SEER: Self-Aligned Evidence Extraction for Retrieval-Augmented Generation

Anonymous EMNLP submission

Abstract

Recent studies in Retrieval-Augmented Generation (RAG) have investigated extracting evidence from retrieved passages to reduce computational costs and enhance the final RAG performance, yet it remains challenging. Existing methods heavily rely on data-level augmentation, encountering several issues: (1) Poor generalization due to hand-crafted context filtering; (2) Semantics deficiency due to rule-based context chunking; (3) Skewed length due to sentence-wise filter learning. To address these issues, we propose a model-level evidence extraction learning framework, **SEER**, optimizing a vanilla model as an evidence extractor with desired properties through self-aligned learning. Extensive experiments show that our method largely improves the final RAG performance, enhances the faithfulness, helpfulness, and conciseness of the extracted evidence, and reduces the evidence length by 9.25 times.

1 Introduction

Recent years have witnessed the prevailing winds of Retrieval-augmented Generation (RAG), which is a prevailing paradigm for improving the performances of Large Language Models (LLMs) in various downstream tasks, such as question answering, making the output more reliable (Lewis et al., 2020; Chen et al., 2023; Jiang et al., 2023b; Ram et al., 2023), interpretable (Guu et al., 2020; Louis et al., 2024), and adaptable (Xu et al., 2023; Zakka et al., 2024). Traditional practices (Karpukhin et al., 2020; Min et al., 2019) often involve providing top-retrieved passages as the input context to LLMs without discrimination. However, imperfect retrieval systems frequently yield irrelevant content. Furthermore, indiscriminately feeding all retrieved content to LLMs will cause input redundancy, imposing a significant computational cost and rendering them prone to hallucination (Shi et al., 2023).

Ideally, LLMs should be grounded on supporting content that is both highly helpful to address user input and sufficiently concise to facilitate inference speed. However, it is practically impossible for imperfect retrieval systems to achieve such an ideal grounding solely (Wang et al., 2023). In fact, top-retrieved passages usually compose supporting and distracting content, inflicting a heavy blow on LLMs trained with high-quality corpora to generate the correct output. This motivates us to develop an evidence extractor, that aims at extracting supporting content while filtering out distracting content.

Recently, a pioneering study, FILCO (Wang et al., 2023), attempts to retrieve chunking document content with sentence precision via three filters, *i.e.,* StrInc, Lexical, and CXMI. Then, it trains a context filtering model, using context filtered by the above three measures as ground truth. Despite effectiveness, current context-filtering methods have several limitations: **(1) Hand-crafted Context Filtering.** Manually designed context-filtering measures typically require domain knowledge, which can hardly be adaptable to diverse downstream tasks with limited supervision. **(2) Disruptive Chunking on Context.** The use of chunking strategies may be ineffective as rule-based splitting on context usually cannot preserve its original semantics and often produces semantically deficient text blocks. **(3) Skewed Distribution in Length.** The length of supporting content in top-retrieved passages may vary largely across different samples. Hence, learning to filter context sentence-wise is biased toward skewed length distribution.

Given these limitations, an interesting question arises: Now that data-level augmentation[1] suffers from several issues, can we develop a model-level augmentation method free of the above problems? Inspired by the recent success of self-alignment (Li et al., 2023a; Zhang et al., 2024; Liang et al., 2024), self-aligned learning utilizes the model to improve itself and aligns its response with desired proper-

[1]Previous methods generally construct training signals via data engineering (we denoted it as data-level augmentation).

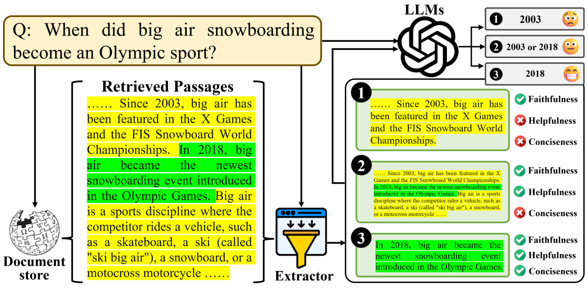


Figure 1: The RAG pipeline with the evidence extractor, in which the supporting content and the distracting content are marked in **green** and **yellow**, respectively.

ties, which is able to mitigate the heavy reliance on hand-crafted context filtering, rule-based context chunking, as well as sentence-wise filter learning.

Given extracted evidence, a question arises again: How to evaluate the quality of evidence properly? In principle, the evidence should be faithful (*i.e.,* avoiding intrinsic hallucination) to the retrieved passages (Rashkin et al., 2021; Maynez et al., 2020), helpful in addressing the user input (Adlakha et al., 2023), and concise to facilitate the inference speed (Ko et al., 2024). Figure 1 shows three representative scenarios: **(1)** When the evidence only favors faithfulness, LLMs may generate an incorrect answer; **(2)** When the evidence further favors helpfulness but lacks conciseness, LLMs' attention may be distracted by noise; **(3)** When the evidence favors all three criteria, LLMs can generate confidently with low computational costs.

In this paper, we propose a model-level evidence extraction learning framework, **SEER**, **S**elf-Aligned **E**vidence **E**xtraction for **R**etrieval-Augmented Generation. Specifically, it consists of three primary stages: **(1) Evidence Extraction:** To mitigate the issues above, we propose extracting diversified evidence with semantic consistency and varying length through response sampling, offering sufficient preference data for alignment. **(2) Expert Assessment:** For each extracted evidence, we construct a quadruple, QuadQARE, made up of query, answer, passage, and evidence. Then, we devise three experts to assess the quality of each extracted evidence *w.r.t.* three primary criteria. Given these scores, we propose smoothing CoV-Weighting, which explicitly leverages the statistics to estimate their relative weighting and result in the CoV-Weighted scores. **(3) Self-Alignment:** With a ranking list of extracted evidence and their smoothing CoV-weighted scores, a question remains: How to optimize extraction preference with the ranking position? To this end, we propose a listwise-aware Lambda Preference Optimization method, LPO, assigning each preference pair with a listwise-aware weight scaled by the gain in Reciprocal Rank from swapping the position of two evidence (Donmez et al., 2009; Burges et al., 2006; Wang et al., 2018).

