RAG methods like Self-Ask (Press et al., 2022) and RQ-RAG (Chan et al., 2024) break down complex queries into subqueries, they primarily address multi-hop questions without accounting for lexical diversity, limiting their effectiveness in mitigating relevance assessment errors due to varied lexical expressions. Additionally, other RAG methods, such as (Shi et al., 2023a) and (Qiu et al., 2024) employ pointwise mutual information and contrastive decoding (Li et al., 2022) to adjust each tokens in generation stage, addressing conflicts between internal and external knowledge within the model.However, these methods overlook the varying degrees of noise interference that different tokens in the output may experience. Calibrating all tokens indiscriminately introduces significant computational overhead and limits the overall performance

In contrast, our method introduces granular relevance assessment criteria that address the lexical diversity of query components, and sparsely calibrate the predicted tokens with high irrelevance risk to mitigate the influence of irrelevant information. This approach facilitates efficient retrieval and utilization of relevant knowledge, without the need for extensive training or reductive summarization.

3 Method

We propose a Lexical Diversity-aware RAG (DRAG) framework, which incorporates a fine-grained relevance evaluation mechanism across both retrieval and generation stages to sufficiently leverage the relevant knowledge. DRAG consists of a Diversity-sensitive Relevance Analyzer (DRA) (Section 3.2) and an Irrelevance Risk-guided Sparse Calibration module (RSC) (Section 3.3), operating with minimal training resources.

3.1 Formalization and Overview

We conform to the standard setup of RAG (Lewis et al., 2020; Asai et al., 2023). In the retrieval stage, DRA module takes the query $\mathbf{x}$ and retrieved documents $\mathbf{D} = \{d_1, d_2, \ldots, d_k\}$ from embedding-based retriever as input, to further extract more related documents. It first decomposes the query into multiple components $\mathbf{C}$ and evaluates the relevance $\mathbf{s}_{i,j}$ between the documents d_i and j-th query component. The top r highest-ranked documents to achieve generation. In the generation stage, the RSC strategy introduces the Irrelevance Risk r_t to quantify the impact of irrelevant noise on each predicted token $\mathbf{y}_t$. Typically, The predicted token at step t based on d_i can be expressed as:

$$\mathbf{y}_t = \mathcal{M}(x, \mathbf{y}_{<t}, \{T^+, T^-\}; \theta_{\mathcal{M}}) \tag{1}$$

where $\mathbf{y}_{<t}$ represents the generation prior to step t, T^+, T^- is the relevant knowledge T^+, and a significant amount of irrelevant noise text in the retrieved documents. Next, we sparsely adjust the decoding distribution of high-risk tokens, thereby mitigating the detrimental effects of irrelevant information and producing reasonable results. The overview process of our DRAG is illustrated in Algorithm 1 in Appendix B.

3.2 Diversity-Sensitive Relevance Analyzer

Existing approaches, which apply a single relevance criterion, fail to capture granular relevance between queries and retrieved documents, disregarding the effects of lexical diversity. Therefore, we propose the Diversity-sensitive Relevance Analyzer (DRA), which decouples the relevance assessment process to accommodate varying degrees of lexical diversity among query components. DRA decomposes the query into distinct components and diversely evaluates the intrinsic relevance between each component and the retrieved documents, employing tailored criteria according to the extent of diversity. This enables a more accurate and fine-grained relevance assessment between the retrieved documents and the full query.

Lexical Diversity-Driven Query Decoupling. To assess relevance granularly, We explore different attributes of lexical diversity and corresponding

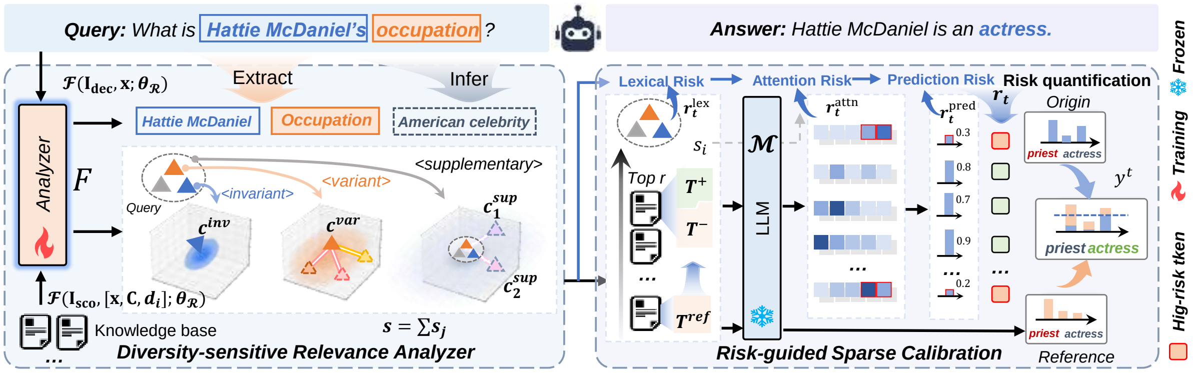

Figure 2: An overview of our DRAG. DRA first decouples different query components based on their lexical diversity and conducts a precise relevance analysis of retrieved documents. RSC then calibrates the model's decoding process by contrasting it with outputs under noise information.

assessment mechanisms, guiding the analyzer to perform query decomposition. Specifically, we categorize the query components into three attributes: A $= \{< Invariant >, < Variant >, < Supplementary >\}$, taking the query "*What is Portland the capital of?*" as an example:

- $< Invariant >$: Components without lexical diversity that are directly extracted from the query. The invariant component of the example is "*Portland*", whose expression cannot be altered.
- $< Variant >$: Components with lexical diversity that are directly extracted from the query. The variant component of the example is "*Capital*", which could be expressed with synonyms such as "*Administrative center*".
- $< Supplementary >$: Components not explicitly mentioned in the query but can be reasonably inferred to supplement relevance assessment, which is not mandatory and demonstrates significant lexical diversity. A possible supplementary component of the example query is "*State or country*".

Based on the attribute setting, we train the DRA module $\mathcal{F}$ to decouple various query components $\mathbf{c} = \{< c_1, a_1 >, < c_2, a_2 >, \ldots, < c_n, a_n >\}$ and assign the attribute $a_j \in$ A to each component c_j. n represents the total number of components, which is determined by $\mathcal{F}$:

$$\mathbf{c} = \mathcal{F}(\mathrm{I_{dec}}, \mathbf{x}; \theta_{\mathcal{F}}) \tag{2}$$

where $\mathrm{I_{dec}}$ is the instruction for $\mathcal{F}$ to generate decoupling components ("*You are an assistant in extracting key components from a given question.*"), and $\theta_{\mathcal{F}}$ is the fine-tuned parameters of $\mathcal{F}$. By decoupling the query and assigning distinct attributes, we can effectively account for the lexical diversity across different components, enabling a tailored relevance assessment.

Granular Relevance Assessment. To accurately assess the relevance between each component and the retrieved documents, we further apply granular assessment criteria tailored to the attributes of different components. As lexical diversity increases, the evaluation criteria are refined and made more stringent, ensuring the precision and comprehensiveness of the retrieved documents. Specifically, we prompt the DRA module with instruction $\mathrm{I_{sco}}$ ("*You are an assistant in scoring documents based on a given question and its components.*") to granularly assess the relevance between retrieval document d and each component c_j:

$$\mathbf{s}_i = \mathcal{F}(\mathrm{I_{sco}}, [\mathbf{x}, \mathbf{C}, d_i]; \theta_{\mathcal{F}}) \tag{3}$$

where $\mathbf{s}_i = \{< s_{i,1}, e_{i,1} >, < s_{i,2}, e_{i,2} >, \ldots, < s_{i,n}, e_{i,n} >\}$, $s_{i,j}$ is the relevance score of the j-th query components to the document d_i, and e_j denotes the corresponding explanation.

For invariant component c_j whose attribute $a_j =< Invariant >$, $\mathcal{F}$ applies strict evaluation criteria σ_1 and assigns a binary score. If retrieved document d explicitly mentions c_j, the score $s_{i,j}$ is set to 1; otherwise, it is 0. For variant and supplementary components c_j with attributes $a_j =< variant >$ or $< supplementary >$, the DRA module $\mathcal{F}$ apply more flexible criteria σ_2, and assigns a continuous score $s_{i,j} \in [0, 1]$. The implementation of σ_1, σ_2 is detailedly discussed in the Appendix B. We compute the weighted sum of the query components $s_i = \sum_{j=1}^{n} w_{i,j} \cdot s_{i,j}$, where the weights for the three component types are set sequentially to 1, α, and β, with $\alpha, \beta \in (0, 1)$. Therefore, $\mathbf{s}_i$ indicates the overall relevance degree between the query $\mathbf{x}$ and the document d_i.

