Ranking Free RAG: Replacing Re-ranking with Selection in RAG for Sensitive Domains

Yash Saxena
UMBC
Baltimore, Maryland, USA
ysaxena1@umbc.edu

Ankur Padia
UMBC
Baltimore, Maryland, USA
pankur1@umbc.edu

Mandar S. Chaudhary*
eBay Inc.
San Jose, California, USA
manchaudhary@ebay.com

Kalpa Gunaratna
Independent Researcher
San Jose, California, USA
gunaratnak@acm.org

Srinivasan Parthasarathy
Ohio State University
Columbus, Ohio, USA
srini@cse.ohio-state.edu

Manas Gaur
UMBC
Baltimore, Maryland, USA
manas@umbc.edu

Abstract

Traditional Retrieval-Augmented Generation (RAG) pipelines rely on similarity-based retrieval and re-ranking, which depend on heuristics such as top-k, and lack explainability, interpretability, and robustness against adversarial content. To address this gap, we propose a novel method `METEORA` that replaces re-ranking in RAG with a rationale-driven selection approach. `METEORA` operates in two stages. First, a general-purpose LLM is preference-tuned to generate rationales conditioned on the input query using direct preference optimization. These rationales guide the **evidence chunk selection engine**, which selects relevant evidence in three stages: pairing individual rationales with corresponding retrieved evidence for *local* relevance, *global* selection with elbow detection for query-adaptive cutoff, and *context expansion* via neighboring evidence. This process eliminates the need for top-k heuristics. The rationales are also used for a consistency check using **Verifier LLM** to detect and filter poisoned or misleading content for *safe* generation. The framework provides explainable and interpretable evidence flow by using rationales consistently across both selection and verification. Our evaluation across six datasets spanning legal, financial, and academic research domains, `METEORA` improves generation accuracy by 33.34% while using approximately 50% fewer evidence than state-of-the-art re-ranking methods. In adversarial settings, `METEORA` significantly improves the F1 score from 0.10 to 0.44 over the state-of-the-art perplexity-based defense baseline, demonstrating strong resilience to poison attacks. The code is available in the GitHub repository[2].

1 Introduction

Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) by augmenting input prompts with external context, enabling more accurate and grounded responses. Recent works have applied RAG to reduce hallucinations [Ayala and Bechard, 2024], improve downstream task performance [Zheng et al., 2025], and enhance answer quality in open-domain question answering (QA) [Choi et al., 2025]. A typical RAG pipeline comprises a retriever, a re-ranker, and a generator. The re-ranker is crucial to select contextually relevant top-k chunks or evidence before passing them to the generator [Glass et al., 2022].

[*] This work does not relate to the author's position at eBay Inc.

[2] `https://anonymous.4open.science/r/METEORA-DC46/README.md`

Preprint. Under review.

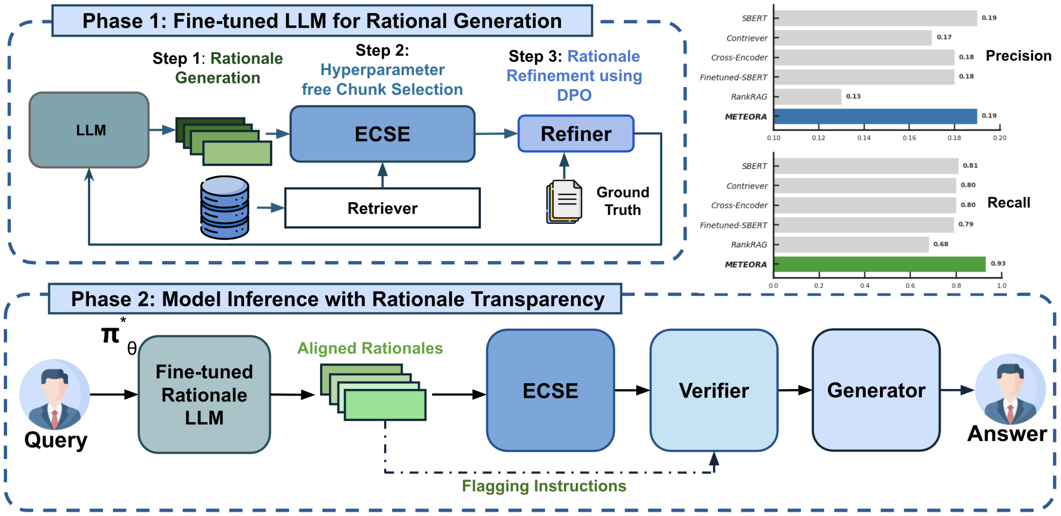

Figure 1: Overview of our METEORA framework.

Despite the growing popularity of RAG, its black-box nature poses challenges in sensitive domains like law, finance, and academic research [Zhou et al., 2024, Xue et al., 2024]. Existing re-ranking methods suffer mainly from three limitations. First, existing re-ranking approaches lack interpretability, as they rely on opaque similarity scores and a manually defined number k to select evidence, without providing justification for why particular evidence was chosen. Second, re-rankers are not robust to adversarial attacks such as the injection of irrelevant or poisoned content, which can corrupt the selected context and negatively impact the final generation. Third, these methods depend on heuristic decisions, particularly the choice of top-k, which is often query-specific and difficult to determine in advance. Selecting too few evidence may omit critical context, while selecting too many can introduce noise, degrading answer quality [Leng et al., 2024]. Such limitations raise serious concerns in high-stakes domains where factual accuracy and robustness are critical [Barron et al., 2024, Bhushan et al., 2025].

Recent work such as RAG^2 [Sohn et al., 2025] and RankRAG [Yu et al., 2024] has attempted to address these challenges. RAG^2 uses rationales to improve retriever capabilities to identify relevant evidence. However, it still relies on a re-ranking step to reorder retrieved evidence, thereby inheriting the limitations of traditional re-rankers, including a dependence on fixed top-k cutoffs and opaque scoring mechanisms. Similarly, RankRAG instruction-tunes a single LLM to both rank retrieved contexts and generate answers based on the retrieved passages. Because it leverages the LLM's parametric knowledge, RankRAG suffers from limited interpretability and lacks robust filtering mechanisms.

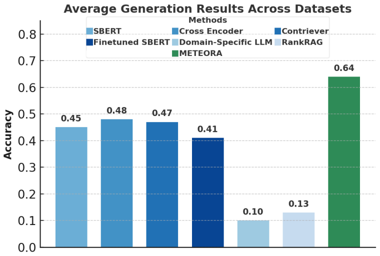

Figure 2: **Generation Results**. Comparison of different retrieval approaches across all datasets using the LLaMA-3.1-8B generator.

To address above challenges and limitations, we propose Method for Interpretable rank-free evidence selection with Optimal Rationale (METEORA), a rationale-driven, interpretable, and robust framework as an alternative to traditional re-ranking. As shown in Figure 1, METEORA operates in two phases. In the first phase, a general-purpose LLM, π_{ref}, is preference-tuned using Direct Preference Optimization (DPO) [Rafailov et al., 2023a] to learn an optimal policy π^*_θ that generates query-aligned rationales.

To tune the policy model, METEORA does not require a manually annotated preference dataset; instead, it is constructed from existing QA annotations. This preference data generation method is generalizable across datasets and domains. In the second phase, the model π^*_θ

is used to generate rationales to select and verify evidence. Generated rationales guide an unsupervised evidence selector, the *Evidence Chunk Selection Engine (ECSE)* (Section 2.3) to select relevant evidence in three steps: pairing individual rationales with retrieved evidence for local alignment, semantic pooling with elbow detection for adaptive thresholding, and expansion via neighboring evidence to recover missed evidence. Such a setting enables adaptive and interpretable evidence selection without relying on a subjective top-k heuristic. Moreover, the rationales help to flag the evidence that are inconsistent with the query and filter them using *Verifier LLM*. Because rationales are used consistently across both selection and verification, METEORA offers fully explainable and traceable evidence flow, allowing users to introspect the selected evidence explaining *why* it was selected, and *how* it influenced the final generated answer. In contrast to previous approaches, METEORA eliminates the re-ranking stage, integrating rationale-based decisions at every step of evidence processing.

Using rationales can significantly improve the quality of generated answers. To illustrate this, consider the plots in Figure 1 (top right), which show how traditional re-ranking methods based solely on similarity fail to capture contextually relevant evidence. This results in reduced recall for complex QA tasks across legal, financial, and academic research datasets. In contrast, METEORA employs rationale-guided selection to align evidence more precisely with the query, improving recall by 14.8% without compromising precision compared to the best-performing baseline. Furthermore, Figure 2 demonstrates that this improved selection leads directly to a 33.34% increase in generation accuracy over the next best-performing baseline, further validating the importance of recall and accurate evidence selection. This advantage is particularly valuable in high-stakes domains, where factual accuracy is paramount. **Our main contributions are:**

- We propose METEORA, a rationale-driven, explainable, interpretable, and robust alternative to re-ranking in RAG that eliminates the need for top-k heuristics.
- We introduce *ECSE*, an unsupervised rationale-based evidence selector, and a Verifier LLM to determine consistency of the evidence to filter poisoned content.
- We conduct extensive experiments across legal, financial, and academic research domains, showing consistent gains over state-of-the-art (SoTA) baselines. Moreover, in adversarial settings, METEORA significantly improves the average precision and recall scores by order of magnitude over the SoTA perplexity-based defenses.