It is worth mentioning that **SEER** is **criterion-agnostic** and can employ any off-the-shelf expert. Here, we use faithfulness, helpfulness, and conciseness, which are regarded as three primary criteria for assessing the quality of evidence (Maynez et al., 2020; Rashkin et al., 2021; Adlakha et al., 2023; Ko et al., 2024). Our **main contributions** are summarized as follows: **(1)** We propose a novel evidence extraction learning framework, **SEER**, leveraging preference data augmented by the model itself to improve performance, being free of the arduous workforce. **(2)** We devise three experts to assess the quality of the evidence, and design smoothing CoV-weighting to get an overall assessment, which supports criterion-agnostic. Besides, we propose a listwise-aware preference optimization method, LPO, seamlessly bringing the ranking position into preference learning. **(3)** Extensive experiments on three datasets show that our method can considerably improve QA performance, enhance the quality of evidence, as well as reduce computational costs.

2 Preliminaries

2.1 Problem Formulation

In this task, we are given a base extractor $\mathcal{E}$, and a fixed generator $\mathcal{G}$, where we choose Llama2-7b-Chat (Touvron et al., 2023) as the backbone for the base extractor $\mathcal{E}$. For a given query q and its corresponding golden answer a, we assume a set of retrieved passages $P = \{p_i\}_{i=1}^{K}$, where K is the retrieved size. Here, we aim to fine-tune the base extractor $\mathcal{E}$ via self-alignment to get the aligned extractor $\tilde{\mathcal{E}}$, for the generator $\mathcal{G}$ to leverage the better evidence and achieve superior performance:

$$e \sim \tilde{\mathcal{E}}(\cdot|q \oplus P), \quad o \sim \mathcal{G}(\cdot|q \oplus e), \tag{1}$$

where e and o denote the extracted evidence and the generated output, respectively; $\oplus$ denotes the concatenation operation; q is the given user query.

2.2 Augmentation Analysis

As stated in Section 1, data-level augmentation suffers from several issues, severely hindering the optimization for context filtering. To verify this, we compare the context relevance between data-level and model-level augmentation, where the **context**

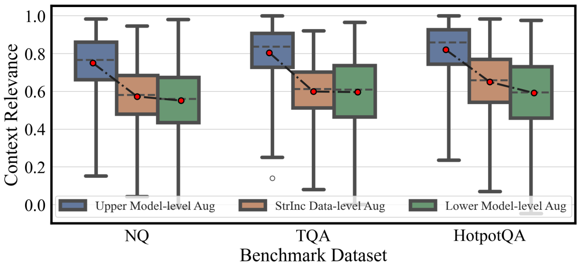


Figure 2: Comparison between model-level and data-level augmentation in terms of their context relevance.

relevance is the cosine similarity between the extracted evidence and the user query[2]. Here, we use StrInc as the representative data-level augmentation method (abbreviated as "**StrInc Data-level Aug**"), as it usually performs best on QA tasks according to (Wang et al., 2023). For another, we perform model-level augmentation by response sampling (More details can be seen in §3.1). We take the best-performing extracted evidence for each QA pair as "**Upper Model-level Aug**" while the worst-performing one as "**Lower Model-level Aug**".

We experiment on three datasets, *i.e.,* NQ, TQA, and HotpotQA. As shown in Figure 2, we find that: **(1)** The context relevance of Upper Model-level Aug is consistently higher than that of StrInc Data-level Aug. **(2)** The context relevance of StrInc Data-level Aug generally lies in the middle of Upper and Lower Model-level Aug. From the above observations, our claim is well-validated, since model-level augmentation shows a larger potential than data-level one. Hence, it is valuable to conduct model-level augmentation for better performance.

3 Methodology

Figure 3 depicts the overall framework of **SEER**, composing three key stages: **(1) Evidence Extraction** (§3.1), which extracts evidence via response sampling. **(2) Expert Assessment** (§3.2), which assesses the quality of evidence. **(3) Self-Alignment** (§3.3), which aligns the extractor with extraction preference. The learning algorithm of our proposed method can be seen in Appendix D in Algorithm 1.

3.1 Evidence Extraction Stage

As stated in Section 1, data-level augmentation (Wang et al., 2023) suffers from several issues. An empirical study (§2.2) further indicates that model-level augmentation is more beneficial for performance improvement than data-level augmentation. Hence, we aim to utilize the base extractor $\mathcal{E}$ to improve itself and align it with desired properties. To this end, we probe into its evidence extraction preference by response sampling for preference data collection. Specifically, given a query q and its retrieved passage P, we generate multiple candidate extracted evidence $\{e_i\}_{i=1}^{M}$ via response sampling $e_* \sim \mathcal{E}(\cdot|q \oplus P)$, where M is the sample size.

However, LLMs often tend to be overconfident in their knowledge (Xiong et al., 2023). As such, the response distribution typically follows a power-law, where head responses occupy a large portion of extracted evidence while long-tail ones are very sparse. Directly using the power-law response distribution for alignment would cause preference optimization to be biased toward head responses. Hence, we remove duplicates and obtain the uniformly distributed set, *i.e.,* $\{e_i\}_{i=1}^{N}$, where we use n-gram similarity (Kondrak, 2005) to detect duplicates and N is the remaining size. In practice, we find using the uniform response distribution does matter for alignment to reach higher performance.

3.2 Expert Assessment Stage

Although the base extractor can extract evidence, its output might be unfaithful, unhelpful, and unconcise, regarded as three primary obstacles hindering the quality of evidence (Maynez et al., 2020; Rashkin et al., 2021; Adlakha et al., 2023; Ko et al., 2024). As such, we devise **three experts** to assess the quality of extracted evidence *w.r.t.* faithfulness, helpfulness, and conciseness[3], respectively. Considering multiple scores for each extracted evidence, we devise a **smoothing CoV-Weighting** schema in order to get the overall assessment score.

Obtaining Oracle Scores. For expert assessment, we first collect a set of QuadQARE $< q, a, P, e >$, a Quadruple composed of Query q, Answer a, Retrieved passage P, and extracted Evidence e. Then, we design three plug-and-play experts to assess the quality of extracted evidence, parallelly.

- **Faithfulness Expert.** It focuses on the faithfulness of each extracted evidence. Toward this end, we adopt an advanced NLI model, ALIGNSCORE[4] (Zha et al., 2023), to evaluate the consistency between the retrieved passage P and extracted evidence e in terms of hallucination. Specifically, we treat the retrieved passage and

[2] We employ the SBERT-NLI-base model Reimers and Gurevych (2019) (denoted as SBERT) to encode the extracted evidence and the user query into sentence embedding vectors.

[3] We use the term "**oracle**" to denote three primary criteria.