Finally, the top r retrieved documents with the highest overall relevance score $\mathbf{s}_i$ are then sent to the model $\mathcal{M}$ to enhance the generation. The DRA module accounts for lexical diversity across query components, facilitating a fine-grained and accurate relevance assessment. It is fine-tuned with data tailored for query decomposition and granular relevance evaluation, ensuring highly relevant document retrieval.

3.3 Risk-guided Sparse Calibration

To address the varying risk degrees posed by irrelevant information in retrieved documents to different predicted tokens, we propose the Risk-guided Sparse Calibration strategy (RSC). It introduces Irrelevance Risk to quantify the impact of irrelevant noise on each predicted token and sparsely adjusts the decoding process of high-risk tokens to mitigate granular noise, while maintaining minimal computational overhead.

Irrelevance Risk Quantification. Irrelevance Risk aims to quantify and identify which generated tokens are significantly influenced by irrelevant text, encompassing the Lexical Risk, the Attention Risk, and the Prediction Risk.

Specifically, Lexical Risk measures the risk associated with the difficulty of extracting relevant information due to the lexical diversity of the query. As the lexical diversity of a query increases, the complexity of extracting pertinent information also rises, thereby amplifying the risk that the output may be distorted by noise. Lexical Risk is calculated through the different components extracted by DRA module:

$$r_t^{\text{lex}} = \sum_{a_j = inv} \lambda_1 + \sum_{a_j = var} \lambda_2 + \sum_{a_j = sup} \lambda_3 \tag{4}$$

where inv, var, sup denotes the three attributes of the query components $< Invariant >, < Variant >, < Supplementary >$. $\lambda_1, \lambda_2, \lambda_3$ is the corresponding weights for these attributes.

Attention Risk measures the risk arising from the dependency of tokens on retrieved documents of varying relevance. Attention to low-relevance documents can elevate the potential for Irrelevance Risk. It is quantified by integrating by integrating the token's attention distribution $A_{i,t}$ across documents with different relevance scores s_i:

$$r_t^{\text{attn}} = \sum_{i=1}^{n} \frac{A_{i,t}}{1 + s_i} \tag{5}$$

where $A_{i,t} = \sum a_{i,t}$ represents the total attention weights assigned by token $\mathbf{y}_t$ to the document d_i.

Prediction Risk represents the uncertainty of the model when predicting a token based on retrieved information. Higher uncertainty indicates a greater risk of the token being influenced by noise. We quantify this risk by using the maximum prediction probability p_t of the token:

$$r_t^{\text{pred}} = 1 - p_t \tag{6}$$

Lower confidence typically suggests that the token's prediction is strongly affected by noise, resulting in higher uncertainty, and consequently, its Irrelevance Risk is higher. Finally, we integrate the three different levels to represent the Irrelevance Risk of a token:

$$r_t = r_t^{\text{lex}} \cdot r_t^{\text{attn}} \cdot r_t^{\text{pred}} \tag{7}$$

Sparse Token Calibration. By leveraging quantified Irrelevance Risk, we sparsely calibrate high-risk tokens by comparing their output distributions with those generated under irrelevant text conditions, thereby mitigating the noise interference.

Since accurately capturing irrelevant text T^- within documents is challenging, we first construct a reference noise text T^{ref} to simulate the T^- during the generation process. Based on the DRA's assessment results, we select the document with the lowest relevance as the noisy reference $T^{\text{ref}} = \arg\min_{d_i \in \mathbf{D}} s_i$. This document represents irrelevant information with similar vocabulary but different semantics from the query, effectively simulating the distribution of irrelevant text T^-.

Subsequently, we set a threshold δ to identify and adjust the decoding distribution of predicted tokens most affected by irrelevant content, according to their Irrelevance Risk.

$$\begin{aligned} \mathbf{y}_t = &\ \mathcal{M}(x, \mathbf{y}_{<t}, \{T^+, T^-\}; \theta_\mathcal{M}) \\ &- \mathbb{I}(r_t \geq \delta) \cdot \gamma \cdot \mathcal{M}(x, \mathbf{y}_{<t}, T^{\text{ref}}; \theta_\mathcal{M}) \end{aligned} \tag{8}$$

where γ denotes the calibration weight and $\mathbb{I}(\cdot)$ is the indicator function, equal to 1 when the condition holds, and 0 otherwise. Through sparse calibration, this approach effectively reduces the noise interference in high-risk tokens while maintaining minimal computational overhead.

Through granular relevance assessment based on lexical diversity, DRA can retrieve more relevant documents, while the RSC refines token predictions by emphasizing pertinent information, ensuring more meaningful and accurate generation.

4 Experiments and Results

In this section, we conduct extensive experiments on widely used open-domain generation tasks to validate the effectiveness of the proposed method.

4.1 Experimental Settings

Datasets and Evaluation. We evaluate the effectiveness of our method on three tasks: **(1)** ***Short-form generation:*** Following prior work (Asai et al., 2023), we evaluate performance on the PopQA (Mallen et al., 2022) and TriviaQA (Joshi et al., 2017) datasets using factual accuracy, which assesses whether the gold answer is included in the model's generation. **(2)** ***Long-form generation task:*** we employ ASQA (Stelmakh et al., 2022) and utilize 948 queries in dev set for evaluation. We adopt the official metrics of str-em, Rouge-L (Chin-Yew, 2004) (R-L), QA-Hit (Pillutla et al., 2023), QA-EM, and QA-F1 scores. **(3)** ***Multi-hop question answering:*** we assess factual accuracy on HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA (2WikiQA) (Ho et al., 2020).

Baselines. We utilize the model Llama3-8B-Instruct (Dubey et al., 2024) as default generator. We compare our work against two baselines: (1) **Baseline without Retrieval**, where LLMs generate answers directly without using retrieval information. (2) **Baseline with Retrieval**, where an LLM generates output based on the query and top retrieved documents. We consider several advanced RAG methods including Self-RAG (Asai et al., 2023), FLARE (Jiang et al., 2023), QD-RAG (Press et al., 2022), REPLUG (Shi et al., 2023b), SuRe (Kim et al., 2024a) and RECOMP (Xu et al., 2024) for comparison. On the same dataset, we use identical settings across all experiments to ensure a fair comparison.

Implementation Details. For training of DRA module, we utilize a small language model Qwen2-0.5B (Yang et al., 2024a) as the base model to avoid introducing substantial computational demands. For inference, following previous work (Asai et al., 2023), we employ the Contriever-MS MARCO retriever model (Izacard and Grave, 2020) to retrieve documents from Wikipedia for the PopQA and TriviaQA tasks. For ASQA, HotpotQA, and 2WikiMultiHopQA, we utilize the author-provided retrieval documents across all baselines to ensure a fair comparison. More experimental details are outlined in the Appendix Section C.

4.2 Main Results

Comparison against Baselines without Retrieval. Table 1 (top) presents the baselines without retrieval. Our method demonstrated significantly superior performance compared to existing fine-tuned LLMs across all datasets. Notably, on the PopQA dataset, our method achieved a retrieval performance gain of 45.5%. This demonstrates the strong capability of our method in granular retrieval and utilization of relevant information.

Comparison against Baselines with Retrieval. The bottom of Table 1 presents the performance comparison between our method and baseline with retrieval. Our method significantly outperforms existing RAG approaches across nearly all tasks:

(1) ***Short-form generation:*** Our method achieves a 4.9% improvement in accuracy on the PopQA dataset and a 4.4% improvement on TriviaQA, compared to the next best approach. This demonstrates that by incorporating lexical diversity, our method effectively enhances the retrieval and utilization of relevant information. The case study in Table 2 further illustrates the effectiveness of our approach. Due to the inability to accurately assess document relevance, the baseline method is adversely affected by irrelevant information, leading to erroneous outputs. Our method addresses this issue through granular relevance assessment and calibration, thereby enhancing the factual accuracy of the generated content. Further analysis on relevance assessment and lexical diversity can be found in Appendix C.