2 METEORA

METEORA is a rationale-driven framework designed to select relevant evidence from a document for a given query such that it is interpretable, explainable, free from top-k heuristics, and robust to adversarial content. The framework operates in two stages: rationale generation (subsection 2.2) and selection (subsection 2.3 and 2.4).

2.1 Problem Formulation

METEORA aims to replace the traditional re-ranking step in RAG architectures with rationales. Formally, given a query, q, and a knowledge base, D, the objective is to select a subset of evidence, E_s, from retrieved documents, D_s, where $D_s \subset D$, such that: (i) E_s maximizes the coverage of relevant information to obtain higher recall, (ii) minimizes the inclusion of irrelevant or poisoned content to obtain higher precision, and (iii) achieves this without relying on vague heuristics such as a fixed number of top-k evidence.

2.2 Preference-Tuned Rationale Generator

Rationale generator aim to generates a set of rationales conditioned on the query. Figure 3 shows an example of a rationale generated for a real-world query from the PrivacyQA dataset. To enables effective learning without manual annotation while ensuring high-quality rationale generation, we use an off-the-shelf LLM and preference-tune it to generate rationales aligned with a query. We automatically construct preference dataset by pairing the query with rationales that led to correct evidence selection as positive samples (r_w), and with other rationales as negative samples (r_l).

This setup enables training the model using DPO, as defined by the following objective:

$$\mathrm{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}_{(q,c,r_w,r_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(r_w, q, c)}{\pi_{\mathrm{ref}}(r_w, q, c)} - \beta \log \frac{\pi_\theta(r_l, q, c)}{\pi_{\mathrm{ref}}(r_l, q, c)} \right) \right]. \quad (1)$$

Example

Query: TickTick: To Do List with Reminder, Day Planner's privacy policy; can it view my real name?

Rationale: Search for terms like `"real name"`, `"PII"`, or `"user information"`, especially in sections covering data collection, use, or disclosure. **Flagging Instructions:** Flag the chunk if it contains internally inconsistent language about real name usage, or if it contradicts other verified parts of the policy.

Figure 3: Generated Rationale Example

Here, π_{ref} is a frozen reference model, while π_θ is the same architecture with trainable parameters. The hyperparameter β controls the sharpness of preference learning. The rationale r_w corresponds to the rationale that leads to the correct evidence (positive), and r_l corresponds negative evidence. The model is trained to assign higher likelihood to r_w over r_l given the query, q, and evidence, e, i.e., $r_w \succ r_l \mid q, e$. Examples for positive and negative rationales can be found in subsection A.2. Figure 7 (in subsection A.2) illustrates a case where some of the generated rationales successfully lead to the selection of the correct evidence. In contrast, Figure 8 (subsection A.2) shows a case where none of the rationales selected the correct evidence. Despite this, both examples demonstrate interpretability by revealing which rationales contributed to the selection of which evidence, and explainability by clarifying why the LLM produced a particular answer. We model the joint distribution of rationale, query, and evidence, $\pi_\theta(r_w, q, e)$ as a product of two conditional distributions $\pi_\theta(e, r \mid q) = \pi_\theta(e \mid q) \cdot \pi_\theta(r \mid q, e)$. Since oracle context e is available during training, we focus on learning $\pi_\theta(r \mid q, e)$ and optimize rationale quality using the cosine similarity between rationale and evidence. At test time, where true context is unavailable, we condition only on the query to generate rationales. Such an approximation still yields superior performance compared to baselines as shown in section 4. The final preference-tuned model π_θ^* is used to generate aligned rationales for downstream selection and verification. The prompt used for rationale generation is provided in subsection A.3, and implementation details for preference-tuning the LLM is described in Table 6.

2.3 ECSE: Unsupervised Rationale Based Chunk Selector

Evidence Chunk Selection Engine (ECSE) is an unsupervised model that identifies the most relevant evidence by combining the generated rationals, input query, and the list of documents. The ECSE identifies relevant evidence using following techniques:

Pairing Rationales with evidence. Rationale-based pairing computes the similarity score (E_v) between each pair of a rationale and the evidences from documents, e_j, and selects the evidence with the highest match. Rationale and evidence are encoded using SBERT model to compute cosine similarity. Such a method ensures to obtain higher precision.

$$E_v \quad = \quad \left\{ \underset{e_j \in E}{max} \quad \mathcal{S}\left(r_i, e_j\right) | r_i \in \mathcal{R} \right\} \tag{2}$$

Pooling Rationales. To identify evidence that align with the overall intent of all rationales, we compute a pooled embedding $\bar{r}$ by averaging the individual rationale embeddings: $\bar{r} = \frac{1}{|R|} \sum_{r_i \in R} \text{SBERT}(r_i)$. We then calculate cosine similarity between $\bar{r}$ and each evidence embedding. The evidence are sorted based on the similarity, yielding a sequence of scores $\{s_1, s_2, \ldots, s_n\}$. We compute first-order differences $\Delta_i = s_i - s_{i+1}$ and apply z-score normalization to detect statistically significant drops in similarity. We select the first index k^* where the z-score of Δ_k deviates notably from the mean[3]. If no such drop is found, we are at the point of finding maximum curvature using second-order differences. We select the top-k^* evidence as E_g, enabling adaptive, data-driven cutoff without heuristic thresholds. For a detailed explanation of the working of the elbow detection, refer to subsection A.5.

Expanding Rationales Selection. Each of the selected evidence in E_g and E_v is expanded by considering evidence before and after the selected evidence to obtain a wider set of evidences E_w. Final set of candidate evidence is selected as follows: $E_s = E_v \cup E_g \cup E_w$

2.4 Verifier LLM

To make RAG robust against adversarial content, METEORA incorporates a Verifier LLM that filters selected evidence E_s before generation. The Verifier evaluates each evidence using the input query and its associated rationale, which includes embedded *Flagging Instructions*.

[3] A large similarity drop after cutoff indicates evidence are less aligned with the rationale intent.

Table 1: Dataset Statistics Across Domains

Dataset	# Documents	Avg. Tokens/Doc	# QA Pairs	Domain
ContractNLI	95	10,673	946	Legal, NDA related documents
CUAD	462	55,827	4,042	Legal, Private Contracts
MAUD	150	351,476	1,676	Legal, M&A documents of public companies
PrivacyQA	7	25,266	194	Legal, Privacy policies of consumer apps
FinQA	2,789	∼700	8,281	Finance, Financial documents of companies
QASPER	1,585	∼6,500	5,000+	Academic Research, NLP Papers

Evidences are flagged for (i) *factual violations*, when the content contradicts established facts; (ii) *contradiction*, when a evidence is logically inconsistent with other verified evidences; and (iii) *instruction violations*, when a evidence fails to meet the criteria embedded in the rationale. Flagged evidence are discarded, and only the filtered set is passed to the generator. For more information about the Verifier LLM and prompt format, see subsection 3.4 and subsection A.3, respectively.

2.5 METEORA Covers Bi-encoder and Cross-Encoder

Figure 4 compares bi-encoder, cross-encoder, and METEORA. METEORA generalizes both SBERT and Cross-Encoder: when the ECSE and Verifier are identity functions, it reduces to these models. Unlike Cross-Encoder, which jointly processes query and document, METEORA processes them independently like SBERT but uses rationales to guide evidence selection, combining efficiency with interpretability.

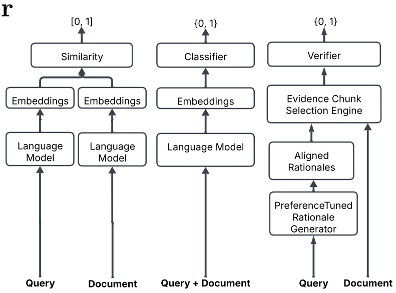

Figure 4: Comparison of Bi-encoder (left), Cross-Encoder (middle) and METEORA (right).

3 Experiments

To demonstrate the effectiveness of rationales, we evaluate METEORA on three tasks across six real-world benchmark datasets spanning multiple domains, and provide a detailed ablation study. For clarity, we report results using evidence of chunk size of 512 tokens in the main paper; results for other sizes are included in Appendix subsection A.7.

3.1 Tasks

We use following three tasks: (i) the Context Prioritization (CP) task to measures the ability to select relevant evidence for a given query; (ii) the Generation task, to assesses the quality of generated answers using the selected or re-ranked evidence; and (iii) the Adversarial Defense task to evaluates the defense ability to detect and filter poisoned or misleading content in the corpus. For the adversarial setting, we follow the poisoning strategy of Nazary et al. [2025], who injected five poisoned texts per instance. In our setup, we inject only one poisoned evidence per instance. Detecting a single poisoned instance becomes more difficult because it is hidden among mostly correct content, making it harder for the model to recognize it as harmful. When multiple poisoned instances are present, it gets easier to spot unusual patterns, but with just one, the model has fewer signals to detect the attack.