[4] We use ALIGNSCORE-large for faithfulness assessment.

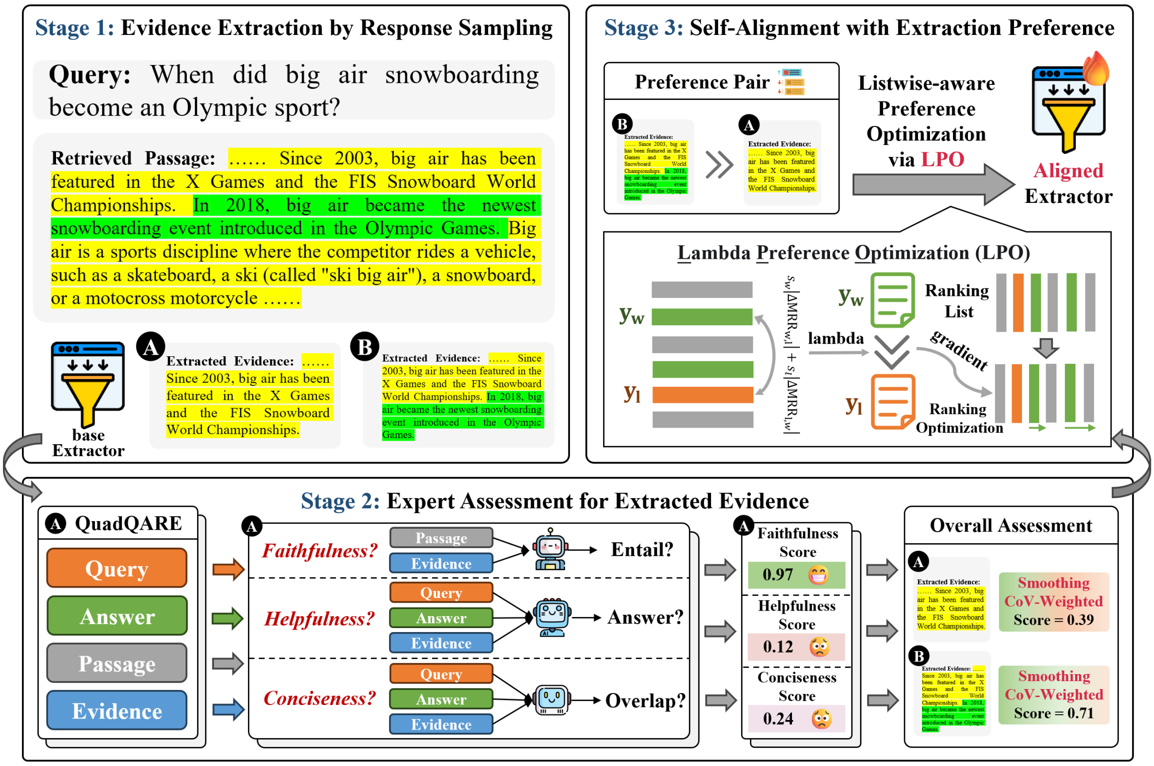


Figure 3: The overall system framework of our **SEER**, which mainly consists of three modeling stages.

extracted evidence as the premise and hypothesis, respectively. Then, we leverage ALIGNSCORE to measure to what extent the extracted evidence e could be entailed by the retrieved passage P:

$$s^f = \text{ALIGNSCORE}(P, e), \qquad (2)$$

where $s^f \in [0, 1]$ is the faithfulness score. If the hypothesis e is faithful to the premise P, then the score is close to 1, otherwise, it is close to 0.

- **Helpfulness Expert.** It examines the helpfulness of each extracted evidence candidate in terms of output improvement. In other words, it checks whether the extracted evidence e contributes to the model's output improvement when utilized as input. Specifically, we assess its potential influence on LLMs by calculating the change in the log probability of generating the golden answer a between the model's output before and after the inclusion of the extracted evidence e:

$$s^h = \text{Sig}\left(\log \frac{\prod f(a|q \oplus e)}{\prod f(a|q)}\right), \qquad (3)$$

where $s^h \in [0, 1]$ is the helpfulness score, $f(\cdot)$ is the helpfulness expert[5], $\text{Sig}(\cdot)$ is the sigmoid function. Similarly, if the extracted evidence e is helpful for LLMs to output the golden answer a, the score is close to 1, otherwise, it is close to 0.

[5]We employ Flan-T5-XL for helpfulness assessment.

- **Conciseness Expert.** If only the above two experts are considered, the aligned extractor can easily be achieved by directly treating the retrieved passage as evidence. To avoid such a trivial solution, we further measure the conciseness of the extracted evidence e. Towards this end, we first convert the query q and the golden answer a into the full-length answer[6] t, which represents minimal information for the need to answer the query. Subsequently, we leverage SBERT (Reimers and Gurevych, 2019) to measure to what extent the semantic overlap between the full-length answer and the extracted evidence:

$$s^c = \text{SBERT}_{\text{cosine}}(t, e), \qquad (4)$$

where $s^c \in [-1, 1]$ is the conciseness score via measuring cosine similarity between the sentence embedding of t and e, t is a full-length answer. Here, we prompt GPT-3.5-turbo to generate a full-length answer t given the query q and its answer a. More details can be seen in Appendix B.

Weighting Oracle Scores. Having obtained the oracle scores, a question naturally arises: How to get the overall assessment for each extracted evidence? A straightforward way is to compute the average of

[6]The full-length answer is generated by transforming the question and its corresponding answer into a declarative statement (Pal et al., 2019; Jain et al., 2021).

the oracle scores. However, equal weighting might not result in optimal alignment, since the learning difficulty is inconsistent. Therefore, the weights should match the learning difficulty to guide the preference optimization process. Given this, we propose smoothing CoV-weighting, leveraging the variability of the scores in relation to the mean:

$$c^f = \sigma^f / \mu^f, \tag{5}$$

where σ^f and μ^f denote the mean and the standard deviation of faithfulness score s^f, c^f is the Coefficient of Variation (CoV) whose value is independent of the magnitude. As such, CoV can decouple the score magnitude from the score weighting, so a type of score with a small magnitude may still be relatively impactful when it is variant (Groenendijk et al., 2021). Analogously, we obtain the CoV of the helpfulness and conciseness score, *i.e.,* c^h and c^c. Moreover, we employ the softmax function with temperature on the coefficient of variation of these scores, which controls the smoothness of the score weight to avoid abnormal score weight:

$$\alpha^f = \frac{\exp(c^f/\tau)}{\sum_* \exp(c^*/\tau)}, \tag{6}$$

where α^f is the faithfulness score weight, τ is the temperature. Analogously, we obtain the helpfulness and conciseness score weight, *i.e.,* α^h and α^c. Then, the CoV-weighted score can be defined as:

$$s = \alpha^f s^f + \alpha^h s^h + \alpha^c s^c, \tag{7}$$

where the score weight increases when the std increases or the mean decreases, ensuring more optimization proceeds when the score is more variant.