(2) ***Muti-hop question answering:*** Our method achieves substantial improvements on multi-hop tasks, demonstrating a 10.6% increase in accuracy on HotpotQA and an 10.6% increase on 2WikiMultiHopQA. Given that our training data is solely based on the PopQA and TriviaQA datasets, these results strongly validate the generalization capability of our method. This success can be attributed to the fact that lexical diversity is a common challenge faced by RAG tasks, and multi-hop tasks typically

Methods	Short-form		Multi-hop		Long-form				
	PopQA	TriviaQA	HotpotQA	2WikiQA	ASQA				
	acc (%)	acc (%)	acc (%)	acc (%)	str-em	R-L	QA-Hit	QA-EM	QA-F1
Baseline w/o retrieval									
ChatGPT*	29.3	74.3	-	-	35.3	-	-	-	-
Llama2-7B-Chat	14.7	57.1	14.6	18.4	7.0	29.1	0.4	4.2	6.4
Llama2-13B-Chat	14.7	59.3	18.7	22.3	9.2	12.4	0.8	5.2	7.9
Llama3-8B-Instruct	22.8	69.4	27.7	45.6	24.4	32.8	1.8	13.4	19.5
Baseline with retrieval									
Llama2-7B-Chat	38.2	42.5	16.4	16.9	7.8	24.5	0.7	4.0	5.7
Llama2-13B-Chat	45.7	47.0	26.3	24.5	16.7	28.1	2.9	8.9	12.6
Llama3-8B-Instruct	63.4	73.0	35.8	44.0	27.6	33.8	3.6	17.3	22.9
Self-RAG (Llama2)$_{7B}$	52.4	66.4	27.4	35.9	30.2	35.7	3.3	18.5	24.0
Self-RAG (Llama2)$_{13B}$	55.8	69.3	28.2	36.0	31.6	**35.9**	2.8	**20.2**	26.3
Self-RAG (Llama3)$_{8B}$	50.2	71.4	14.9	32.9	26.7	32.8	2.3	14.4	19.5
FLARE	16.7	53.4	19.5	25.6	13.1	9.3	0.4	9.5	12.8
QD-RAG	35.4	48.7	21.3	36.4	-	-	-	-	-
REPLUG	37.4	60.8	16.2	19.9	20.9	11.2	1.1	14.7	20.2
SuRe	54.8	53.2	18.5	16.6	20.5	5.8	0.7	13.6	19.3
RECOMP	62.8	60.2	25.2	32.0	24.4	8.0	1.3	15.0	21.1
Ours	**68.3**	**77.4**	**46.4**	**54.6**	**35.0**	35.2	**4.0**	20.1	**26.9**

Table 1: State-of-the-art comparison on various open-domain question answering datasets. We re-implement the baselines and report their performance as the maximum value between the original scores and our reproduced results. An asterisk * indicates results copied from (Asai et al., 2023) for reference. A dash "-" denotes results that are either not reported in the original paper or are not applicable. The best performance is highlighted in ***bold***.

Self-RAG	**Ours**
Question: *What star sign is Jamie Lee Curtis? [Ground Truth: "Scorpio"]*	
Jamie Lee Curtis is a Cancer.	Jamie Lee Curtis is a Scorpio, born on November 22, 1958.
Question: *Who was known by his stage name Aladin and helped organizations improve their performance as a consultant? [Ground Truth: "Eenasul Fateh"]*	
James P. Comer	Eenasul Fateh, also known by his stage name Aladin ...

Table 2: Case study on TriviaQA and HotpotQA. Blue text indicates correct output, while red text represents incorrect output.

DRA	**RSC**	**PopQA**	**TriviaQA**	**HotpotQA**
		63.4	73.0	35.8
	✓	65.8 (↑**2.4%**)	74.0 (↑**1.0%**)	36.5 (**0.7%**)
✓		64.1 (↑**0.7%**)	76.5 (↑**3.5%**)	44.9 (↑**9.1%**)
✓	✓	**68.3** (↑**4.9%**)	**77.4** (↑**4.4%**)	**46.4** (↑**10.6%**)

Table 3: Ablation study on the impact of DRA and RSC. Our full model yields superior performance, and each module contributes to the proposed method.

involve more complex queries and external information retrieval requirements. These results further highlight the effectiveness of our lexical diversity-based relevance assessment approach.

(3) ***Long-form generation:*** On the ASQA dataset, the str-em metric, which quantifies the alignment between generated content and ground truth, indicates that our method attained optimal performance, highlighting its precise knowledge extraction and calibration capabilities. In the comprehensive evaluation offering an objective assessment of the generated content through a question-answer framework, our approach demonstrates superior performance in QA-Hit and QA-F1, further validating its generalization capability for complex generation tasks. The slight discrepancy in QA-EM may be attributed to the limited number of irrelevant documents in the official dataset, which may have constrained our model's ability to fully exploit its information assessment capabilities.

4.3 Ablation Study

Ablation of Modules. We first conduct ablation experiments on PopQA, TriviaQA, and HotpotQA to separately investigate DRA and RSC modules in Table 3. The baseline model achieves 35.8% accuracy on HotpotQA. Simply employing the DRA module will bring huge 3.1% accuracy gains on HotpotQA. It reveals that introducing distinct relevance assessment criteria facilitates precise document relevance evaluation, thereby improving the retrieval of relevant information. Additionally, solely deploying the RSC module will bring a 2.4% improvement on PopQA. This indicates that adjusting generation by eliminating the noisy decoding distribution effectively promotes the aggregation and utilization of relevant information. Therefore, the DRA and RSC modules should work synergistically to fully enhance the overall method.

Ablation of Hyper-parameter. We first fix the RSC module and analyze two parameters that regulate relevance assessment in DRA module: the weight of component scores α and β. Figure 3 (a) shows the variation in model accuracy on PopQA as parameters α and β change. It can be observed that, compared to supplementary components, variable components have a more significant impact on our model's performance. As α increases, the accuracy exhibits an inverted U-shaped trend.

Additionally, we evaluated the impact of the calibration threshold δ in the RSC module on model accuracy, as shown in Figure 3(b). As δ increases, the proportion of calibrated tokens decreases, leading to a drop in model performance for most of the range. At the early part of the curve, an increase in δ results in a slight performance improvement, possibly due to the excessive calibration affecting the generation of subsequent tokens.

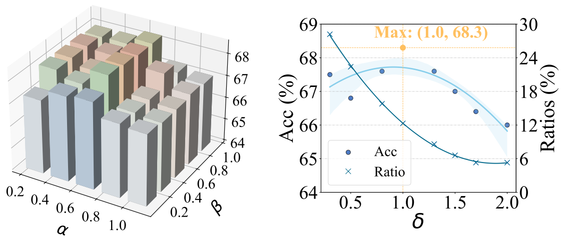

Figure 3: Analysis of hyper-parameters. (a) The influence of component scores weights α and β in DRA. As α increases, the accuracy exhibits an inverted U-shaped trend. (b) As calibration threshold δ increases, the ratio of calibrated tokens decreases, and the model performance decreases over most of the range.

4.4 Deep Analysis

Generation Computation. Additionally, we compare the computation in the generation stage among our method, a decoding-based baseline CAD (Shi et al., 2023a) and the strategy with all tokens calibration via Latency and Throughput. The results in Table 4 demonstrate that the computational overhead introduced by our method during the generation phase is significantly lower than that of other decoding-based RAG methods and only incurs a slight increase in computational cost. Considering the outstanding performance of our method, this increase is manageable.

Methods	s/iter	iter/s	acc
Llama3$_{8B}$	6.02	0.17	63.40
CAD	19.53	0.05	67.33
w/ all tokens calibration	20.93	0.05	68.50
Ours	10.38	0.10	68.26

Table 4: Impact of decoding strategy on performance. "s/iter" refers to the time required to process a single iteration and "iter/s" denotes the number of inferred iterations per unit.

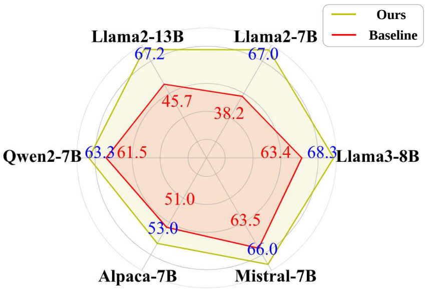

Figure 4: Our method on different baselines.

Performance on Different LLMs. To further validate the adaptability of our method to other LLMs, we select other representative fine-tuned LLMs as generator models and conduct experiments on PopQA. As shown in Figure 4, the results demonstrate that our method outperforms the baseline for all LLMs. It is worth mentioning that our method has significantly improved the Llama models, notably boosting the accuracy of Llama2-7B-Chat from 38.2% to 67.0%. It confirms that our method is compatible with various LLMs and can effectively enhance their performance.

5 Conclusion

In this paper, we introduced Lexical Diversity-aware RAG, a retrieval-augmented generation framework designed to address the limitations of existing RAG methods by incorporating granular relevance assessment and calibrating risky tokens. Our approach effectively improves the retrieval of semantically aligned documents and promotes the aggregation of relevant knowledge, leading to significant advancements in RAG. Extensive experiments on several open-domain question-answering benchmarks validate the superiority of our method.

6 Limitations

A limitation of our approach is its reliance on open-source LLMs designed for next-token prediction, which limits its applicability to models with different architectures. Moreover, for tasks requiring high levels of domain-specific expertise, such as medical report analysis, our method requires additional domain data to enrich the model's specialized knowledge and better capture lexical diversity. In future work, we aim to extend this approach to a broader range of specialized domains.

7 Acknowledgements

This work was supported in part by the Beijing Municipal Science and Technology Project (No. Z231100010323002).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Lin Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hang Gao and Yongfeng Zhang. 2024. Vrsd: Rethinking similarity and diversity for retrieval in large language models. *arXiv preprint arXiv:2407.04573*.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernandez, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

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

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024a. Sure: Summarizing retrievals using answer candidates for open-domain qa of llms. *arXiv preprint arXiv:2404.13081*.

Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024b. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. *arXiv preprint arXiv:2408.01084*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2023. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. *arXiv preprint arXiv:2305.13269*.

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*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. 2023. Mauve scores for generative models: Theory and practice. *Journal of Machine Learning Research*, 24(356):1–92.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. 2024. Entropy-based decoding for retrieval-augmented large language models. *arXiv preprint arXiv:2406.17519*.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F Siu, Byron C Wallace, and Ani Nenkova. 2024. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *arXiv preprint arXiv:2403.00553*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023a. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. *arXiv preprint arXiv:2204.06092*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jonas Waldendorf, Barry Haddow, and Alexandra Birch. 2024. Contrastive decoding reduces hallucinations in large multilingual machine translation models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2539.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. 2023. Magicoder: Source code is all you need. *arXiv preprint arXiv:2312.02120*.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. Instructrag: Instructing retrieval-augmented generation with explicit denoising. *arXiv preprint arXiv:2406.13629*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.

A More Related Work

Contrastive Decoding. Contrastive decoding (Li et al., 2022) is a technique for enhancing open-ended text generation without requiring additional training, achieved by maximizing the difference in log probabilities between an expert LLM and an amateur LLM. This method has demonstrated strong performance across various domains, including reasoning (O'Brien and Lewis, 2023) and neural machine translation (Waldendorf et al., 2024). (Shi et al., 2023a) and (Qiu et al., 2024) employ pointwise mutual information to adjust the output probability distribution, addressing conflicts between internal and external knowledge within the model. (Chuang et al., 2023) propose a decoding strategy that contrasts different layers of the same LLM to more effectively highlight factual knowledge acquired during pre-training. ACD (Kim et al., 2024b) applies contrastive decoding to enhance the RAG in noisy environments.

In contrast to these contrastive decoding approaches, our Irrelevance Risk-guided Sparse Calibration sparsely calibrates decoding process of the tokens significantly influenced by irrelevant text by contrasting with a perturbed reference. Compared to standard contrastive decoding and other contrastive-decoding-based RAG methods, our approach selectively quantifies and calibrates the Irrelevance Risk of generated tokens most impacted by irrelevant content. This results in a significant improvement in the real-time accuracy of the model's output, with only a minimal increase in computational cost, as discussed in Section C.

The Difference between Query Expansion /Decomposition-Based RAG and DRAG. The objectives and processes of our DRA module and Other Query Expansion/Decomposition-Based RAG are fundamentally distinct. Firstly, the Query Expansion/Decomposition-Based RAG method does not account for lexical diversity, which limits its ability to address the challenges of inaccurate relevance assessment caused by diverse lexical expressions. Secondly, our DRA module decouples a query into multiple distinct components (at the word or phrase level) based on lexical diversity, making it applicable to both single-hop and multi-hop questions. In contrast, Query Decomposition-Based RAG methods (such as Self-Ask (Press et al., 2022) and the query decomposition operation in RQ-RAG (Chan et al., 2024)) primarily focus on splitting multi-hop questions into multiple sub-queries for step-by-step reasoning. Moreover, our DRA module can be integrated with Query Expansion/Decomposition-Based RAG methods, allowing for fine-grained decoupling and relevance evaluation of different components within the expanded queries or decomposed sub-queries. This combination enhances the overall performance by leveraging the strengths of both approaches.

The Difference between Robustness-based RAG and DRAG. Several existing methods (Yoran et al., 2023; Yu et al., 2023) enhance the robustness of RAG models under noisy conditions by employing techniques such as noise-aware training, aiming to mitigate the adverse effects of entirely irrelevant information on the generation process. However, in real-world RAG applications, the retrieved documents often contain content that is lexically similar to the query but semantically unrelated to its true answer. Such distractive information is more subtle and difficult to identify, yet it poses a significant hidden challenge to the generation process. Moreover, most prior approaches primarily focus on improving robustness during the generation phase. In contrast, DRAG introduces a correlation-based assessment of Lexical Diversity alongside output calibration, which jointly targets both the retrieval and generation phases. This dual-stage strategy enables DRAG to more effectively suppress the influence of strongly distracting, yet superficially relevant, irrelevant information.

B More Details of Method

The overview process of our DRAG is illustrated in Algorithm 1.

Theoretical Proof of Lexical Diversity The theoretical foundation of DRAG is rooted in the observation that lexical diversity captures a broader spectrum of semantic variations, thereby enhancing both information retrieval and generation processes. Specifically, prior studies (Clarke et al., 2008; Gao and Zhang, 2024) have demonstrated that expression and lexical diversity play a critical role in influencing the complexity and effectiveness of information retrieval. Moreover, language models often struggle to fully capture the nuances and variations inherent in human language, underscoring the importance of incorporating diversity to achieve improved performance (Shaib et al., 2024; Giulianelli et al., 2023). These findings clearly indicate that introducing lexical diversity is highly

Data Type	Training Sample Size	Data Source	Selected Sample Size from Source
Component Extraction	1200	PopQA Training Set	2100
		TriviaQA Training Set	1756
Consistency Evaluation	5543	PopQA Training Set	1990
		TriviaQA Training Set	4553

Table 5: The sources and statistics of training data.

Algorithm 1 DRAG: Lexical Diversity-Aware Retrieval-Augmented Generation

```
Require: Query $x$, Documents $\mathbf{D} = \{d_1, d_2, \dots, d_k\}$, DRA
    Module $\mathcal{F}$, Generation Model $\mathcal{M}$, Risk threshold $\delta$;
    /* The Retrieval Stage */ (§ 3.2)
Generate query components $\mathbf{c} \leftarrow \mathcal{F}(\mathrm{I}_{\mathrm{dec}}, \mathbf{x}; \theta_{\mathcal{F}})$
for each $d_i \in \mathbf{D}$ do
    Analyze relevance $\mathbf{s}_i \leftarrow \mathcal{F}(\mathrm{I}_{\mathrm{sco}}, [x, \mathbf{C}, d_i]; \theta_{\mathcal{F}})$
    Calculate weighted score sum $s_i = \sum_{j=1}^{n} w_{i,j} \cdot s_{i,j}$
end for
Sort documents $\mathbf{D}^{'} = \{d_1^{'}, d_2^{'}, \dots, d_k^{'}\} \leftarrow \mathrm{Sort}(\mathbf{D}; s)$
    according to final score $s$ in descending order
Select top $r$ documents $\mathbf{D}^{\mathrm{rel}} = \{d_1^{'}, d_2^{'}, \dots, d_r^{'}\} =$
    $\{T^{+}, T^{-}\}$, which comprises the relevant text $T^{+}$ and
    the irrelevant text $T^{-}$.
    /* The Generation Stage */ (§ 3.3)
Construct noise text $T^{\mathrm{ref}} = d_k^{'}$
while Generation is not ending do
    Quantify irrelevance risk $\mathrm{r}_t$
    if $\mathrm{r}_t \geq \delta$ then
        $\mathbf{y}_t \leftarrow \mathcal{M}(x, \mathbf{y}_{<t}, \mathbf{D}^{\mathrm{rel}}; \theta_{\mathcal{M}}) - \mathcal{M}(x, \mathbf{y}_{<t}, T^{\mathrm{ref}};$
        $\theta_{\mathcal{M}})$
    else
        $\mathbf{y}_t \leftarrow \mathcal{M}(x, \mathbf{y}_{<t}, \mathbf{D}^{\mathrm{rel}}; \theta_{\mathcal{M}})$
    end if
end while
```

valuable for enhancing RAG's performance in both retrieval and generation stages.

Data Collection for DRA. The instance $(\mathbf{i}, \mathbf{o})$ of the DRA training data consists of two different types: (1) Query decomposition data. The input of DRA $\mathbf{i}$ is the query $\mathbf{x}$ and the instruction $\mathrm{I}_{\mathrm{dec}}$ (*"You are an assistant in extracting key components from a given question."*), the output $\mathbf{o}$ is the decoupled components set C extracted based on $\mathbf{x}$; (2) relevance assessment data. The input $\mathbf{i}$ is a combination of the query $\mathbf{x}$, the retrieved document d, the decoupled components set C, and the output $\mathbf{o}$ is an analysis of the relevance between the retrieved document d and each component in C, including both match scores and explanations. Following the approach of works (Asai et al., 2023), we utilize the state-of-the-art LLM GPT-4 to generate training data for both the components set construction and the relevance assessment processes. Specifically, we prompt GPT-4 with type-specific instructions followed by few-shot demonstrations of the original task input $\mathbf{x}$ to generate the decoupling components set C. Next, we prompt GPT-4 with instructions followed by few-shot demonstrations of the original task input $\mathbf{x}$, the generated components set C, and the retrieved documents D to predict the semantic matching analysis. Manual evaluation indicates that GPT-4's predictions align well with human assessments. We collected a total of 1200 instances for decoupling components set construction data and 5543 instances for progressive relevance assessment data to form the supervised training dataset for the analyzer.