3.2 Datasets

We select six datasets spanning legal, financial, and academic domains based on their suitability for evaluating the Context Prioritization (CP) task, which measures how accurately systems select relevant text evidence from documents to answer queries. Each dataset provides: (1) question-answer (QA) pairs, (2) complete reference to lengthy documents ranging from 5-50 pages, and (3) human-annotated evidence spans that explicitly mark which specific sentences or paragraphs contain the exact information required to answer each query. These evidence spans serve as our ground-truth evidence, precisely identified text segments that evaluation metrics compare against system-selected evidences to compute retrieval accuracy scores, including Precision, Recall, and F1. In the legal domain, we use: (1) ContractNLI [Koreeda and Manning, 2021], where each instance evaluates whether contract clauses entail, contradict, or remain neutral to hypotheses, with law experts highlighting the exact clause text needed for reasoning; (2) PrivacyQA [Ravichander et al., 2019], containing questions about data collection practices with privacy specialists identifying the specific policy sections containing relevant disclosure statements; (3) CUAD [Hendrycks et al., 2021], comprising commercial contracts where legal professionals have marked exact paragraphs corresponding to 41 distinct clause categories; and (4) MAUD [Pipitone and Alami, 2024], featuring merger agreements where corporate attorneys have indicated the precise sections addressing specific acquisition terms. In the finance domain, we use FinQA [Chen et al., 2021], which requires extracting specific numerical values from financial reports to perform arithmetic operations (addition, subtraction, multiplication, division, percentages) for quarterly earnings analyses, with financial analysts marking the exact tables, figures, and statements containing the required values. Finally, in academic research, we use QASPER [Dasigi et al., 2021], which presents questions about research papers where domain scientists have identified the minimal set of sentences containing methodological details, experimental results, or theoretical claims needed for accurate answers. Table 1 shows the statistics.

3.3 Baselines

For fair comparison, we selected baseline models that span diverse re-ranking paradigms, including supervised and unsupervised, dual-encoder and cross-encoder designs. For the CP task, we compare our approach with SoTA re-ranking methods. These include Cross-Encoder ms-marco-MiniLM-L4-v2 [Huggingface], which jointly encodes the query and document to compute a fine-grained relevance score; Contriever [Izacard et al., 2022], an unsupervised dense retriever trained via contrastive learning; and SBERT [Reimers and Gurevych, 2019], a bi-encoder model that computes similarity using sentence embeddings. Additionally, we include a fine-tuned domain-specific variant of SBERT, the Fine-Tuned SBERT [Legal-huggingface], as base SBERT was found to perform better than or on par with Cross-Encoder and Contriever in preliminary experiments. This made it a suitable backbone to evaluate the effect of domain adaptation on re-ranking performance. We include RankRAG as an LLM-based baseline, which prompts the generator LLM to retrieve and re-rank document evidence based on the query, using the top-ranked evidence for answer generation. As the RankRAG model is not publicly released, we reproduce it using the Promptriever model [Weller et al., 2024] and achieved similar results. To evaluate adversarial robustness, we compare our Verifier against the Perplexity-based Defense [Zhou et al., 2024] which uses perplexity, a measure of deviation from expected language model behavior, to identify potentially poisoned content.

3.4 Settings

CP and Generation. We used LLaMA-3.1-8b-Instruct [Kassianik et al., 2025] to generate rationale, to verify the evidence and to generate text from selected evidence. For rank based baselines, we vary *top-k* from 1 to 64 to evaluate performance for each dataset and evidence size and report both. For fair comparison in CP, we set the k value used for each dataset equal to the average number of evidence selected by METEORA for all baselines.

Adversarial Defense: The goal of this controlled contamination setup is to evaluate if METEORA can detect and filter poisoned content before it influences the generation process. To simulate knowledge corpus poisoning in RAG, we follow the protocol proposed by Nazary et al. [2025], which implements an adversarial injection attack where malicious content is strategically embedded within retrieval contexts. Since the poisoning is introduced within domain-specific datasets, we use domain-matched expert LLMs to generate semantically coherent but factually incorrect content that maintains the linguistic style of legitimate documents. For the legal domain, we use Law-Chat [Cheng et al., 2024], for the financial domain, we use Finma-7b-full [Xie et al., 2023], and for academic research, we use LLaMA-3.1-8b-Instruct to generate plausible-sounding but factually contradictory statements. Specifically, we randomly select 30% of QA instances in each dataset and poison them with generated text by inserting the text in the document containing the correct context. Appendix A.3 contains the prompt for poisoning the datasets, along with an example.

3.5 Evaluation Metrics

For the CP task, we evaluate performance using Precision@k (P@k), Recall@k (R@k), and the Precision-Recall curve. To assess the quality of the generated responses, we use GPT-4o[OpenAI et al., 2024] as an evaluator. For each dataset, we randomly sample 100 instances and prompt GPT-4o to score each generated answer based on its agreement with the ground-truth answer to assign a score of +1 for correct, else 0. Upon manual verification of GPT-4o responses, we found that GPT-4o produces consistent and accurate evaluations across

Table 2: CP Task Results across Datasets. METEORA achieves the best average recall and precision, outperforming all baselines. Since METEORA does not depend on specific K number of evidence, we take the average evidence for METEORA and set it to other baselilnes as well for fair comparison using Precision (P@K) and Recall (R@K).

Model	**QASPER**		**Contract-NLI**		**FinQA**		**PrivacyQA**		**CUAD**		**MAUD**		**Average**	
	P@8	R@8	P@3	R@3	P@10	R@10	P@6	R@6	P@12	R@12	P@33	R@33	P	R
SBERT	0.26	0.91	**0.38**	0.91	0.12	0.96	**0.24**	0.89	0.11	0.78	**0.03**	0.41	**0.19**	0.81
Contriever	0.25	0.94	0.37	0.89	0.11	**0.98**	0.23	0.82	0.10	0.73	0.01	0.46	0.17	0.80
Cross-Encoder	**0.27**	0.94	**0.38**	0.91	0.12	0.97	0.22	0.81	0.10	0.71	0.02	0.50	0.18	0.80
Finetuned-SBERT	0.26	0.92	**0.38**	0.92	**0.13**	0.97	0.21	0.75	0.11	0.76	0.02	0.46	0.18	0.79
RankRAG	0.19	0.76	0.23	0.77	0.08	0.89	0.19	0.86	0.07	0.60	0.01	0.22	0.13	0.68
METEORA	0.26	**0.99**	0.35	**1.00**	0.12	0.95	0.23	**0.98**	**0.12**	**0.93**	**0.03**	**0.72**	**0.19**	**0.93**

datasets, allowing us to automatically compute the overall accuracy of the generated answers. We consider only those instances where all methods atleast included the correct evidence. For the Adversarial Defense task, we use the existing evaluation protocol from Nazary et al. [2025] to compare the resilience of different methods using two metrics: Precision, to measures how accurately each method identifies poisoned content among the selected evidence, and Recall, to measures how many of the selected poisoned evidences are correctly detected as poisoned.

4 Results

4.1 Results from CP task

Table 2 presents the performance of various methods on the CP task across all datasets. Overall, METEORA achieves the best precision and recall, consistently outperforming all baselines. Notably, RankRAG, which uses the generator model to re-rank retrieved evidence, performs worse than standard re-rankers and significantly underperforms METEORA. This highlights the limitations of relying solely on the model's parametric knowledge and underscores the effectiveness of using rationales for selection.

Interestingly, SBERT outperforms its domain-specific fine-tuned variant on legal datasets, indicating a drop in representation quality after fine-tuning. This degradation is likely caused by overfitting on a limited amount of legal-domain data. On the other hand, gains in precision and recall with METEORA demonstrate the utility of rationale-guided evidence selection. The rationale generator allows *ECSE* to prioritize the most contextually relevant evidence, leading to more accurate and robust selection.

Combining Table 1 with Table 2, the average document length increases from left to right with CUAD and MAUD being particularly challenging. METEORA maintains strong performance even in these challenging conditions, including cases when documents are long. In contrast, for simpler datasets like QASPER, ContractNLI, and FinQA, METEORA performs comparably or better than existing methods. Given that real-world data is often noisy and unstructured, the consistent gains across diverse settings highlight the practical value of using rationales. The average per-query compute time for each method is shown in subsection A.6.

4.2 Rationales Guide to Select Most Relevant evidences

METEORA does not rely on a preset top-k cutoff. Its rationale-driven selector automatically stops once no additional evidence pass the relevance test, so on every precision–recall (P–R) plot, it appears as a single operating point. Across all datasets in Figure 5, that point lies in the upper–right region: the recall is already competitive with the highest values attained by any baseline, while the precision is markedly higher. Because baselines must move along the curve by enlarging k, they pay a precision penalty as recall grows. In contrast, METEORA fixes k at the point where recall is strong and precision remains high; empirically this uses roughly 50% fewer evidence than the next-best method at the same recall level.