3.3 Self-Alignment Stage

After obtaining the preference data over all candidates $\mathcal{D} = \{(q \oplus P, e_i, e_j) | 1 \le i, j \le N, s_i > s_j\}$, where each tuple represents a choice preference between winning and losing extracted evidence, we proceed to the stage of alignment tuning for improving faithfulness, helpfulness, and conciseness. For alignment training, previous works commonly adopt Proximal Policy Optimization (PPO) (Schulman et al., 2017) or Direct Preference Optimization (DPO) (Rafailov et al., 2023). However, PPO cannot perceive the ranking position and DPO treats all preference pairs indiscriminately. Due to the above drawbacks, both of them cannot result in optimal alignment. Inspired by the Lambdaloss method (Donmez et al., 2009; Burges et al., 2006; Wang et al., 2018), we propose a listwise-aware Lambda Preference Optimization algorithm, LPO, which seamlessly brings the ranking position into DPO by assigning a lambda weight to each pair:

$$\mathcal{L}(\pi_\theta; \pi_{\text{ref}}, \lambda_{w,l})_{\text{LPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\lambda_{w,l} \log \text{Sig} \left(\beta \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right], \tag{8}$$

where $\pi_\theta = \tilde{\mathcal{E}}, \pi_{\text{ref}} = \mathcal{E}, x = q \oplus P, y_w, y_l = e_i, e_j$. We implement the lambda weight $\lambda_{w,l}$ for Mean Reciprocal Rank (MRR), *i.e.,* measuring the gain in Reciprocal Rank from swapping the position of two candidates, which can be formulated as follows:

$$\lambda_{w,l} = s_w \Delta\text{MRR}_{w,l} + s_l \Delta\text{MRR}_{l,w}, \tag{9}$$

where $\Delta\text{MRR}_{w,l} = \frac{1}{r_w} - \frac{1}{r_l}$, r_w is the rank position of y_w in the ranking permutation induced by the smoothing CoV-weighted score s. Thus, by introducing the lambda weight, LPO becomes a listwise-aware method. LPO is designed to work with any ranking metric, as long as the lambda weight can be defined, *e.g.,* NDCG (Liu et al., 2024). Here, we implement LPO to optimize a well-founded ranking metric MRR because it is simple yet effective.

4 Experiments

In this section, we conduct extensive experiments on three QA benchmark datasets to answer the following Research Questions (**RQs**): **RQ1:** How does our model contribute to QA accuracy compared with other state-of-the-art methods? **RQ2:** Can LPO facilitate the generation of more faithful, helpful, and concise evidence? **RQ3:** Can our model perform robustly to noise from irrelevant passages? **RQ4:** How effective are the key settings in our model, such as smoothing CoV-weighting?

4.1 Experimental Settings

Datasets and Metrics. We experiment on three benchmark QA datasets, NaturalQuestions (NQ) (Kwiatkowski et al., 2019), TriviaQA (TQA) (Joshi et al., 2017), and HotpotQA (Yang et al., 2018). Following Wang et al. (2023), we use the processed version (Lee et al., 2019) of NQ for experiments, discarding answers with more than 5 tokens. As NQ and TQA belong to the extractive QA task, we use Exact Match (EM) as their evaluation metric, where a score of 1 is assigned if at least one among multiple correct answers appears in the response

Datasets	Generators	Metrics	WE	CGE		FGE			
			Zero	Full	SeleCtx	LLM-Embedder	Bge-Reranker	FilCo	SEER
NQ	Flan-T5	EM	0.0934	0.4137	0.2853	0.3953	0.4089	0.3809	**0.4322**
		Tok	0	732	290	147	148	**62**	89
	Llama2	EM	0.2695	0.4382	0.3850	0.4208	0.04202	0.4061	**0.4549**
		Tok	0	804	319	160	162	**67**	95
TQA	Flan-T5	EM	0.2621	0.6320	0.5022	0.5689	0.6227	0.6431	**0.6503**
		Tok	0	760	306	152	153	130	**121**
	Llama2	EM	0.4898	0.6571	0.6061	0.6239	0.6581	0.6599	**0.6711**
		Tok	0	813	331	161	163	137	**133**
HotpotQA	Flan-T5	F_1	0.5289	**0.5702**	0.5127	0.5532	0.5608	0.5535	0.5615
		Tok	0	765	313	154	153	**56**	58
	Llama2	F_1	0.6467	0.6978	0.6658	0.6940	0.7106	0.7132	**0.7312**
		Tok	0	821	337	165	164	**59**	62

Table 1: QA performance comparison, where the best results are **boldfaced** and the second-best results are underlined, in each row. 'Tok' is the average length of extracted evidence fed into generators, where the smaller the value, the lower the computational cost. All improvements are significant with p-value < 0.01 according to t-test.

of the QA model; otherwise, the score is 0. While HotpotQA belongs to the abstractive QA task, we employ unigram F_1 to evaluate answer correctness. As the test set for HotpotQA is unavailable, we report the dev set results. The detailed statistics of datasets are summarized in Appendix A in Table 3.

Baseline Methods. There are three types of baselines: **(1) Without Evidence (WE)** includes **(i) Zero-shot (Zero)** that does not pass any evidence to LLMs. **(2) Coarse-grained Evidence (CGE)** includes **(i) Full Passage (Full)** that directly passes the top-retrieved passage to LLMs, **(ii) Select-Context (SeleCtx)** (Li et al., 2023b) that identifies and prunes redundancy in the top-retrieved passage based on perplexity. **(3) Fine-grained Evidence (FGE)** includes **(i) LLM-Embedder** (Zhang et al., 2023) that extracts the sub-passages with the highest similarity to the query from the top-retrieved passage, **(ii) Bge-Reranker-Large (Bge-Reranker)** (Xiao et al., 2023) that reorders all sub-passages in the top-retrieved passage and uses the top-ranked sentence as evidence, **(iii) FilCo** (Wang et al., 2023) that learns to filter the top-retrieved passage with sentence precision leveraging data-level augmentation to label ground-truth.