Table 5 presents the sources and statistics of our training data. Specifically, considering the characteristics of open-domain question-answering tasks, we select a subset of data from the PopQA and TriviaQA training sets without losing generality. For each query, we employ type-specific instructions accompanied by 2-3 example prompts to guide GPT-4 in generating component decoupling data. Subsequently, we use Contriever as the retriever to obtain 10 documents similar to the query and prompt GPT-4 to generate relevance analysis results for each document and the query's decoupled components. The prompt and examples for generating component decoupling data are shown in Table 13, while the prompt and examples for generating relevance analysis results and scores are presented in Table 14.

Relevance Assessment Criteria. For the relevance assessment criteria σ_1, a score of 1 is assigned if the document explicitly matches the extracted invariant component; otherwise, it is set to 0. As for the evaluation criteria σ_2, the score is a continuous value between [0,1], reflecting the document's relevance to both the variant and supplementary components of the query. A higher score indicates stronger relevance to the variant component, while the supplementary component is evaluated more leniently, requiring only partial relevance to achieve a score within the same range.

Details of DRA Training. We employ the commonly used cross-entropy loss for supervised fine-tuning of the analyzer:

$$\mathcal{G}_\theta = \arg\min_{\theta} \mathbb{E}_{(\mathbf{i},\mathbf{o})\sim\mathcal{D}}[\mathrm{CE}(\mathcal{G}(\mathbf{o};\theta), \mathbf{i})] \quad (9)$$

Where θ denotes the learned parameter of g and CE refers to the cross-entropy loss.

To minimize additional computational overhead during inference, we employed instruction fine-tuning to train a small language model Qwen-2-0.5B (Yang et al., 2024a) as our DRA module. This training process requires only a small amount of data and computational resources.

C Experimental Details

This section provides a comprehensive analysis of the experimental setup, methodologies, and results to evaluate the effectiveness and generalizability of our proposed approach. We detail the implementation strategies, analyze components with different lexical diversity, and examine the influence of retrieved document volume, different LLMs, and the size of the training data. The findings highlight the efficiency, adaptability, and robustness of our method in addressing challenges such as lexical diversity and relevance evaluation, underscoring its potential for broader applications in retrieval-augmented generation tasks.

More Implementation. By default, we assess relevance using 10 documents per query and select the top 5 for augmentation during model generation. For PopQA and TriviaQA, we follow prior work by using Wikipedia as the retrieval corpus. For HotpotQA, 2WikiQA, and ASQA, we use the official retrieval documents provided by each dataset to ensure fair comparisons.

For comparison with baseline without retrieval, we involve publicly available models like Llama2-7B, Llama2-13B (Touvron et al., 2023), Llama3-8B (Dubey et al., 2024), and private models such as ChatGPT (Ouyang et al., 2022) for comparison.

Comparison with Robustness-based RAG we conduct comparative experiments of two Robustness-based RAG methods RetRobust (Yoran et al., 2023) and Chain-of-Note(CoN) (Yu et al., 2023) on popQA to further illustrate the effectiveness of DRAG. The experimental results, presented in Figure 5, show that our method achieves superior performance. This is because Robustness-based RAG methods primarily focus on enhancing RAG's robustness to noise during the generation phase. In contrast, our approach leverages Lexical Diversity for relevance evaluation and output calibration, competently improving both the retrieval and generation phases, thereby more efficiently mitigating the impact of irrelevant information.

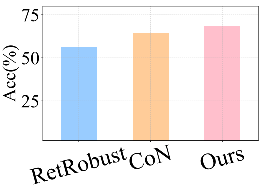

Figure 5: Accuracy comparison with Robustness-based RAG methods.

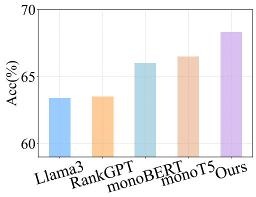

Figure 6: Performance comparison with other reranking strategies.

Comparison with RAG Rerankers. To further validate the effectiveness of our DRA module in assessing the relevance of retrieved documents, we compared it with embedding-based retrieval reranking methods. We selected three typical rerankers, RankGPT (Sun et al., 2023), monoBERT (Nogueira et al., 2019) and MonoT5 (Nogueira et al., 2020), replacing our DRA module with these methods, and applied the RSC based on the resulting re-ranking. Figure 6 presents the experimental results on PopQA, where our approach significantly outperformed both rerankers in terms of accuracy. This further supports our motivation: calculating similarity between the entire query and the retrieved documents to represent relevance is inherently biased, whereas our DRA reasoning and analysis enable a more accurate assessment of relevance.

Performance on Larger Models. We further provide experimental results on larger models (LLaMA2-13B, LLaMA3-70B) in Figure 7. The results show that our method also leads to significant performance improvements on these larger models, demonstrating that the proposed approach and the challenges it addresses in RAG are equally applicable to larger models.

Generalization of DRAG As analyzed in Section 4.2, our training data is solely based on PopQA and TriviaQA, yet our method achieves exceptional performance on other datasets such as ASQA, HotpotQA, and 2WikiMultiHopQA. This demonstrates the strong generalization capability of our approach. Since our method primarily focuses on query de-

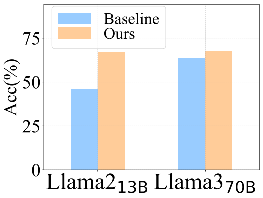

Figure 7: Performance on larger models.

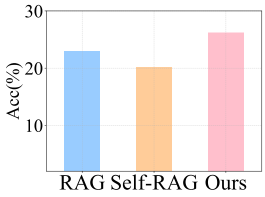

Figure 8: Accuracy comparison on FreshQA.

composition to incorporate lexical diversity for fine-grained relevance assessment, and lexical diversity is a common challenge across different types of queries, our approach is not constrained by factors such as generation format or retrieval database. Thus, it exhibits broad applicability across various generation tasks. In this section, to further validate the generalization of our method, we conducted validation on the FreshQA dataset, a non-Wikipedia style dataset, as shown in Figure 8. The results indicate that our method delivers significant performance improvements on FreshQA, significantly surpassing the baseline Llama3-8B with retrieval (referred to as "RAG"). This highlights the strong generalization capabilities of our approach across different datasets. In fact, the non-Wikipedia style of a dataset does not significantly affect the performance of DRAG. This is because DRAG primarily focuses on decoupling the query within the dataset, and as long as the different components of the query can be accurately identified, DRAG can function effectively. However, it is important to clarify that for more complex or specialized domains, such as medical report analysis or multi-turn dialogues, a certain amount of domain-specific training data is essential to further enhance model performance.

Ablation of Different Components in DRA Module. We validated on the PopQA dataset the performance contributions of the different types of components extracted from the query in DRA to demonstrate that these components help retrieve information with varying lexical diversity, as shown in Table 6. It can be observed that the performance associated with each type of query component significantly exceeds that of the approach without considering lexical diversity (w/o DRA). This demonstrates that our method's consideration of lexical diversity effectively aids in retrieving genuinely relevant documents and generating accurate answers.

Methods	PopQA
Llama3-RAG	63.4
w/o DRA	65.8
Invariant only	67.9
Variant only	66.2
Supplementary only	65.6
Ours	**68.3**

Table 6: Effectiveness of individual components in DRA. Each category contributes to the overall performance of our method.

Lexical Risk	Attention Risk	Prediction Risk	Acc
			64.1
		✓	66.3
	✓	✓	67.6
✓	✓	✓	**68.3**

Table 7: Ablation study of Three Types of Risks in RSC Module. Each risk type contributes meaningfully to output calibration.

Ablation of Three Types of Risks in RSC Module. Additionally, we present in the Table 7 the results of detailed ablation experiments on these three types of risks. The results demonstrate that each risk type contributes meaningfully to output calibration, further validating the necessity of their design.

Case Study about Lexical Diversity In Table 8, we present a case study illustrating how our method addresses the challenge of lexical diversity. In the example, a query and two relevant document passages are analyzed. The baseline retrieval fails to maintain the lexical integrity of "Arcangelo Ghisleri," instead being misled by repeated occurrences of "Arcangelo" in unrelated contexts, resulting in irrelevant documents. Our method, by analyzing lexical diversity and applying refined evaluation criteria, identifies the passage that exactly matches "Arcangelo Ghisleri" and implicitly references "occupation" through the term "journalist." In this case, our method effectively addresses the issue of document relevance caused by lexical diversity, ensuring the retrieval of the correct information and enabling accurate model outputs.