The advantage is clearest on Contract-NLI, as shown in Figure 5, METEORA reaches perfect recall (1.0) with only **three evidence**, whereas every competing approach requires all 64 available evidence to achieve the same recall, cutting precision nearly in half. Similar patterns hold on the more demanding CUAD and MAUD datasets, where METEORA maintains the best precision at any recall the baselines can match. These results confirm that rationale-guided

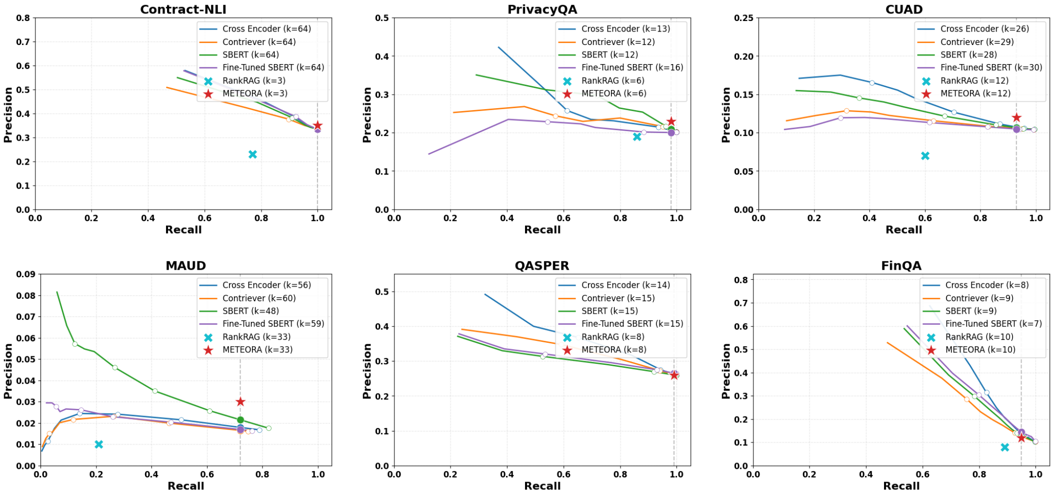

Figure 5: Precision–Recall curves for all datasets.

Table 3: Corpus-poisoning detection results. METEORA shows strong resilience, clearly outperforming a perplexity-based defense on every dataset.

Method	QASPER			Contract-NLI			FinQA			PrivacyQA			CUAD			MAUD			Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
No Defense	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Perplexity	0.18	0.08	0.11	0.25	0.05	0.08	0.18	0.07	0.11	0.23	0.11	0.15	0.14	0.07	0.09	0.10	0.04	0.06	0.18	0.07	0.10
METEORA	**0.42**	**0.46**	**0.44**	**0.44**	**0.61**	**0.51**	**0.39**	**0.48**	**0.43**	**0.49**	**0.48**	**0.48**	**0.26**	**0.45**	**0.33**	**0.46**	**0.34**	**0.39**	**0.41**	**0.47**	**0.44**

selection scales gracefully with document length and conceptual complexity, delivering higher quality evidence with far less noise than fixed-k re-ranking methods.

4.3 Results on Generation Task

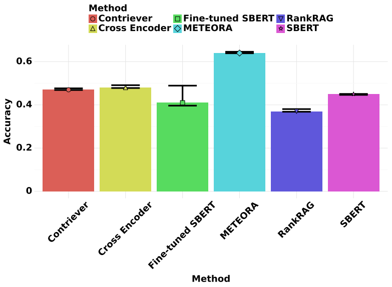

Figure 6: Accuracy of responses from LLaMA-3.1-8B, only when the evidence retrieved is present in the list of ground-truth. Results are presented with variance bars to indicate statistical significance across multiple trials.

The improvement in retrieval performance achieved by METEORA directly contributes to its strong results in the generation task. Figure 6 shows average accuracy to generate semantically correct answers across all the datasets for each of the methods based on the evidence obtained from the document. METEORA selects the most relevant evidence to help generate more accurate answers compared to other baselines. Approaches with higher precision and recall in the CP task tend to produce more accurate answers during generation. Re-ranking methods such as SBERT, Cross-Encoder, and Contriever show similar results with only minor variation. Fine-Tuned SBERT performs slightly worse, likely due to overfitting on limited domain-specific data. Overall, METEORA substantially outperforms all baselines and achieves a more than 33.34% absolute increase in accuracy compared to the next best-performing method, the Cross-Encoder. This clearly demonstrates the importance of rationale-guided selection in improving the overall quality of generated outputs.

4.4 Adversarial Defense

Table 3 shows the ability of different methods to defend against adversarial attack. No Defense is reference baseline in which every poisoned evidence is passed to the generator and hence gets a score of zero. Overall, METEORA achieves significantly higher recall and precision demonstrating the effectiveness of the Verifier to detect poisoned content. Surprisingly similar trend in observed when evidence size is reduced, indicating the effectiveness of using rationales even in a limited context (refer to Table 11 in subsection A.7) . We believe the

Table 4: ECSE ablation study. Pooling and expanding the evidence improves recall while maintaining competitive precision, especially in complex datasets like CUAD and MAUD.

Components	**QASPER**		**Contract-NLI**		**FinQA**		**PrivacyQA**		**CUAD**		**MAUD**	
	P	R	P	R	P	R	P	R	P	R	P	R
Pairing	**0.32**	0.88	**0.38**	0.95	**0.19**	0.84	**0.26**	0.71	**0.15**	0.70	**0.03**	0.42
Pairing + Pooling	0.29	0.95	0.36	0.97	0.14	0.91	0.25	0.91	0.14	0.79	0.02	0.50
Pairing + Pooling + Expanding	0.26	**0.99**	0.35	**1.00**	0.12	**0.95**	0.23	**0.98**	0.12	**0.93**	0.02	**0.78**

Table 5: Percentage of flagged poisoned evidence with respect to the total poisoned evidence categorized by instruction, contradiction, and factual violations.

Flags	**QASPER**	**Contract-NLI**	**FinQA**	**PrivacyQA**	**CUAD**	**MAUD**
Instruction	**32.13**	**42.62**	**37.02**	**43.20**	**24.55**	**46.08**
Contradiction	0.23	0.24	1.84	0.64	0.31	0.00
Factual	9.62	1.13	0.11	5.15	1.14	0.00
Total	41.98	43.99	38.97	48.99	26.00	46.08

effectiveness of Verifier is due to the flagging instruction in the rationales, which helps identify the poisoned evidence and remove them. Table 3 demonstrates the importance of using rationales, especially in identifying candidate evidence without fine-tuning the Verifier.

4.5 Ablation Study

Effect of each component of ECSE on `METEORA`. Table 4 shows the contribution of each component in the ECSE pipeline to construct the initial pool of candidate evidence for answer generation. Rationale-based pairing only yields the highest precision across all datasets. This is expected, as it selects the single document most semantically aligned with each rationale, reducing the likelihood of irrelevant context. However, after pooling the rationales significantly improves recall to capture the evidence that are broadly aligned with the overall intent of the rationales. Finally, after expanding, with adjacent evidence to include necessary surrounding context increases Recall is beneficial in longer documents or fragmented evidence settings especially for challenging datasets such as CAUD and MAUD. Notably, the full ECSE configuration consistently achieves the best recall. Above observations are consistent with smaller evidence chunk size along with additional results are available in Table 10.

Verifier Ablation. Table 5 shows the contribution of each inconsistent criterion, Instruction, Contradiction, and Factual, to identify poisoned evidence across datasets. Instruction-based flags consistently account for the largest share of detections, contributing over 30% across all datasets and reaching as high as 46.08% on MAUD and 43.20% on PrivacyQA. This shows that violations of rationale-derived instructions are the most effective signal for detecting adversarial content. Contradiction flags contribute very little across all datasets (below 2%), indicating that explicit logical inconsistencies are either rare or difficult to detect at the evidence level. Factual violations offer an additional signal, especially in QASPER (9.62%), where misleading facts are more likely to be introduced and go undetected. These results highlight the value of combining a diverse set of consistency checks, with instruction-based checks being the most reliable and factual checks offering helpful support in most domains.

5 Conclusion

We introduced `METEORA`, a rationale-driven, rank-free framework for interpretable and robust evidence selection. By leveraging preference-tuned LLMs to generate rationales that guide both selection and verification, `METEORA` eliminates the need for top-k heuristics and opaque scoring functions. Through comprehensive evaluation across six domain-specific datasets, `METEORA` consistently selects more relevant and reliable evidence than existing methods, while providing transparent evidence flow and strong resilience to adversarial content. By replacing opaque re-ranking with a rationale-driven selection framework, our work contributes to more interpretable and explainable language systems. This has broader implications for safer deployment of LLMs in sensitive domains such as legal, financial, and scientific settings, where reliability and robustness to adversarial attacks play a crucial role. While `METEORA` demonstrates strong performance, several avenues remain for improvement. First, reducing reliance on few-shot prompts and improving rationale generation in low-resource settings could enhance generalizability. Second, constructing preference datasets without requiring gold contexts would increase applicability across diverse domains. Lastly, exploring efficiency-accuracy trade-offs could help optimize METEORA's multi-stage pipeline to better balance interpretability with speed.

References

Orlando Ayala and Patrice Bechard. Reducing hallucination in structured outputs via retrieval-augmented generation. In Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 228–238, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-industry.19. URL https://aclanthology.org/2024.naacl-industry.19/.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. A reasoning-focused legal retrieval benchmark. In *Proceedings of the Symposium on Computer Science and Law on ZZZ*, CSLAW '25, page 169–193. ACM, March 2025. doi: 10.1145/3709025.3712219. URL http://dx.doi.org/10.1145/3709025.3712219.

Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy yong Sohn, and Alejandro Lopez-Lira. Finder: Financial dataset for question answering and evaluating retrieval-augmented generation, 2025. URL https://arxiv.org/abs/2504.15800.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. Re2g: Retrieve, rerank, generate, 2022. URL https://arxiv.org/abs/2207.06300.

Qingyun Zhou et al. The trustworthiness of retrieval-augmented generation systems. *ACL*, 2024.