Generators for QA. To measure the efficacy of the evidence extracted by **SEER** and other competitive baselines[7], we employ two different generators, *i.e.*,, **Flan-T5-XL** (Chung et al., 2024) and **Llama2-7B-Chat** (Touvron et al., 2023), for QA evaluation.

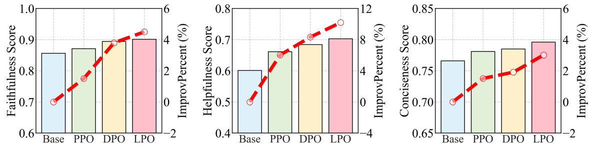
[7]In what follows, we use Flan-T5 and Llama2 to represent Flan-T5-XL and Llama2-7B-Chat, respectively, for brevity.

Implementation Details. Following Wang et al. (2023), we use the adversarial Dense Passage Retriever (DPR) (Karpukhin et al., 2020) to retrieve the top-5 passages from all Wikipedia passages. For each <user query q, retrieved passage P> pair, we set the sample size M as 10. For the temperature coefficient of smoothing CoV-weighting, we tune it within the range of $\{0.2, 0.5, 1.0, 2.0, 5.0\}$. We employ Llama2-7B-Chat (Touvron et al., 2023) as the base extractor $\mathcal{E}$ and fine-tune it on the constructed preference data for 2 epochs to get the aligned extractor $\tilde{\mathcal{E}}$. We adopt greedy decoding for evidence extraction and output generation. More implementation details are shown in Appendix A.

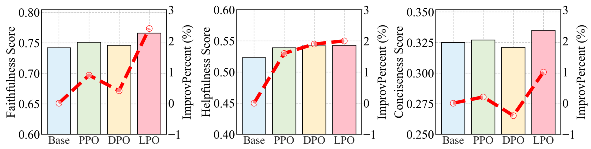
4.2 Model Comparison (RQ1)

To examine the impact of evidence extraction on the final RAG performance, we experiment on three benchmark QA datasets, where we prepend the extracted evidence before the user query and then input them together into the generator. Besides, we use the tokenizer of Flan-T5 and Llama2 to convert the extracted evidence into a list of subwords and then calculate the length of the list, where the length is adopted as a metric (denoted by '**Tok**') measuring the computational burden to a large extend. Table 1 shows the final RAG performance of different baseline evidence extraction methods and our proposed **SEER**. From the experimental results, we mainly have the following observations:

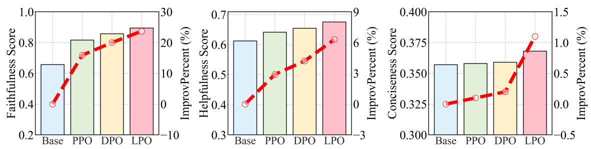
- In all cases, **SEER** outperforms FilCo by a large margin, indicating the superiority of model-level



(a) NQ dataset.



(b) TQA dataset.



(c) HotpotQA dataset.

Figure 4: Alignment performance *w.r.t.* faithfulness, helpfulness, and conciseness. The bar represents the oracle scores, while the line denotes the percentage of performance improvement in comparison with the Base.

augmentation that can provide more informative samples than data-level one. For example, **SEER** achieves 13.5% and 12.0% improvements in the NQ dataset with Flan-T5 and Llama2 generators, while the average evidence length is very close.

- Optimizing the three primary criteria for evidence extraction (*i.e.,* faithfulness, helpfulness, and conciseness) yields such impressive performance improvements, considering most baselines come from studies in recent two years. This demonstrates that these three properties strongly agree with the evidence quality in RAG, while current methods might not satisfy all of them simultaneously, which results in inferior results.

- Comparing different baselines, it is not surprising the method without evidence performs the worst. Secondly, methods with fine-grained evidence do not always perform better than ones with coarse-grained evidence. Specifically, the '**Full**' method generally performs well, as it preserves retrieved passages complete, while some FGE methods (*e.g.,* LLM-Embedder and Bge-Ranker) might lose key information in the process of evidence extraction, but it takes much more time for generation due to the long context. Last but not least, our **SEER** considerably outperforms the '**Full**' method in most cases, where the average improvement on the three datasets is 2.76% *w.r.t.* QA accuracy, but the average length of evidence fed into generators is reduced by a factor of 9.25.

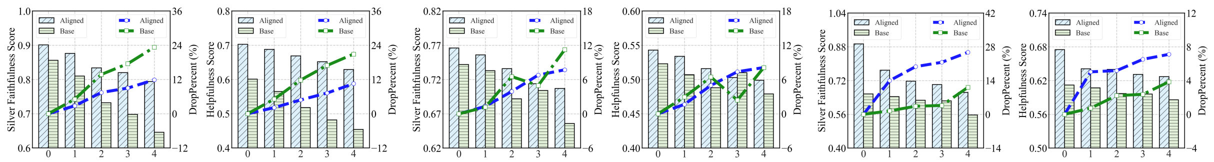
4.3 Alignment Study (RQ2)

To verify the effectiveness of the proposed LPO, we implement **SEER** with different types of PO methods to optimize the three primary criteria: (1) Base, *i.e.,* the base extractor; (2) PPO (Schulman et al., 2017); (3) DPO (Rafailov et al., 2023); (4) LPO (§3.3). In Figure 4, we present the oracle scores made by each method, the performance percentage of improvement *w.r.t.* the Base method. From the results, we find that: **(1)** Unsurprisingly, the Base without alignment performs the worst in 11 out of 12 cases, indicating the necessity of alignment for evidence extraction. **(2)** The PPO usually performs worse than the DPO one, as it directly optimizes the reward signal, *i.e.,* the oracle scores in our work, and thus neglects the pairwise signals between the extracted evidence corresponding to the same query. **(3)** Our LPO consistently outperforms the DPO, indicating the superiority of supplementing DPO with a listwise-aware weight. **(4)** After self-alignment, the average improvements of our LPO over the Base on three datasets are 10.2%, 6.16%, and 1.70% regarding the three primary criteria, showing huge potential to enhance the final RAG performance and quicken up the inference.