Influence of Retrieved Document Volume. We conduct an analysis of how the number of retrieved documents affects the model's performance. Figure 9 compares the outputs of our method with those of the Llama3 retrieval-based approach across different numbers of retrieved documents. The results demonstrate that our method consistently outperforms the baseline with retrieval in all scenarios. Initially, the relative performance improvement increases as more documents are retrieved. This im-

Question	What is Arcangelo Ghisleri's occupation?
The top retrieval documents of baseline	S. Michele Arcangelo, archangel in Jewish, Christian, and Islamic teachings; Andrea di Cione Arcangelo (1308–1368), Italian painter, sculptor, and architect active in Florence; Antonio di Arcangelo, Italian painter, active in Florence in a Renaissance style, between 1520 and 1538; Arcangelo Califano (1730–1750), baroque composer and cellist; Arcangelo Placenza da Calatafimi, (1390–1460) venerated Italian Franciscan friar and preacher...
Answer of Baseline	Arcangelo Ghisleri is a geographer who created numerous maps of Africa.
The top retrieval documents of Ours	Ghislieri; Michele Ghislieri (1504–1572), also known as Pope Pius V; Ghislieri College; Ghislieri Choir and Consort, Giulio Prandi Ghisleri; **Arcangelo Ghisleri (1855–1938), an Italian journalist.** The Ghislieri, or less commonly Ghisleri, were an ancient Bolognese aristocratic family.
Answer of ours	Arcangelo Ghisleri (1855–1938) is an Italian journalist.

Table 8: Case Study about Lexical Diversity. Our method effectively addresses the issue of document relevance caused by lexical diversity, ensuring the retrieval of the correct information and enabling accurate model outputs.

provement can be attributed to our method's superior evaluation of document relevance, which ensures that the retrieved documents are genuinely relevant to the query, thereby enhancing model performance. The performance gain reaches its maximum when five documents are retrieved, after which a gradual decline is observed. This decline occurs because, beyond a certain threshold, the proportion of truly relevant documents decreases as more documents are retrieved, resulting in diminishing returns. Nonetheless, our method continues to deliver positive performance improvements, even as the number of retrieved documents increases.

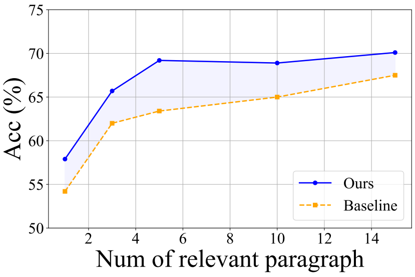

Figure 9: Accuracy with relevant parameter data

Influence of Training Data Size. We analyze the impact of DRA's training data on model accuracy using the PopQA dataset. The training data consists of two parts: data for query decomposition (shown in Figure 10) and data for relevance evaluation (shown in Figure 11). The results show that model performance gradually improves as the amount of two kinds of training data increases, with significant performance gains achieved even with a relatively small dataset. It can be observed that for data related to query decomposition, only around 1,000 samples are needed to achieve significant performance improvement. Similarly, for data related to relevance evaluation, fewer than 5,000 samples are sufficient to realize a substantial performance boost. This highlights the efficiency and low resource demands of our approach. The slight performance drops in the curve may be attributed to domain differences between the training data and PopQA, which could be explored in future work by increasing the diversity of the training data.

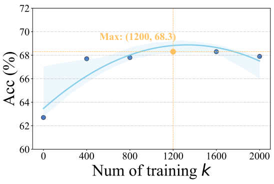

Figure 10: Influence of query decomposition data.

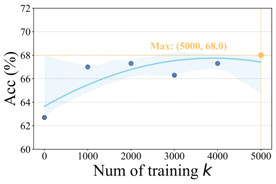

Figure 11: Influence of relevance assessment data.

Influence of Decoding Strategy. We further compare the effects of different contrastive decoding strategies, as shown in Table 9. Despite calibrating only 15% of the tokens, we achieve performance comparable to full-token contrastive decoding, demonstrating the efficiency of the Irrelevance Risk-guided sparse calibration. Moreover, our method outperforms both the noise-free contrastive decoding setting and the contrastive decoding setting with completely irrelevant noise references. This suggests that our strongly interfering noise documents better simulate the irrelevant noise encountered in retrieved documents under real-world conditions, leading to more effective calibration of the model's generation.

Training efficiency A significant advantage of our approach is its ability to deliver substantial per-

Methods	PopQA	TriviaQA
w/o irrelevant docs	67.3	76.1
w/ fixed irrelevant doc	67.4	75.8
w/ all tokens calibration	68.5	77.0
Ours	**68.3**	**77.4**

Table 9: Impact of decoding strategy on performance. Our method achieves performance comparable to full-token calibration with only 15% calibration.

formance gains without requiring extensive training resources or time. Table 10 compares the data size and training cost of our method with the training-based RAG approach (Self-RAG), demonstrating the low resource consumption of our method.

Methods	Number of Instances	Model Size
Self-RAG	145,619	7B
Ours	**6,743**	**0.5B**

Table 10: Comparison of number of instances, model size, and training costs.

D Human validation of Training Data

In this section, we conducted further validation to ensure that the generated data reasonably reflects our design regarding lexical diversity. We sampled a subset consisting of 100 query decoupling data points and 500 relevance assessment data points for cross-validation by human evaluators. The evaluators included three data annotation experts and two of our co-authors. They were tasked with evaluating the data quality based on both the correctness of the data and the logical consistency of the GPT reasoning process. Each evaluator's assessment was cross-validated by two other evaluators. The results showed that the generated data largely meets the reasoning requirements of DRAG. We provide 10 examples of query decoupling data and relevance assessment data. From these samples, it is evident that our generated data is generally accurate and reasonable, effectively capturing the logical reasoning processes involved in query decoupling and relevance assessment.

In fact, our method does not rely on highly precise and strictly accurate generated data, eliminating the need for extensive manual annotation and validation. This is also one of the key advantages of our approach. The purpose of using generated data for training is to equip our first module (DRA) with the ability to perform query decoupling and relevance analysis based on lexical diversity. As long as the generated data reasonably reflects the designed analysis and processing framework, minor inaccuracies in the data will not significantly impact the performance of our method.

E More Case Study.

We present a case study illustrating how our method addresses the challenge of lexical diversity in Table 15. In the example, a query and two relevant document passages are analyzed. The baseline retrieval fails to maintain the lexical integrity of "Arcangelo Ghisleri," instead being misled by repeated occurrences of "Arcangelo" in unrelated contexts, resulting in irrelevant documents. Our method, by analyzing lexical diversity and applying refined evaluation criteria, identifies the passage that exactly matches "Arcangelo Ghisleri" and implicitly references "occupation" through the term "journalist." This ensures the retrieval of the correct document, enabling accurate model outputs.

Additionally, we provide further results on the PopQA, TriviaQA, HotpotQA, 2WikiMultiHopQA, and ASQA datasets, as well as an example of an irrelevant paragraph along with our corresponding result.

input: What sport does Nobuhiro Ishizaki play?
output: Nobuhiro Ishizaki (invariant); Sport (variant).
system: You are an assistant in extracting key elements from a given question.

input: What was Cleopatra's favourite colour?
output: Cleopatra (invariant); favourite colour (variant).
system: You are an assistant in extracting key elements from a given question.

input: What disease claimed the life of notorious American gangster, Al Capone?
output: Al Capone (invariant); disease claimed the life (variant); notorious American gangster (supplementary)
system: You are an assistant in extracting key elements from a given question.

input: The Torres Strait lies between Australia and which island?
output: The Torres Strait (invariant); lies between Australia (variant); which island (initial)
system: You are an assistant in extracting key elements from a given question.

input: In Britain, by an act of 1541 (repealed in 1845) artificers, labourers, apprentices and servants were forbidden to play which sport at any time except Christmas, and then only in their master's house and presence, or be fined 6s. 8d?
output: Britain, 1541 act (invariant); forbidden sport, Christmas, master's house and presence (variant); artificers, labourers, apprentices, servants (supplementary)
system: You are an assistant in extracting key elements from a given question.

input: Question: What sport does Roland Zajmi play?
Element: Roland Zajmi (invariant); Sport (variant).
Paragraph: Roland Zajmi (born 6 November 1973 in Tiranë) is an Albanian retired footballer who played the majority of his career in Greece with Proodeftiki, Atromitos, Kassandra, Agios Dimitrios and Pierikos. He also played for Dinamo/Olimpik Tirana and Partizani Tirana in Albania, Apollon Limassol in Cyrpus as well as the Albania national team.
output: Roland Zajmi (invariant): 1.0 (Roland Zajmi is explicitly mentioned)
Sport (variant): 1.0 (Football is a sport and it's well detailed)
system: You are an assistant in scoring paragraphs based on a given question and its associated elements.

input: Question: What is Quebec City the capital of?
Element: Quebec City (invariant); Capital (variant); Province (supplementary).
Paragraph: Culture of Quebec City, the provincial capital (albeit dubbed La capitale nationale, national capital, in French), is best known as the first permanent settlement and the only fortified city in North America north of Mexico. The old city, partially encircled within the centuries-old walls, has a European flair.
output: Quebec City (essential): 1.0 (Quebec City is explicitly mentioned) Capital (initial): 1.0 (Provincial capital is related to capital) Province (supplementary): 0.0 (Province is not mentioned)
system: You are an assistant in scoring paragraphs based on a given question and its associated elements.