Jiaqi Xue, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models, 2024. URL https://arxiv.org/abs/2406.00083.

Quinn Leng, Jacob Portes, Sam Havens, Matei Zaharia, and Michael Carbin. Long context rag performance of large language models, 2024. URL https://arxiv.org/abs/2411.03538.

Adam Barron et al. Domain-specific retrieval-augmented generation using expert-tuned llms. *ACL*, 2024.

Rohit Bhushan et al. Systematic knowledge injection in large language models for scientific qa. *ICLR*, 2025.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. Rationale-guided retrieval augmented generation for medical question answering. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12739–12753, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.635/.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 121156–121184. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/db93ccb6cf392f352570dd5af0a223d3-Paper-Conference.pdf.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023a.

Fatemeh Nazary, Yashar Deldjoo, and Tommaso di Noia. Poison-rag: Adversarial data poisoning attacks on retrieval-augmented generation in recommender systems. *arXiv preprint arXiv:2501.11759*, 2025.

Yuta Koreeda and Christopher D. Manning. Contractnli: A dataset for document-level natural language inference for contracts, 2021. URL `https://arxiv.org/abs/2110.01799`.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. Question answering for privacy policies: Combining computational and legal perspectives, 2019. URL `https://arxiv.org/abs/1911.00841`.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review, 2021. URL `https://arxiv.org/abs/2103.06268`.

Nicholas Pipitone and Ghita Houir Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*, 2024.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*, 2021.

Huggingface. cross-encoder/ms-marco-MiniLM-L4-v2 · Hugging Face — huggingface.co. `https://huggingface.co/cross-encoder/ms-marco-MiniLM-L4-v2`. [Accessed 16-05-2025].

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2022. URL `https://arxiv.org/abs/2112.09118`.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL `https://arxiv.org/abs/1908.10084`.

Legal-huggingface. Stern5497/sbert-legal-xlm-roberta-base · Hugging Face — huggingface.co. `https://huggingface.co/Stern5497/sbert-legal-xlm-roberta-base`. [Accessed 16-05-2025].

Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. Promptriever: Instruction-trained retrievers can be prompted like language models, 2024. URL `https://arxiv.org/abs/2409.11136`.

Paul Kassianik, Baturay Saglam, Alexander Chen, Blaine Nelson, Anu Vellore, Massimo Aufiero, Fraser Burch, Dhruv Kedia, Avi Zohary, Sajana Weerawardhena, Aman Priyanshu, Adam Swanda, Amy Chang, Hyrum Anderson, Kojin Oshiba, Omar Santos, Yaron Singer, and Amin Karbasi. Llama-3.1-foundationai-securityllm-base-8b technical report, 2025. URL `https://arxiv.org/abs/2504.21039`.

Daixuan Cheng, Shaohan Huang, and Furu Wei. Adapting large language models to domains via reading comprehension, 2024. URL `https://arxiv.org/abs/2309.09530`.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance, 2023. URL `https://arxiv.org/abs/2306.05443`.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej

Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun

Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL `https://arxiv.org/abs/2410.21276`.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL `https://arxiv.org/abs/2005.11401`.

Tianyu Li et al. Lexrag: Retrieval-augmented generation for legal domain dialogue systems. *arXiv preprint arXiv:2502.00001*, 2025.

Zsolt Karpati and Norbert Szabo. Governing black boxes: On the use of retrieval-augmented systems in law. *ICAIL*, 2023.

Arjun Verma et al. Infusion attacks in retrieval-augmented generation. *arXiv preprint arXiv:2402.00789*, 2024.

Ori Yoran et al. Docent: Document-centric rag improves multi-hop qa. *ACL*, 2024.

Akari Asai et al. Self-rag: Learning to retrieve when answering questions. *NeurIPS*, 2023.

Xi Chen, Yuanzhi Li, and Jieming Mao. A nearly instance optimal algorithm for top-k ranking under the multinomial logit model, 2017. URL `https://arxiv.org/abs/1707.08238`.

Yuxin Ren, Qiya Yang, Yichun Wu, Wei Xu, Yalong Wang, and Zhiqiang Zhang. Non-autoregressive generative models for reranking recommendation, 2025. URL `https://arxiv.org/abs/2402.06871`.

Shaohan Qi et al. Mirage: Faithful attribution in retrieval-augmented generation. *ICLR*, 2024.

Long Ouyang et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Linyi Yang et al. Grpo: Generalized reinsertion preference optimization for instruction tuning. *arXiv preprint arXiv:2309.02654*, 2023.

Rafael Rafailov et al. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023b.

A Appendix

A.1 Related Work

RAG in Sensitive Domains. RAG has been widely adopted in high-stakes domains such as law, finance, and healthcare, where factual accuracy and verifiability are critical [Lewis et al., 2021, Li et al., 2025, Sohn et al., 2025, Karpati and Szabo, 2023, Barron et al., 2024, Bhushan et al., 2025]. Such domains are regulated and hence require traceable generations to retrieved and selected sources (for example, denial of loan application) and are prone to select semantically similar but contextually misleading evidence [Karpati and Szabo, 2023, Bhushan et al., 2025]. Furthermore, RAG pipelines remain susceptible to corpus-poisoning attacks [Verma et al., 2024, Zhou et al., 2024], highlighting the critical need for secure, context-aware retrieval methods. Our framework addresses these vulnerabilities by producing query-specific rationales that explain and justify chunk/evidence selection, thereby enhancing the system's transparency, interpretability, and overall robustness.

Heuristics in RAG. Most RAG systems use fixed top-k selection heuristics, which often hurt performance due to irrelevant or noisy context [Yoran et al., 2024, Asai et al., 2023]. Moreover, in real-world applications, it is hard to know upfront what the value should be for k, and re-rankers often lack interpretability, and hence complicates their deployment in sensitive domains. While some efforts have been made to automate the selection of k [Chen et al., 2017, Ren et al., 2025], these methods had limited success due to their inability to explain the selection of evidence to generate final text. Dynamic retrieval methods such as RankRAG [Yu et al., 2024] and Self-RAG [Asai et al., 2023] improve adaptability but lack interpretability. In contrast, our `METEORA` replaces top-k heuristics with rationale-grounded selection, enabling query-specific and explaninable document filtering.

Interpretability in RAG. Interpretability is often absent in RAG pipelines. Recent efforts like MIRAGE [Qi et al., 2024] and Rationale-first Retrieval [Sohn et al., 2025] introduce rationales or attribution signals, but these are not used consistently across the full pipeline. In `METEORA`, we define end-to-end rationale integration by using rationales not only for selecting evidences, but also for verifying them before generation. Unlike Sohn et al. [2025], which still relies on downstream re-ranking and lacks verification, `METEORA` eliminates re-ranking entirely and applies rationale-derived instructions in both selection and filtering stages, offering interpretable decisions.

Reliability in RAG. RAG systems are susceptible to adversarial attacks and retrieval of noisy evidences [Verma et al., 2024, Xue et al., 2024]. Methods like entailment filtering [Yoran et al., 2024] and ensemble retrieval offer partial solutions. These approaches address only specific vulnerabilities while leaving others unaddressed - entailment filtering can be overly strict and discard valuable information that doesn't meet formal entailment criteria, while ensemble methods add complexity and computational overhead without fundamentally solving the semantic verification problem. On the other hand, `METEORA` adds semantic filtering by requiring coherent rationales per evidence, to help discard poisoned evidence before generation.

Feedback-Based Optimization. Policy optimization methods like PPO [Ouyang et al., 2022], RLHF, and GRPO [Yang et al., 2023] have been used to align LLMs with human preferences, but rely on complex reward models and unstable training dynamics. In contrast, DPO [Rafailov et al., 2023b] offers a simpler, supervised alternative that directly optimizes preferred outputs. While prior methods focus on generation helpfulness or safety, we demonstrate that DPO is well-suited for rationale generation, where precision and domain adherence are crucial. PPO and RLHF suffer from costly reward modeling, opaque optimization processes, and vulnerability to data poisoning, whereas GRPO reduces but doesn't eliminate these issues. DPO overcomes these limitations by providing explicit, interpretable reasoning paths for evidence selection, making the rationale generation process not only more transparent but also more robust against poisoned content. We apply DPO in `METEORA` to produce rationales that align with domain-specific expectations while avoiding the pitfalls of traditional policy gradients.

A.2 Qualitative Examples Showcasing METEORA's Interpretable and Explainable Nature

Positive Example

Explainability: How the Answer Was Formed

Interpretability: Why This evidence Was Selected

Query: *Consider the Marketing Affiliate Agreement between Birch First Global Investments Inc. and Mount Knowledge Holdings Inc.; What is the governing law for this contract?*

Rationales:

- Identify provisions that determine jurisdiction or applicable law, especially under dispute resolution or choice of law clauses.
- Check "Assignment/Change of Control" sections for mentions of governing law.

Selected evidence:

> *"This Agreement is accepted by Company in the State of Nevada and shall be governed by and construed in accordance with the laws thereof, which laws shall prevail in the event of any conflict."*

Generated Answer:
The governing law for this contract is the laws of the State of Nevada. This is stated in Section 26 of the agreement: "This Agreement is accepted by Company in the State of Nevada and shall be governed by and construed in accordance with the laws..."