4.4 Robustness Analysis (RQ3)

In real-world scenarios, RAG systems usually suffer from data noise issues (Gao et al., 2023; Ding et al., 2024) caused by imperfect retrieval systems, etc. To simulate this scenario, we randomly add a certain proportion (0%, 100%, 200%, 300%, and 400%) of irrelevant passages to each test query. We use Noise-to-Signal Ratio (NSR) to denote the ratio of irrelevant passages to the relevant retrieved ones. Figure 5 shows the results on silver faithfulness[8] and helpfulness, while conciseness is omitted as the noise issue does not affect it much. The results show that: **(1)** The performance of both aligned and base extractors decreases, while the aligned one can consistently outperform the base under any NSR except for 1 case. **(2)** The performance drop percent of the aligned model is generally lower than the base in 2 out of 3 datasets. Besides, with 100% noise proportion, the aligned model can even outperform the base without noise data on all datasets. These observations manifest that **SEER** can endow the backbone with more robustness to noise issues.

[8]The silver faithfulness measures the entailment degree between the relevant retrieved passage (rather than the mixture of it and the irrelevant passages) and the extracted evidence.



(a) NQ dataset. (b) TQA dataset. (c) HotpotQA dataset.

Figure 5: Model performance *w.r.t.* NSR ratio. The bar denotes the silver faithfulness or helpfulness score, while the line denotes the performance drop percent compared to the model provided with only relevant retrieved passage.

Model	**Dataset**					
	NQ			**HotpotQA**		
	FS	HS	CS	FS	HS	CS
(A) SEER	0.901	**0.703**	0.796	0.894	**0.674**	**0.369**
(B) w/o Dup	0.896	0.675	**0.800**	0.881	0.657	0.365
(C) w/o CoV	**0.904**	0.696	0.787	**0.903**	0.668	0.363
(D) w/o Lam	0.894	0.684	0.785	0.857	0.654	0.359

Table 2: Ablation study with key settings of **SEER**, where we use FS, HS, and CS to indicate the Faithfulness, Helpfulness, and Conciseness scores, respectively.

4.5 Ablation Study (RQ4)

In Table 2, we conduct an ablation study to verify the effectiveness of key settings in our method, where w/o denotes without, (A) represents **SEER**, (B) removes the deduplication operation, (C) removes smoothing CoV-weighting by uniformly setting α^f, α^h, and α^c to $1/3$ in Eq. (7), (D) removes the lambda weight $\lambda_{w,l}$ in Eq. (8). From this table, we can find that (A) achieves the best or second-best results in all datasets, indicating all key settings are effective for **SEER**. By comparing (A) and (B), removing duplicates can largely improve helpfulness, as it effectively avoids preference optimization overwhelmed by head responses. By comparing (A) and (C), weighting the oracle scores based on their statistical properties can match the learning difficulty well. By comparing (C) and (D), we see that weighting the preference pairs plays a more key role than weighting the oracle scores. The reason might be that equally treating all preference pairs causes less attention to the crucial ones.

5 Related works

5.1 Context Refinement for RAG

Recently, many works have emerged, aiming at identifying the supporting content from retrieved passages. The common method is to rerank the retrieved passages and feed the top-ranked ones into generators (Zhang et al., 2023; Xiao et al., 2023). Thereafter, some methods leverage the capabilities of LLMs to summarize retrieved passages to identify key information (Ko et al., 2024; Laskar et al., 2023; Kim et al., 2024; Sarthi et al., 2024). Furthermore, a few methods leverage agent models to calculate perplexity as an important indicator to filter out low-information content (Li et al., 2023b; Jiang et al., 2023a). Other works use manually designed data-level augmentation to construct training signals for fine-tuning LLMs, to enhance their capacity to identify key information (Wang et al., 2023; Jin et al., 2024). In contrast to previous works heavily relying on hand-crafted augmentation, we use data augmented by the model itself to boost performance, free of the arduous workforce.

5.2 Self-Aligned Learning

Recently, a few works have attempted to utilize the model to improve itself and align its response with desired properties (Li et al., 2023a; Zhang et al., 2024; Liang et al., 2024; Sun et al., 2023). For example, (Zhang et al., 2024) utilizes the self-evaluation capability of LLMs to provide training signals steering the model towards actuality. (Liang et al., 2024) utilizes the model's self-awareness to align the model for hallucination mitigation. To the best of our knowledge, our study is the first to explore self-aligned learning for evidence extraction.

6 Conclusion

This work explores the method that learns to extract high-quality evidence to assist model generation and reduce computational cost. Different from previous works heavily relying on heuristics, we introduce a novel evidence extraction learning framework, **SEER**, which utilizes the model to calibrate its extraction preference via self-alignment. To this end, we first probe into model extraction preferences via response sampling, then assess the quality of extracted evidence via experts, and finally optimize the vanilla model as an evidence extractor via self-alignment. Extensive experiments show that **SEER** considerably improves the final RAG performance. Moreover, it can extract more faithful, helpful, and concise evidence, and also shows higher robustness against data noise issues.

Limitations

Despite our discoveries and improvements, we must acknowledge certain limitations in our work:

Firstly, computing resource constraints restrict our experiment to LLMs with limited and moderate scale, *i.e.,* Flan-T5-XL (Chung et al., 2024) and Llama2-7B-Chat (Touvron et al., 2023). We will explore the use of our method on larger models such as Llama2-70B in future work. The EM and F_1 metrics used in our experiments might overestimate the correctness of responses, even if the response does not convey equivalent semantics to the ground truth, since these metrics mechanically verify whether the answer exists in the response.

Secondly, our method still requires domain knowledge for devising experts to assess the quality of evidence, though it has considerably lightened the arduous workforce in data engineering. We experiment solely on Dense Passage Retriever (Karpukhin et al., 2020) with Wikipedia passages, while de facto RAG applications commonly involve multi-source retrieval with varied writing styles.

Thirdly, there are a few cases where the aligned extractor is vulnerable to data noise issues. As demonstrated in Table 5(c), with the NSR increases, the performance drop percent of the aligned extractor is higher than that of the base one, although it still outperforms the base one. Therefore, we are currently conducting further research to propose a more powerful evidence extractor, which is not only skilled at refining retrieved passages but also has higher robustness against noisy passages.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *CoRR*, abs/2307.16877.