Table 11: Examples of training data (Part 1).

input: Question: Who is the author of Occupy?
Element: Occupy (invariant); Author (variant).
Paragraph: Occupy (book) The book includes an editor's note, a brief section providing legal advice for American Occupy activists, and five sections written by Chomsky himself. Occupy opens with an editor's note written by Greg Ruggiero, in which he explains the basics to Chomsky's views on the Occupy movement, drawing quotes from his various public speeches in order to do so. Ruggiero also discusses Occupy's success in the United States, stating that it has helped to change media discussions by introducing terms like "the 99%" into popular discourse and also by bringing national attention to the plight of the impoverished. He remarks that the protest movement has."
output: Occupy (essential): 1.0 (Occupy is explicitly mentioned)
Author (initial): 0.8 (Author Chomsky is mentioned but not directly as the author)
system: You are an assistant in scoring paragraphs based on a given question and its associated elements.

input: Question: Rimini Miramare airport has been renamed in honour of which noted film director?
Element: Rimini Miramare airport (invariant); has been renamed in honour of (variant); noted film director (variant).
Paragraph: Federico Fellini by Nino Rota during the funeral ceremony. Five months later, on 23 March 1994, Fellini's widow, actress Giulietta Masina died of lung cancer. Fellini, Masina and their son, Pierfederico, are buried in a bronze sepulchre sculpted by Arnaldo Pomodoro. Designed as a ship's prow, the tomb is located at the main entrance to the Cemetery of Rimini. The Federico Fellini Airport in Rimini is named in his honour. Fellini was raised in a Roman Catholic family and considered himself a Catholic. As an adult, however, he avoided formal activity in the Catholic Church. Films by Fellini included Catholic themes;
output: Rimini Miramare airport (invariant): 1.0 (Rimini Miramare airport is explicitly mentioned as T̈he Federico Fellini Airport in Rimini)̈ has been renamed in honour of (variant): 0.6 (The paragraph does not explicitly mention the renaming, but it implies it by stating the airport in Rimini is named after Fellini) noted film director (variant): 1.0 (The paragraph mentions that the airport is named after Italian filmmaker Federico Fellini) No supplementary element provided.
system: You are an assistant in scoring paragraphs based on a given question and its associated elements.

input: Question: What is Delhi the capital of?
Element: Delhi (invariant); Capital (variant); Country (supplementary).
Paragraph: Delhi Delhi, officially known as the National Capital Territory (NCT) of Delhi, is a city and a union territory of India containing New Delhi, the capital of India. It is bordered by the state of Haryana on three sides and by Uttar Pradesh to the east. The NCT covers an area of 1484 km2. According to the 2011 census, Delhi's city proper population was over 11 million, the second-highest in India after Mumbai, while the whole NCT's population was about 16.8 million. Delhi's urban area is now considered to extend beyond the NCT boundaries, and include the neighbouring satellite cities of Ghaziabad, Faridabad, Gurgaon and Noida in an area called the National Capital.
output: Delhi (invariant): 1.0 (Delhi is explicitly mentioned) Capital (variant): 1.0 (Capital is explicitly mentioned and discussed) Country (supplementary): 1.0 (Country, India, is explicitly mentioned)" **system:** You are an assistant in scoring paragraphs based on a given question and its associated elements.

Table 12: Examples of training data (Part 2).

A prompt designed to guide GPT-4 for generating data of query decoupling.

You will receive a series of instructions but do not answer them directly. Follow these steps:

1. For each instruction, extract key elements for retrieval from the instruction. Each element should be a word or a phrase. In the output, separate each element extracted from the same instruction with a semicolon.

2. Categorize the extracted elements into three types and annotate them in parentheses after every extracted element in the output:

- Invariant elements: Explicitly extracted from the instruction without any changes. Subsequent verification of the retrieved document will require that the retrieved document explicitly contain Invariant elements. Therefore, the extracted Invariant elements must be able to represent the most basic information in the original instruction, such as personal names, place names, and other entities. Limit the number of invariant elements to a maximum of 1.

- Variant elements: Explicitly extracted from the instruction, but the retrieved documents don't need to explicitly include these elements, just be implicitly related to them. No quantity limit. Ensure the combination of all Invariant and Variant elements fully represents the original instruction.

- Supplementary elements: Inferred based on the instruction's context to clarify the search, not explicitly mentioned, and the retrieved documents don't need to explicitly include them, just be implicitly related to them. Supplementary elements are not necessary and should be minimized in number. They can only be added if the combination of Invariant and Variant elements is not clear when used for retrieval.

Finally, list the key elements separately for each instruction, indicating their category (invariant, variant, or supplementary).

Example and output format:

Example 1:

Instruction: FDA gives fast track status to AstraZeneca's diabetes drug Farxiga.

Elements: Farxiga (invariant); Fast track status (variant); FDA (variant); AstraZeneca (variant); Diabetes drug (variant); Drug approval (supplementary); Regulatory process (supplementary); Pharmaceuticals (supplementary);

Example 2:

Instruction: Does a surgical mask help avoid COVID-19?

Elements: COVID-19 (invariant); Surgical mask (variant); Help avoid (variant)

Table 13: A prompt designed to guide GPT-4 for generating data of query decoupling. The original query is decoupled into three kinds of components: invariant, variant, and supplementary.

A prompt designed to guide GPT-4 for generating data of relevance assessment.

You are tasked with evaluating the relevance of some given paragraphs to a specific question based on the following elements: invariant, variant, and supplementary.

Scoring Standard:

Score of Invariant Element: Check if the invariant element is explicitly mentioned in the paragraph. If it is, assign an invariant score of 1.0; otherwise, assign a score of 0.0.

Score of Variant Element or Supplementary Element: Consider how well the paragraph discusses or relates to the concept or entity represented by the variant element or supplementary element. Assign a variant score ranging from 0 to 1, where 1.0 indicates a strong relevance and 0 indicates no relevance.

Output Format:

paragraph_id: <ID>

invariant_score: <Score>

variant_score: <Score>

supplementary_score: <Score>

Example:

Question: What is Henry Feilden's occupation? Elements: Henry Feilden (Invariant); Occupation (Variant)

Paragraphs: { id: 11341299, title: Henry Feilden (Conservative politician), text: Henry Master Feilden (21 February 1818 – 5 September 1875) was an English Conservative Party politician. }

Output:

id: 11341299. Invariant Score: 1.0 (Henry Feilden is explicitly mentioned). Variant Score: 0.8 (Politician is related to occupation, but not fully detailed). Supplementary Score: 0.0 (No supplementary element provided)

Table 14: A prompt designed to guide GPT-4 for generating data of relevance assessment.

Question: What is Kyaw Swe's occupation?

Retrieval Document (Baseline):

Paragraph 1: S. Michele Arcangelo, archangel in Jewish, Christian, and Islamic teachings ; Andrea di Cione Arcangelo (1308–1368), Italian painter, sculptor, and architect active in Florence ; Antonio di Arcangelo, Italian painter, active in Florence in a Renaissance style, between 1520 and 1538 ; Arcangelo Califano (1730–1750), baroque composer and cellist ; Arcangelo Placenza da Calatafimi, (1390–1460) venerated Italian Franciscan friar and preacher ; Arcangelo Canetoli (1460–1513), venerated Catholic priest ; Arcangelo Cascieri (1902–1997), influential sculptor, major figure in Boston Architectural College in Boston, Massachusetts ; Arcangelo di Cola (active 1416-1429) Italian late-Gothic painter ; Arcangelo Corelli (1653–1713), Italian violinist and composer of Baroque music ; Arcangelo Ghisleri (1855–1938), geographer who created numerous maps of Africa ; Arcangelo Guglielmelli (c. 1650–1723), Italian

Paragraph 2: Arcangelo Guglielmelli (c.1650—1723) was an Italian architect and painter, active in his native Naples, Italy, in a late-Baroque style. He was involved in the building and reconstruction of churches, many of which had been damaged by the earthquakes of 1688 and 1694.

Answer (Baseline): Arcangelo Ghisleri is a geographer who created numerous maps of Africa.

Retrieval Document (Ours):

Paragraph 1: Ghislieri ; Michele Ghislieri (1504–1572), also known as Pope Pius V. ; Ghislieri College ; Ghislieri Choir and Consort, Giulio Prandi Ghisleri ; **Arcangelo Ghisleri (1855–1938), an Italian journalist.** The Ghislieri, or less commonly Ghisleri, were an ancient Bolognese aristocratic family.