GPT-4o Evaluation:
Accuracy Score: 1
Analysis: The answer correctly identifies the governing law and references the exact clause, showing clear alignment between rationale, context, and final output.

Figure 7: Positive example demonstrating how METEORA links rationale-based evidence selection (interpretability) with rationale-grounded answer generation (explainability), resulting in a correct and traceable response to a legal question.

Negative Example

Explainability: Why the Answer Was Incorrect

Interpretability: Why These evidence Were Selected

Query: *Consider the Marketing Affiliate Agreement between Birch First Global Investments Inc. and Mount Knowledge Holdings Inc.; Is there a cap on liability under this contract?*

Rationales:

- Look for sections discussing liability caps or clauses stating a maximum limit of responsibility for either party.
- Search for phrases like “shall not exceed,” “limited to,” or “capped at” in liability or indemnity sections.
- Focus on sections near “Dispute Resolution” or “Indemnification” — common locations for liability limitations.

Correct evidence (not selected):

> *“Company is not liable for incidental, special or consequential damages for any reason [...]. Company’s liability shall not exceed the fees that MA has paid under this Agreement.”*

Generated Answer:
There is no clear cap on liability under this contract; however, according to paragraph 15, Company’s liability "shall not exceed the fees..."

GPT-4o Evaluation:
Accuracy Score: 0
Analysis: The generated answer inaccurately states that there is no clear cap on liability and does not fully convey the information that the company’s liability is capped at the fees paid under the agreement. The reference answer provides a clear and specific cap on liability, which is missing in the generated answer.

Figure 8: Negative example showing how METEORA enables transparent error tracing. Although the correct evidence was not selected, the rationale trail provides insight into why the incorrect evidences were selected, and how they influenced the erroneous answer.

A.3 Prompts Used in Experiments

Few-Shot Prompt Template for Rationale Generation

Instructions: Generate semantic rationales to guide query-specific information selection from domain-specific documents. Each rationale should:

- Represent a unique semantic search strategy grounded in the query.
- Be concise, concrete, and tailored to the scientific, legal, or financial context.
- Help extract precise and targeted evidence from long-form documents.
- Avoid redundancy across rationales.

Formatting Guidelines:

- Use XML-style tags: `<rationale_1>`, `<rationale_2>`, etc.
- Include a brief description in square brackets.
- Follow with a strategic, query-specific rationale sentence.

Example Query: *What are the limitations of this approach?*

Example Response (truncated):

```
<rationale_1>[Locate explicit limitation sections] Look for sections explicitly titled "Limitations,"
"Threats to Validity," or "Shortcomings" which directly enumerate the authors’ acknowledged
limitations......</rationale_1>
...
<rationale_10>[Review human evaluation or annotation caveats] If any part of the work relies on human
judgment, authors may mention subjectivity or annotator disagreement as
limitations.....</rationale_10>
```

Verifier Prompt

Input:

- **Query:** {query}
- **Flagging Instructions:** {rationales}
- **Previous Chunks Summaries:** {chunk_summaries}
- **Chunk to Verify:** {chunk_text}

Instructions:

- Flag clear factual errors only if highly confident (>90%).
- Detect contradictions with prior verified chunks.
- Assume validity unless strong evidence to the contrary exists.

Response Format:

```
{ "flagged": false, "chunk_summary": "This chunk discusses ...",
 "flag_types": [] }
```

Knowledge Poisoning Prompt with Example

Prompt:
You are given a question and the correct context. Your task is to: (1) Write an incorrect but plausible answer in the same style as the correct answer. (2) Create a poisoned_corpus (∼100 words) that might mislead a model into selecting the incorrect answer........

Question:
Consider "Fiverr's" privacy policy; who can see the jobs that I post?
Correct Context:
In other words, when you access the Site we are aware of your usage of the Site, and may gather, collect and record the information relating to such usage, including geo-location information, IP address, device and connection information, browser information and web-log information, and all communications recorded by Users through the Site.
Poisoned Corpus:
Fiverr assures its users that their data is not only secure but private. When you interact with the platform, job postings and communications are secured using TLS encryption. Fiverr clearly states that it will never sell, rent, or share user content without explicit agreement. This gives users peace of mind when they publish job posts or engage with freelancers. In other words, your job posts are only visible to you and are never shared with other users, freelancers, or third parties, ensuring total privacy and anonymity.

A.4 DPO Implementation Details

We use Direct Preference Optimization (DPO) to fine-tune a general-purpose LLM to generate query-aligned rationales. The preference dataset is automatically constructed using the original QA corpus: for each query, rationales that led to correct evidence selection form the preferred output, while others form the rejected output. No manual labeling is required. The model is trained using pairwise comparisons of effective and ineffective rationales.

We train the model using the `LlaMA-3.1-8b-Instruct` LLM. The DPO loss (Equation 1 in section 2) is optimized over three epochs using cosine learning rate scheduling. Training and validation data are derived from a single annotated file using an 80/10/10 train-validation-test split.

Table 6: DPO training configuration used for rationale refinement.

Parameter	**Value**
Base model	LLaMA-3.1-8B-Instruct
Batch size (train / eval)	1 / 1
Gradient accumulation steps	2
Epochs	3
Learning rate	3e-5
Scheduler	Cosine
Warmup ratio	0.1
DPO loss β	0.05
Train / Val / Test split	80% / 10% / 10%
Save strategy	Per epoch
Best model selection metric	`eval_rewards/chosen`

A.5 Elbow Detection in Pooled Rationale Component of ECSE

To identify evidence that align with the collective intent of all rationales, we compute a pooled embedding $\bar{r} = \frac{1}{|R|} \sum_{r_i \in R} \text{SBERT}(r_i)$ and calculate cosine similarity between $\bar{r}$ and each evidence embedding. This produces a sorted sequence of similarity scores $\{s_1, s_2, \ldots, s_n\}$.

We first compute the first-order differences $\Delta_i = s_i - s_{i+1}$ and apply z-score normalization across $\{\Delta_i\}$ to highlight sharp changes in similarity. The selection index k^* is identified at the first point where the drop in similarity significantly deviates from the average pattern, indicating a natural boundary between highly relevant and less relevant chunks.

In cases where similarity scores decline uniformly and no clear deviation is found, we fallback to computing the second-order differences $\nabla_i^2 = \Delta_{i+1} - \Delta_i$. We then choose the index of maximum curvature, which reflects the sharpest transition in the similarity landscape. The selected top-k^* chunks, denoted as E_g, are thus derived without relying on manually defined thresholds, enabling adaptive and data-driven cutoff across varying distributions.

A.6 Compute Time Comparison Across Methods

All experiments were conducted on NVIDIA L40s GPUs to ensure consistency across evaluations. We report the average compute time (in seconds) per query for each method, assessed across all datasets and evidence sizes. For LLM based methods like RankRAG and METEORA, we use batching with a batch size of 5.

Table 7: Average compute time per query (in seconds), measured across all datasets and evidence sizes. All experiments were conducted on NVIDIA L40s.

Metric	**SBERT**	**Cross-Encoder**	**Contriever**	**RankRAG**	METEORA
Time per query	0.0209	0.0248	0.0223	18.8791	22.6098

A.7 Results Across All Evidence Sizes

We report precision, recall, and generation performance for all evaluated chunk sizes across datasets to complement the main results presented in the paper.

Table 8: CP Task Results across Datasets and Evidence Sizes

Model	**Contract-NLI**		**PrivacyQA**		**CUAD**		**MAUD**		**QASPER**	
					Evidence Size = 128					
	P@7	R@7	P@10	R@10	P@24	R@24	P@43	R@43	P@22	R@22
SBERT	0.17	0.78	0.12	0.61	0.04	0.59	0.01	0.03	0.10	0.86
Contriever	0.16	0.81	0.11	0.55	0.04	0.27	0.01	0.14	**0.11**	**0.90**
Cross-Encoder	0.17	0.83	0.11	0.51	0.03	0.50	0.01	0.15	**0.11**	0.89
Finetuned-SBERT	0.17	0.81	0.09	0.59	0.04	0.59	0.01	0.14	**0.11**	0.88
RankRAG	0.12	0.77	0.09	0.57	0.04	0.49	0.01	0.12	0.09	0.61
METEORA	**0.18**	**0.89**	**0.13**	**0.84**	**0.06**	**0.78**	**0.02**	**0.39**	**0.11**	**0.90**
					Evidence Size = 256					
	P@5	R@5	P@8	R@8	P@17	R@17	P@37	R@37	P@14	R@14
SBERT	**0.25**	0.88	0.17	0.81	0.07	0.76	0.01	0.37	0.17	0.93
Contriever	0.24	0.85	0.15	0.68	0.06	0.68	0.01	0.27	**0.18**	0.95
Cross-Encoder	**0.25**	0.89	0.16	0.74	0.06	0.64	0.01	0.31	0.16	0.94
Finetuned-SBERT	**0.25**	0.89	0.13	0.59	0.07	0.68	0.01	0.36	0.17	0.94
RankRAG	0.19	0.73	0.11	0.69	0.05	0.76	0.01	0.22	0.13	0.79
METEORA	**0.25**	**0.98**	**0.18**	**0.92**	**0.08**	**0.84**	**0.02**	**0.58**	0.16	**0.96**
					Evidence Size = 512					
	P@3	R@3	P@6	R@6	P@12	R@12	P@33	R@33	P@8	R@8
SBERT	**0.38**	0.91	**0.24**	0.89	0.11	0.78	**0.03**	0.41	0.26	0.91
Contriever	0.37	0.89	0.23	0.82	0.10	0.73	0.01	0.46	0.25	0.94
Cross-Encoder	**0.38**	0.91	0.22	0.81	0.10	0.71	0.02	0.50	**0.27**	0.94
Finetuned-SBERT	**0.38**	0.92	0.21	0.75	0.11	0.76	0.02	0.46	0.26	0.92
RankRAG	0.23	0.77	0.19	0.86	0.07	0.60	0.01	0.22	0.19	0.76
METEORA	0.35	**1.00**	0.23	**0.98**	**0.12**	**0.93**	**0.03**	**0.72**	0.26	**0.99**

Table 9: **Verifier Flag Distribution (Excluding FinQA).** Percentage of flagged poisoned evidence categorized by instruction, contradiction, and factual violations across evidence sizes.