Christopher J. C. Burges, Robert Ragno, and Quoc Viet Le. 2006. Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 193–200. MIT Press.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023. Complex claim verification with evidence retrieved in the wild. *CoRR*, abs/2305.11859.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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

Pinar Donmez, Krysta M. Svore, and Christopher J. C. Burges. 2009. On the local optimality of lambdarank. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 460–467. ACM.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Rick Groenendijk, Sezer Karaoglu, Theo Gevers, and Thomas Mensink. 2021. Multi-loss weighting with coefficient of variations. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1468–1477. IEEE.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Manas Jain, Sriparna Saha, Pushpak Bhattacharyya, Gladvin Chinnadurai, and Manish Kumar Vatsa. 2021. Natural answer generation: From factoid answer to full-length answer using grammar correction. *CoRR*, abs/2112.03849.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023a. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *CoRR*, abs/2310.06839.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7969–7992. Association for Computational Linguistics.

Jiajie Jin, Yutao Zhu, Yujia Zhou, and Zhicheng Dou. 2024. BIDER: bridging knowledge inconsistency for efficient retrieval-augmented llms via key supporting evidence. *CoRR*, abs/2402.12174.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

SungHo Ko, Hyunjin Cho, Hyungjoo Chae, Jinyoung Yeo, and Dongha Lee. 2024. Evidence-focused fact summarization for knowledge-augmented zero-shot question answering. *CoRR*, abs/2403.02966.

Grzegorz Kondrak. 2005. *N*-gram similarity and distance. In *String Processing and Information Retrieval, 12th International Conference, SPIRE 2005, Buenos Aires, Argentina, November 2-4, 2005, Proceedings*, volume 3772 of *Lecture Notes in Computer Science*, pages 115–126. Springer.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Md. Tahmid Rahman Laskar, Mizanur Rahman, Israt Jahan, Enamul Hoque, and Jimmy Huang. 2023. Cqsumdp: A chatgpt-annotated resource for query-focused abstractive summarization based on debatepedia. *CoRR*, abs/2305.06147.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6086–6096. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023a. Self-alignment with instruction backtranslation. *CoRR*, abs/2308.06259.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023b. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6342–6353. Association for Computational Linguistics.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation. *CoRR*, abs/2401.15449.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J. Liu, and Xuanhui Wang. 2024. Lipo: Listwise preference optimization through learning-to-rank. *CoRR*, abs/2402.01878.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 22266–22275. AAAI Press.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Knowledge guided text retrieval and reading for open domain question answering. *CoRR*, abs/1911.03868.

Vaishali Pal, Manish Shrivastava, and Irshad Bhat. 2019. Answering naturally: Factoid to full length answer generation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 1–9, Hong Kong, China. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *CoRR*, abs/2302.00083.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 704–718. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: recursive abstractive processing for tree-organized retrieval. *CoRR*, abs/2401.18059.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David D. Cox, Yiming Yang, and Chuang Gan. 2023. SALMON: self-alignment with principle-following reward models. *CoRR*, abs/2310.05910.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. 2018. The lambdaloss framework for ranking metric optimization. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 1313–1322. ACM.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md. Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. *CoRR*, abs/2311.08377.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *CoRR*, abs/2309.07597.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *CoRR*, abs/2306.13063.

Benfeng Xu, Chunxu Zhao, Wenbin Jiang, Pengfei Zhu, Songtai Dai, Chao Pang, Zhuo Sun, Shuohuan Wang, and Yu Sun. 2023. Retrieval-augmented domain adaptation of language models. In *Proceedings of the 8th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2023, Toronto, Canada, July 13, 2023*, pages 54–64. Association for Computational Linguistics.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with A unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11328–11348. Association for Computational Linguistics.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *CoRR*, abs/2402.09267.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *CoRR*, abs/2403.13372.

A More Implementation Details

Dataset	Task	Metric	#Train	#Dev	#Test
NQ	Extractive QA	EM	79.1k	8.7k	3.6k
TQA	Extractive QA	EM	78.7k	8.8k	11.3k
HotpotQA	Abstractive QA	F_1	88.9k	5.6k	5.6k

Table 3: Statistics and task metrics for three datasets.

Statistics of datasets. We conduct extensive experiments on three benchmark datasets, *i.e.,* NaturalQuestions (NQ) (Kwiatkowski et al., 2019), TriviaQA (TQA) (Joshi et al., 2017), and HotpotQA (Yang et al., 2018), for evaluating our proposed method and the competitive baselines. We show the detailed statistics of these datasets in Table 3.

Response sampling details. Given the query and the retrieved passages, we prompt the base extractor to generate 10 candidate response samples and we remove duplicates. To fully probe the evidence extraction preferences of the base extractor, we have modified the generation configuration to make the responses more varied. Specifically, we set top-p, top-k, temperature, and the repetition penalty as 1.0, 80, 1.0, and 1.0 respectively, for collecting diverse preference data, used to align the responses of the based extractor with the desired properties.

Fine-tuning details. We use the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $eps = 1e^{-8}$. The learning rate is $1e^{-5}$ with 1.5% warmup ratio and cosine scheduler. The batch size, gradient accumulation step, and number of epochs are set as 16, 2, and 2.0, respectively. We leverage the parameter-efficient fine-tuning technique, specifically LoRA (Hu et al., 2022), where we employ the Llama-Factory[9] fine-tuning framework (Zheng et al., 2024) to implement all the preference optimization methods for fair comparisons.

Context relevance details. In Section 2, we use context relevance as the metric to measure how well the extracted evidence fits the current user query and can be effectively used to augment the quality of generation. To this end, we naturally define context relevance as the cosine similarity between the extracted evidence and the user query:

$$s^{cr} = \mathrm{SBERT}_{\mathrm{cosine}}(q, e), \tag{10}$$

where $s^{cr} \in [-1, 1]$ is the context relevance score; q and e denote the query and evidence, respectively.

Silver faithfulness details. In Section 4.4, we devise a metric, silver faithfulness, to measure the robustness of the evidence extractor against data noise issues commonly existing in real-world scenarios. Specifically, we fed the mixture of the relevant retrieved passage and the randomly sampled irrelevant passages into the extractor. Then, we treat the relevant retrieved passage and extracted evidence as the premise and hypothesis, respectively, measuring how well the extractor is robust to irrelevant context, which can be formulated as:

$$s^{sf} = \textsc{AlignScore}(\hat{p}, e), \quad e = \tilde{\mathcal{E}}(\cdot|q \oplus \check{P}),$$

where $s^{sf} \in [0, 1]$ is the silver faithfulness score; $\hat{p}$ is the relevant retrieved passage; $\check{P}$ is the mixture of $\hat{p}$ and those randomly sampled irrelevant passages.