Paragraph 2: S. Michele Arcangelo, archangel in Jewish, Christian, and Islamic teachings ; Andrea di Cione Arcangelo (1308–1368), Italian painter, sculptor, and architect active in Florence ; Antonio di Arcangelo, Italian painter, active in Florence in a Renaissance style, between 1520 and 1538 ; Arcangelo Califano (1730–1750), baroque composer and cellist ; Arcangelo Placenza da Calatafimi, (1390–1460) venerated Italian Franciscan friar and preacher ; Arcangelo Canetoli (1460–1513), venerated Catholic priest ; Arcangelo Cascieri (1902–1997), influential sculptor, major figure in Boston Architectural College in Boston, Massachusetts ; Arcangelo di Cola (active 1416-1429) Italian late-Gothic painter ; Arcangelo Corelli (1653–1713), Italian violinist and composer of Baroque music ; Arcangelo Ghisleri (1855–1938), geographer who created numerous maps of Africa ; Arcangelo Guglielmelli (c. 1650–1723), Italian

Answer (Ours): Arcangelo Ghisleri (1855–1938) is an Italian journalist.

Table 15: An example of Llama3-8B and our method on PopQA. The Case illustrates that our method addresses the challenge of lexical diversity.

Question: What is Kyaw Swe's occupation?

Self-RAG: Kyaw Swe is a Burmese politician who currently serves as a House of Nationalities member of parliament for Magway Region No. 11 constituency.

LLaMA-3: Kyaw Swe has had multiple occupations throughout his career. Initially, he was a school teacher. Later, he became a politician and served as a member of the National League for Democracy Party. He was elected as a representative in the Amyotha Hluttaw from Magway Region No. 11 parliamentary constituency in the 2015 Myanmar general election.

Ours: There are five individuals with the name Kyaw Swe:

Politician 1: Kyaw Tint Swe (born 19 March 1945) He is a former Minister for the Office of the State Counsellor of Myanmar.

Teacher/Politician 2: Kyaw Swe (born 1 December 1960) He graduated as a school teacher and is also a politician.

Politician 3: Kyaw Swe (born 1 December 1979) He is a House of Nationalities member of parliament for Magway Region No. 11 constituency and a member of the National League for Democracy.

Minister: Lt. General Kyaw Swe (no birthdate mentioned) He was the former Minister of Home Affairs of Myanmar, in office from 2016 to 2020, and previously served as Commander of Southwest Command in Pathein, Ayeyarwady Region.

Actor: Kyaw Swe (10 February 1924 – 15 August 1982) He was a Burmese actor and film director.

These individuals have distinct occupations and dates of birth, indicating that they are not the same person.

Table 16: Examples of Self-RAG, Llama3-8B, and our method on PopQA.

Question: Who was the man behind The Chipmunks?

Self-RAG: The Chipmunks were created by Ross Bagdasarian Sr.

LLaMA-3: The man behind The Chipmunks was Ross Bagdasarian Sr., also known as David Seville. He was an American singer-songwriter and actor who created the novelty musical group The Chipmunks (now better known as "Alvin and the Chipmunks"). He produced and directed "The Alvin Show", which aired on CBS in 1961-62, and wrote and recorded many of the group's hit songs, including "Witch Doctor" and "The Chipmunk Song (Christmas Don't Be Late)". Bagdasarian died of a heart attack in 1972, bringing all Chipmunk activity to an abrupt halt. His legacy lives on through the beloved characters he created.

Ours: Ross Bagdasarian Sr. was the man behind The Chipmunks. He was an American singer-songwriter and actor who created the novelty musical group in 1958. Initially a stage and film actor, he rose to prominence with the songs "Witch Doctor" and "The Chipmunk Song (Christmas Don't Be Late)", which both became "Billboard" number-one singles. He produced and directed "The Alvin Show", which aired on CBS in 1961–62. He created the group consisting of three singing animated anthropomorphic chipmunks: Alvin, Simon, and Theodore, managed by their human adoptive father, David (Dave) Seville, who was actually Bagdasarian's stage name. The group was initially known as David Seville and the Chipmunks or simply the Chipmunks before being renamed Alvin and the Chipmunks.

Table 17: Examples of Self-RAG, Llama3-8B, and our method on TriviaQA.

Question: What distinction is held by the former NBA player who was a member of the Charlotte Hornets during their 1992-93 season and was head coach for the WNBA team Charlotte Sting?
Self-RAG: Muggsy Bogues
LLaMA-3: Muggsy Bogues, a former NBA player, was a member of the Charlotte Hornets during their 1992-93 season and later became the head coach of the WNBA team Charlotte Sting.
The texts can be categorized into two groups: those related to the Charlotte Hornets (NBA) and those related to the Charlotte Sting (WNBA). The instruction requires finding the connection between the Charlotte Hornets and the Charlotte Sting, which is Muggsy Bogues.
Ours: Tyrone Curtis "Muggsy" Bogues, the shortest player ever to play in the National Basketball Association, was a member of the Charlotte Hornets during their 1992-93 season and later served as head coach of the WNBA team Charlotte Sting.

Table 18: Examples of Self-RAG, Llama3-8B, and our method on HotpotQA.

Question: Who is the mother of the director of film Polish-Russian War (Film)?
Self-RAG: The mother of the director of the film "Polish-Russian War" is not publicly known.
LLaMA-3: The texts do not mention the mother of the director of the film Polish-Russian War. However, the film Polish-Russian War is mentioned in the provided texts, but it is not clear who the director is.
The texts mention the following individuals:
- Maheen Khan, a Pakistani fashion and costume designer - Viktor Yeliseyev, a Russian general, orchestra conductor, and music teacher - Alice Washburn, an American stage and film actress - Minamoto no Chikako, the mother of Prince Morinaga
None of these individuals are mentioned as the director of the film Polish-Russian
Ours: Małgorzata Braunek is the mother of Xawery Żuławski, the director of the film Polish-Russian War (Wojna polsko-ruska).

Table 19: Examples of Self-RAG, Llama3-8B, and our method on 2WikiQA.

Question: The first foreigner to receive bharat ratna was?
Self-RAG: The first foreigner to receive the Bharat Ratna was Dr. C. V. Raman, an Indian physicist, in 1954.
LLaMA-3: The first foreigner to receive the Bharat Ratna was Khan Abdul Ghaffar Khan, a Pakistani national, who was awarded in 1987.
Ours: The first foreigner to receive the Bharat Ratna was Khan Abdul Ghaffar Khan, a Pakistani national, in 1987.

Table 20: Examples of Self-RAG, Llama3-8B, and our method on ASQA.

Question: What is Richard Harris's occupation?
Irrelevant Para: J. Richard Harris
John Richard Harris (born 1910) was an Irish entomologist, fishing consultant, tackle merchant, and author. He was a keen angler and tier of flies from boyhood. He was a sometime merchant seaman, journalist, and freshwater biologist. He was a demonstrator in limnology at Trinity College, Dublin. He was a director of Garnetts & Keegan's Ltd, Dublin, gunsmiths and suppliers of fishing tackle, until his retirement in 1984. He wrote An Angler's Entomology, a book about mayflies for fly fishermen. He has been called, "perhaps the greatest living Irish angler-entomologist". He has also been described as, "a large affable man 'with a sharing attitude towards his whiskey and a colourful manner of expressing his trenchant views on fishing, fishermen, journalists, rugby, life and other matters'
LLaMA-3: Richard Harris has multiple occupations.
Richard Harris is a television writer, primarily writing for the crime and detective genres, and has worked on various TV shows and stage productions. He is also a cave diver with over 30 years of experience, having led expeditions and made significant discoveries. Additionally, Richard Harris is an anaesthetist, having completed his medical training and worked as an aeromedical consultant and anaesthetist for the South Australian Ambulance Service. He is also known for his role in the Tham Luang cave rescue and was jointly awarded 2019 Australian of the Year.
There are four different individuals with the name Richard Harris: a television writer, a cave diver and anaesthetist, a former professional footballer, and another anaesthetist who is also a cave diver.
Ours: There are four individuals named Richard Harris with different occupations: television writer, footballer, anaesthetist/cave diver, and composer.
Television Writer Richard Harris is a British writer most active from the early 1960s to the mid-1990s, contributing episodes of crime and detective genres and winning the prize for best situation comedy from the Writers' Guild of Great Britain for Outside Edge.
Footballer Richard Harris played in the Football League for Crystal Palace and Wycombe Wanderers during the late 1990s and early 2000s and made an appearance in the FA Youth Cup final against Leeds United.
Two Anaesthetists named Richard "Harry" Harris appear in the provided texts. One anaesthetist played a crucial role in the Tham Luang cave rescue and was jointly awarded 2019 Australian of the Year as a result, while the other anaesthetist/cave diver has over 30 years of experience, leading teams to record depths and participating in recovery missions and receiving awards for his exceptional contributions to cave diving exploration.
Composer Richard Frank Keith Harris studied composition and orchestration at the University of Edinburgh, co-founded Piano Circus, and commissioned and performed works by Arvo Pärt, Brian Eno, Philip Glass, and Steve Reich. His compositions feature on Argo CDs Loopholes and Landscapes Of The Heart, and he produced successful arrangements of works by Terry Riley and Thomas Ades. His work Hexada was featured in the UK television programme The Score..

Table 21: Examples of irrelevant documents.