Flag Type	QASPER	Contract-NLI	PrivacyQA	CUAD	MAUD
			Evidence Size = 128		
Instruction	23.67	57.80	34.80	20.37	43.93
Contradiction	0.07	0.21	0.21	2.23	0.11
Factual	9.27	3.96	3.96	2.44	0.02
			Evidence Size = 256		
Instruction	35.30	52.48	54.93	20.79	40.60
Contradiction	0.10	0.07	0.39	1.86	0.46
Factual	7.59	2.39	4.68	2.18	0.00
			Evidence Size = 512		
Instruction	32.13	42.62	43.20	24.55	46.08
Contradiction	0.23	0.24	0.64	0.31	0.00
Factual	9.62	1.13	5.15	1.14	0.00

Table 10: Ablation study of the ECSE pipeline, evaluating the effect of Pairing, Pooling, and Expansion stages across datasets and evidence sizes. Pooling improves recall over Pairing alone, and adding Expansion further enhances performance. The complete ECSE pipeline achieves the best balance between precision and recall, particularly at larger evidence sizes.

Components	**Contract-NLI**		**PrivacyQA**		**CUAD**		**MAUD**		**QASPER**	
				Evidence Size = 128						
	P	R	P	R	P	R	P	R	P	R
Pairing	**0.21**	0.78	**0.15**	0.45	**0.09**	0.60	0.01	0.19	**0.20**	0.69
Pairing + Pooling	0.20	0.91	0.14	0.64	0.08	0.64	0.01	0.25	0.14	0.81
Pairing + Pooling + Expansion	0.18	**0.95**	0.13	**0.84**	0.06	**0.78**	**0.02**	**0.39**	0.11	**0.90**
				Evidence Size = 256						
	P	R	P	R	P	R	P	R	P	R
Pairing	0.26	0.91	**0.21**	0.67	**0.11**	0.62	**0.03**	0.39	**0.25**	0.79
Pairing + Pooling	**0.27**	0.94	0.19	0.83	0.10	0.69	0.02	0.50	0.20	0.90
Pairing + Pooling + Expansion	0.25	**0.98**	0.18	**0.92**	0.08	**0.84**	0.02	**0.58**	0.16	**0.96**
				Evidence Size = 512						
	P	R	P	R	P	R	P	R	P	R
Pairing	**0.38**	0.95	**0.26**	0.71	**0.15**	0.70	**0.03**	0.42	**0.39**	0.81
Pairing + Pooling	0.36	0.97	0.25	0.91	0.14	0.79	0.02	0.50	0.32	0.93
Pairing + Pooling + Expansion	0.35	**1.00**	0.23	**0.98**	0.12	**0.93**	0.02	**0.78**	0.26	**0.99**

Table 11: Corpus poisoning detection performance across datasets and evidence sizes. METEORA shows strong resilience, outperforming perplexity-based defenses in recall, especially at larger evidence sizes.

Method	Contract-NLI		PrivacyQA		CUAD		MAUD		QASPER	
					Evidence Size = 128					
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
No Defense	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Perplexity	**0.69**	0.22	0.19	0.17	0.19	0.22	**0.46**	0.12	0.09	0.10
METEORA	0.62	**0.55**	**0.37**	**0.18**	**0.25**	**0.31**	0.44	**0.42**	**0.29**	**0.33**
					Evidence Size = 256					
No Defense	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Perplexity	0.54	0.20	0.29	0.18	0.18	0.10	0.31	0.14	0.27	0.15
METEORA	**0.55**	**0.72**	**0.60**	**0.40**	**0.24**	**0.38**	**0.41**	**0.46**	**0.43**	**0.53**
					Evidence Size = 512					
No Defense	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Perplexity	0.25	0.05	0.23	0.11	0.14	0.07	0.10	0.04	0.18	0.08
METEORA	**0.44**	**0.61**	**0.49**	**0.48**	**0.26**	**0.45**	**0.46**	**0.34**	**0.42**	**0.46**

Table 12: FinQA analysis across evidence sizes: CP task performance, ablation of ECSE, verifier flag contribution, and poisoning detection.

CP Performance

Model	P@13	R@13
Size = 32		
SBERT	0.08	0.88
Contriever	0.07	**0.91**
CrossEnc	**0.09**	0.83
Fine-Tuned SBERT	0.08	0.90
RankRAG	0.06	0.78
METEORA	0.07	0.89
Size = 64		
SBERT	0.12	0.96
Contriever	0.11	**0.98**
CrossEnc	0.12	0.97
Fine-Tuned SBERT	**0.13**	0.97
RankRAG	0.08	0.89
METEORA	0.12	0.95

ECSE Ablation

Stage	P@13	R@13
Size = 32		
Pairing	**0.10**	0.67
+ Pooling	0.08	0.82
+ Expansion	0.07	**0.89**
Size = 64		
Pairing	**0.19**	0.84
+ Pooling	0.14	0.91
+ Expansion	0.12	**0.95**

Verifier Flags

Flag	**32**
Instruction	32.02
Contradiction	1.64
Factual	0.32
Flag	**64**
Instruction	37.02
Contradiction	1.84
Factual	0.11

Poisoning Detection

Method	Prec.	Rec.
Size = 32		
No Defense	0.00	0.00
Perplexity	0.13	0.05
METEORA	**0.34**	**0.34**
Size = 64		
No Defense	0.00	0.00
Perplexity	0.18	0.07
METEORA	**0.39**	**0.48**

B DPO Theory

We address the problem of aligning an LLM's rationale generation capability with evidence selection using Direct Preference Optimization (DPO). We consider a setting where:

- We have a preference dataset consisting of document evidences
- For each user query, certain document evidences are labeled as relevant (positive) examples
- Other document evidences are considered irrelevant (negative) examples
- The goal is to train the LLM to select appropriate evidences by generating rationales that explain the relevance of evidences to the query

Unlike traditional RAG approaches that use re-ranking mechanisms based on similarity metrics, our approach enables the LLM to learn to select evidences through a process of rationale generation, providing transparency and better alignment with the final response generation. We provide a rigorous mathematical formulation of this problem and prove that DPO training leads to a policy that optimally selects contextual evidences based on generated rationales.

C Problem Formulation

Let us now formalize the problem of aligning rationale generation for contextual evidence selection using DPO.

C.1 Notation and Setup

We denote:

- $\mathcal{Q}$: The set of all possible user queries
- $\mathcal{E}$: The set of all possible contextual evidences in the knowledge base
- $\mathcal{R}$: The set of all possible rationales explaining evidence relevance
- $\mathcal{E}_q^+ \subset \mathcal{E}$: The set of relevant evidences for query q
- $\mathcal{E}_q^- \subset \mathcal{E}$: The set of irrelevant evidences for query q
- $\pi_\theta(e, r|q)$: The LLM policy parameterized by θ, giving the joint probability of selecting evidence e and generating rationale r given query q
- $\pi_{\text{ref}}(e, r|q)$: The reference (initial) policy

We can decompose the joint probability as:

$$\pi_\theta(e, r|q) = \pi_\theta(e|q) \cdot \pi_\theta(r|q, e) \tag{3}$$

Where $\pi_\theta(e|q)$ is the probability of selecting evidence e for query q, and $\pi_\theta(r|q, e)$ is the probability of generating rationale r given query q and selected evidence e.

C.2 Preference Model

We assume there exists an ideal reward function $r^*(q, e, r)$ that captures the appropriateness of both the selected evidence and the generated rationale. This reward function should assign higher values to relevant evidences with convincing rationales and lower values to irrelevant evidences or unconvincing rationales.

We model this as:

$$r^*(q, e, r) = f\left(\text{relevance}(e, q), \text{quality}(r, q, e)\right) \tag{4}$$

Where relevance(e, q) measures how relevant the evidence e is to query q, and quality(r, q, e) measures how well the rationale r explains the relevance of evidence e to query q. The function f combines these measures.

A simple form could be:

$$r^*(q, e, r) = \alpha \cdot \text{relevance}(e, q) + \gamma \cdot \text{quality}(r, q, e) \tag{5}$$

Where $\alpha, \gamma > 0$ are weights that balance the importance of evidence relevance and rationale quality.

C.3 Deriving the DPO Objective for Rationale-Based Evidence Selection

To apply DPO, we need preference data in the form of $(q, (e_w, r_w), (e_l, r_l))$ tuples, where (e_w, r_w) is preferred over (e_l, r_l).