B Full-length Answer Generation

To assess the conciseness of the extracted evidence, we propose measuring the information gap between it and the full-length answer. The full-length answer is generated by transforming the question and its corresponding answer into a declarative statement, as shown in Table 4. Towards this end, we prompt GPT-3.5-turbo to transform each question-answer pair into a full-length answer. Additionally, we prepared a few-shot examples to encourage well-organized output. The prompt for full-length answer generation can be found in Table 5.

[9]https://github.com/hiyouga/LLaMA-Factory

Question: Which branch of philosophy is concerned with fundamental questions about the nature of reality? **Answer:** Metaphysics **Full-length answer:** Metaphysics is the branch of philosophy concerned with fundamental questions about the nature of reality.	
Question: What country used the Drachma as its currency, before switching to the Euro in 2001? **Answer:** Greece **Full-length answer:** Greece used the Drachma as its currency before switching to the Euro in 2001.	
Question: Californian rock band Lit recorded A Place in the Sun in 1995, but what's their best known song? **Answer:** My Own Worst Enemy **Full-length answer:** The Californian rock band Lit recorded their album A Place in the Sun in 1995, and their best known song is My Own Worst Enemy.	

Table 4: Three examples of full-length answers from the NQ, TQA, as well as HotpotQA datasets, respectively.

Full-length Answer Generation Prompt

[Instruction]
You are given a question and its answer. Your task is to transform this question-answer pair into a declarative sentence with lossless fidelity to the original semantics.
[Here are three examples]
[Question]: What profession does Nicholas Ray and Elia Kazan have in common?
[Answer]: director
[Full-length answer]: Nicholas Ray and Elia Kazan have the profession of director in common.
[Question]: When is season seven of game of thrones coming out?
[Answer]: July 16, 2017
[Full-length answer]: Season seven of Game of Thrones is coming out on July 16, 2017.
[Question]: What is the moon festival called in Chinese?
[Answer]: Mid-Autumn Festival
[Full-length answer]: The moon festival is called the Mid-Autumn Festival in Chinese.
[Now complete the following]
[Question]: When did the genre of installation art start to gain acceptance?
[Answer]: in the 1970s
[Full-length answer]:

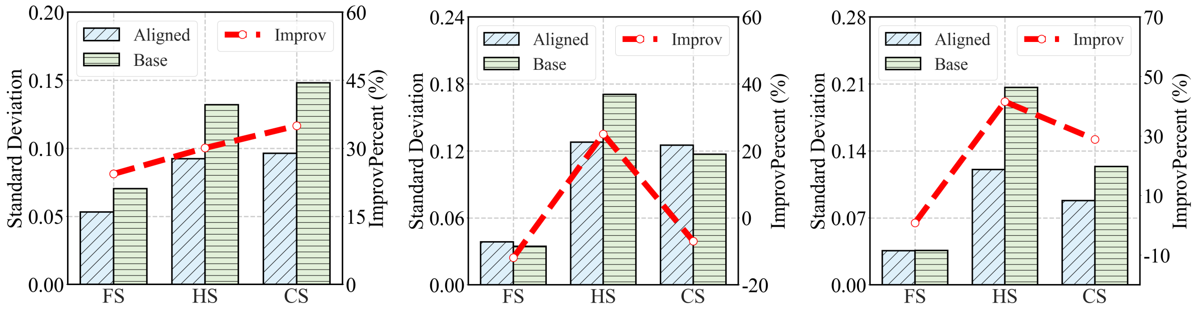
Table 5: The prompt for full-length answer generation.

C Stability Analysis

In Figure 6, we experiment to verify whether the stability of model generation is improved after self-alignment. Specifically, we generate ten pieces of evidence for each test query by response sampling with the same generation configuration as Section 3.1. Then, we measure the oracle scores, calculate the standard deviation, and compute the average. The results show that: **(1)** The generation stability of the aligned model performs much better than that of the base one in most cases. More precisely, the average improvement of the aligned model over the base one on the three datasets is 18.5%. **(2)** The generation stability in terms of helpfulness has seen greater improvements compared to the other two properties, with an average improvement of 32.2%, showing the huge potential to enhance the final RAG performance. These observations fully manifest that **SEER** is able to endow the backbone with superior generation stability during the inference.

D Learning Algorithm of SEER

Algorithm 1 demonstrates the learning algorithm of the proposed **SEER** framework. The algorithm can be divided into three stages, *i.e.,* **(1) Evidence Extraction** (line 3-6), **(2) Expert Assessment** (line 7-10), as well as **(3) Self-Alignment** (line 11-14).



(a) NQ dataset. (b) TQA dataset. (c) HotpotQA dataset.

Figure 6: Model stability *w.r.t.* faithfulness, helpfulness, and conciseness. The bar represents the standard deviation results, while the line represents the stability improvement percent of the aligned model compared to the base model. We use FS, HS, and CS to denote the Faithfulness, Helpfulness, and Conciseness scores, respectively, for simplicity.

Algorithm 1 Learning algorithm of **SEER**

Input: Trainig dataset with queries q, answers a, and retrieved passages $P = \{p_i\}_{i=1}^{K}$; the base evidence extractor $\mathcal{E}$; the sample size M; total number of iterations T.

Output: The aligned evidence extractor $\tilde{\mathcal{E}}$

Initialize the model parameter $\tilde{\mathcal{E}}$ with $\mathcal{E}$
for each $i \in [1, T]$ **do**
  # Stage1: Evidence Extraction
  Sample a mini-batch of (q, a, P) query-answer-passage triples from the dataset.
  Get evidence candidates $\{e_j\}_{j=1}^{M}$ via response sampling $e \sim \mathcal{E}(\cdot | q \oplus P)$.
  Obtain uniformly distributed set $\{e_j\}_{j=1}^{N}$ by removing duplicates in $\{e_j\}_{j=1}^{M}$.
  # Stage2: Expert Assessment
  Construct a QuadQARE for each evidence candidate $< q, a, P, e >$.
  Get the oracle scores for each evidence candidate (s^f, s^h, s^c) with Eq. (2-4).
  Get the smoothing CoV-weighted score s with Eq. (5-7).
  # Stage3: Self-Alignment
  Get the lambda weight $\lambda_{w,l}$ for each preference pair (x, y_w, y_l) with Eq. (9).
  Compute the preference optimization loss $\mathcal{L}_{\text{LPO}}$ with Eq. (8).
  Update the model parameter of $\tilde{\mathcal{E}}$ using gradient descent.
end for