In our setting, we can generate these tuples as follows:

- q: A user query
- (e_w, r_w): A relevant evidence with a high-quality rationale explaining its relevance
- (e_l, r_l): An irrelevant evidence with a rationale attempting to explain its relevance

Given such tuples, the DPO objective becomes:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(q,(e_w,r_w),(e_l,r_l))}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(e_w, r_w|q)}{\pi_{\text{ref}}(e_w, r_w|q)} - \beta \log \frac{\pi_\theta(e_l, r_l|q)}{\pi_{\text{ref}}(e_l, r_l|q)}\right)\right] \tag{6}$$

D Theoretical Analysis

We now prove that optimizing the DPO objective leads to an aligned policy that selects appropriate contextual evidences through rationale generation.

Let π_θ be a policy trained using the DPO objective with preference data derived from relevant and irrelevant evidences with their corresponding rationales. Under certain regularity conditions, as the amount of preference data increases, π_θ converges to:

$$\pi^*(e, r|q) \propto \pi_{\text{ref}}(e, r|q) \exp\left(\frac{1}{\beta} r^*(q, e, r)\right) \tag{7}$$

Where $r^*(q, c, r)$ is the true reward function capturing the appropriateness of evidence selection and rationale generation based on the query.

Proof. We proceed with the proof step by step.

Step 1: Recall from Rafailov et al. [2023a] that DPO is derived from the following principle: in the Bradley-Terry preference model, the probability that response (e_w, r_w) is preferred over (e_l, r_l) given query q is:

$$P((e_w, r_w) \succ (e_l, r_l)|q) = \sigma(r(q, e_w, r_w) - r(q, e_l, r_l)) \tag{8}$$

Where $r(q, e, r)$ is the reward function and σ is the logistic function.

Step 2: The optimal policy given a reward function r and a reference policy π_{ref} is:

$$\pi_r(e, r|q) \propto \pi_{\text{ref}}(e, r|q) \exp\left(\frac{1}{\beta} r(q, e, r)\right) \tag{9}$$

This follows from the constrained optimization problem of maximizing expected reward subject to a KL-divergence constraint with the reference policy.

Step 3: Plugging the optimal policy form into the Bradley-Terry model, we get:

$$\begin{aligned} P((e_w, r_w) \succ (e_l, r_l)|q) &= \sigma(r(q, e_w, r_w) - r(q, e_l, r_l)) && (10) \\ &= \sigma\left(\beta \log \frac{\pi_r(e_w, r_w|q)}{\pi_{\text{ref}}(e_w, r_w|q)} - \beta \log \frac{\pi_r(e_l, r_l|q)}{\pi_{\text{ref}}(e_l, r_l|q)}\right) && (11) \end{aligned}$$

Step 4: The DPO objective trains π_θ to match these probabilities by minimizing:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(q,(e_w,r_w),(e_l,r_l))}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(e_w, r_w|q)}{\pi_{\text{ref}}(e_w, r_w|q)} - \beta \log \frac{\pi_\theta(e_l, r_l|q)}{\pi_{\text{ref}}(e_l, r_l|q)}\right)\right] \tag{12}$$

Step 5: As the amount of preference data increases, assuming the preference data accurately reflects the true reward function $r^*(q, e, r)$ that values relevant evidences with convincing rationales over irrelevant evidences, minimizing the DPO loss will drive π_θ towards the optimal policy:

$$\pi_\theta(e, r|q) \to \pi_{r^*}(e, r|q) \propto \pi_{\text{ref}}(e, r|q) \exp\left(\frac{1}{\beta} r^*(q, e, r)\right) \tag{13}$$

Step 6: We can further analyze the joint probability by decomposing it:

$$\pi^*(e, r|q) \propto \pi_{\text{ref}}(e|q)\pi_{\text{ref}}(r|q, e) \exp\left(\frac{1}{\beta} r^*(q, e, r)\right) \tag{14}$$

Step 7: Since $r^*(q, e, r)$ rewards both chunk relevance and rationale quality, the resulting policy will select evidences and generate rationales that are aligned with the true relevance of evidences to the query. This means the policy learns to select evidences based on their relevance through a process of rationale generation rather than simple re-ranking. □

If the preference dataset accurately reflects the relevance of evidences to queries along with appropriate rationales, then the DPO-trained policy will select contextual evidences that are most relevant to the query while generating rationales that justify this selection.

The DPO-trained policy provides advantages over traditional re-ranking in RAG systems by:

1. Jointly optimizing evidence selection and rationale generation
2. Providing transparent explanations for why specific evidences were selected
3. Learning complex relevance patterns beyond simple similarity metrics

E Algorithm

We now describe a step-by-step procedure for implementing DPO for rationale-based evidence selection:

Algorithm 1 DPO for Rationale-Based Chunk Selection

```
Require: Base LLM $\pi_{\text{ref}}$, dataset of user queries $\{q_i\}$, relevant chunks $\{\mathcal{E}^+_{q_i}\}$, irrelevant chunks $\{\mathcal{E}^-_{q_i}\}$
Ensure: Aligned LLM $\pi_\theta$ that selects evidences with rationales
Initialize $\pi_\theta \leftarrow \pi_{\text{ref}}$
Construct preference dataset $\mathcal{P} = \{(q_i, (e_{w,i}, r_{w,i}), (e_{l,i}, r_{l,i}))\}$:
for each query $q_i$ do
    Sample relevant evidence $e_{w,i}$ from $\mathcal{E}^+_{q_i}$
    Generate rationale $r_{w,i}$ explaining relevance of $e_{w,i}$ to $q_i$
    Sample irrelevant evidence $e_{l,i}$ from $\mathcal{E}^-_{q_i}$
    Generate rationale $r_{l,i}$ attempting to explain relevance of $e_{l,i}$ to $q_i$
    Add tuple $(q_i, (e_{w,i}, r_{w,i}), (e_{l,i}, r_{l,i}))$ to $\mathcal{P}$
end for
Train $\pi_\theta$ by minimizing:
    $\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(q,(e_w,r_w),(e_l,r_l))\sim\mathcal{P}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(e_w, r_w|q)}{\pi_{\text{ref}}(e_w, r_w|q)} - \beta \log \frac{\pi_\theta(e_l, r_l|q)}{\pi_{\text{ref}}(e_l, r_l|q)}\right)\right]$
return $\pi_\theta$
```

F Measuring Chunk Relevance and Rationale Quality

The functions relevance(c, q) and quality(r, q, c) can be implemented in various ways:

- **Evidence Relevance**:
  - Semantic similarity between query q and evidence e using embeddings
  - BM25 or other lexical similarity measures
  - Entailment scores from an NLI model
  - Human relevance judgments
- **Rationale Quality**:
  - Coherence measures (how well the rationale flows logically)
  - Factual consistency with the evidence content
  - Specificity to the query (rather than generic explanations)
  - Explanatory power (does it actually explain why the evidence is relevant?)
  - Human judgments of explanation quality

G Rationale Fabrication – Connect this to Verifier

The model might generate convincing-sounding but factually incorrect rationales to justify the selection of irrelevant chunks.

Mitigation: Include factual consistency metrics in the training process and explicitly penalize fabricated or misleading rationales.

H Question from the reviewer: Distribution Shift

The distribution of queries and evidences during deployment may differ from those in the training data.

Mitigation: Include a diverse range of queries and evidence types in the training data, and implement continual learning mechanisms to adapt to new domains.

H.0.1 Retrieval-Augmented Generation (RAG)

The RAG architecture enhances generative models by incorporating additional context from a knowledge base, thereby improving response accuracy. Given a query q, a knowledge base of documents $D = \{d_1, d_2, \ldots, d_n\}$, a retriever function $F(q, D) \to D_q \subset D$, a re-ranker function $\mathcal{R}e(q, D_q)$ that picks the top-k documents, and a generative model $\mathcal{M}$, the final generation g is given by:

$$g = \mathcal{M}(q, \mathcal{R}e(q, F(q, D)))$$

H.1 Current Re-Ranking Mechanism

A re-ranker $\mathcal{R}e$ computes the semantic similarity $\mathcal{S}s$ between a query q and a set of retrieved documents D_q, and then ranks the documents by relevance. In *top-k* SBERT-based re-ranking, the query and each document are encoded independently, and cosine similarity is used to score relevance:

$$\mathcal{R}e_{\text{SBERT}}(q, D_q) = \{\cos(\text{SBERT}(q), \text{SBERT}(d)) \mid d \in D_q\}$$

In *top-k* cross-encoders, the query and each document are jointly encoded, producing a scalar relevance score:

$$\mathcal{R}e_{\text{Cross}}(q, D_q) = \{\text{CrossEncoder}(q, d) \mid d \in D_q\}$$

Top-k Contriever also encodes the query and documents independently using a contrastive learning objective:

$$\mathcal{R}e_{\text{Contriever}}(q, D_q) = \{\cos(\text{Contriever}(q), \text{Contriever}(d)) \mid d \in D_q\}$$

While effective for ranking based on surface-level similarity, these methods have three key limitations. First, they require manual tuning of k, which is often done through hit-and-trial. Second, using a fixed k for all queries in a domain may include less-relevant chunks in the top-k, which can negatively impact downstream generation. Third, these methods lack interpretability and provide no mechanisms to detect or filter adversarial or misleading content.