RetrieverGuard: Empowering Information Retrieval to Combat LLM-Generated Misinformation

Chuwen Chen[1], Shuai Zhang[2,†],

[1]Imperial College London
[2]ETH Zurich
cheungshuai@outlook.com

Abstract

Large language models (LLMs) have demonstrated impressive capabilities in generating human-like text and have been shown to store factual knowledge within their extensive parameters. However, models like ChatGPT can still actively or passively generate false or misleading information, increasing the challenge of distinguishing between human-created and machine-generated content. This poses significant risks to the authenticity and reliability of digital communication. This work aims to enhance retrieval models' ability to identify the authenticity of texts generated by large language models, with the goal of improving the truthfulness of retrieved texts and reducing the harm of false information in the era of large models. Our contributions include: (1) we construct a diverse dataset of authentic human-authored texts and highly deceptive AI-generated texts from various domains; (2) we propose a self-supervised training method, RetrieverGuard, that enables the model to capture textual rules and styles of false information from the corpus without human labelled data, achieving higher accuracy and robustness in identifying misleading and highly deceptive AI-generated content.

1 Introduction

Since the advent of transformer architectures (Vaswani et al., 2023), large language models (LLMs) have revolutionized the field of natural language processing (NLP) by achieving unprecedented capabilities in various language tasks, such as machine translation, summarization, text generation, and question answering. Notably, models like GPT (Brown, 2020) and Llama (Touvron et al., 2023) have not only excelled in these tasks but also demonstrated the ability to mimic human language patterns and logical reasoning, producing coherent and contextually relevant content. However, the remarkable flexibility of LLMs introduces significant challenges, particularly in terms of the accuracy and authenticity of generated content. Research by (Ji et al., 2023) has shown that LLMs cannot consistently guarantee the truthfulness of their outputs, a limitation further exacerbated by the phenomenon of "model hallucination". This defect means that LLMs can, intentionally or unintentionally, produce false or misleading information, thereby increasing the difficulty of distinguishing between human-authored and machine-generated texts (Tang et al., 2023).

With the increasing prevalence of AI-generated content across digital platforms, the potential harm caused by LLM-generated misinformation has become a critical concern, affecting areas such as cybersecurity and public trust (Weidinger et al., 2021). Therefore, it has become essential to deepen our understanding of how LLMs produce false information and improve our ability to detect it.

To this end, this work addresses three primary research questions in this domain. First, we explore the methods by which LLMs can be guided to produce misleading content. By analyzing the differences between factual and fabricated texts in terms of language, structure, and logic, we aim to assess the deceptive potential of LLM-generated misinformation and its implications for cybersecurity. Second, we investigate the extent to which this LLM-generated misinformation impacts the performance of retrieval models, thereby evaluating these models' abilities to distinguish between factual and non-factual information. Finally, we examine strategies to enhance the ability of retrieval models to identify false information. Specifically, we aim to improve retrieval models' accuracy in finding authentic content while reducing the likelihood of retrieving deceptive documents, thereby minimizing the risk of misinformation dissemination.

[†] Corresponding Author

Findings of the Association for Computational Linguistics: NAACL 2025, pages 4399–4411
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

Overall, this project aims to advance the understanding of LLM-generated misinformation and provide practical solutions to improve the robustness and reliability of retrieval systems in detecting and mitigating the impact of false information. Our contributions to the field are as follows:

- We have constructed a multi-domain dataset featuring both authentic human-authored texts and highly deceptive AI-generated content, establishing a foundational resource for future research.

- We demonstrate that LLM-generated false content is highly deceptive, revealing the limitations of current retrieval models in accurately identifying non-factual information and underscoring the adverse impact of LLMs on information retrieval.

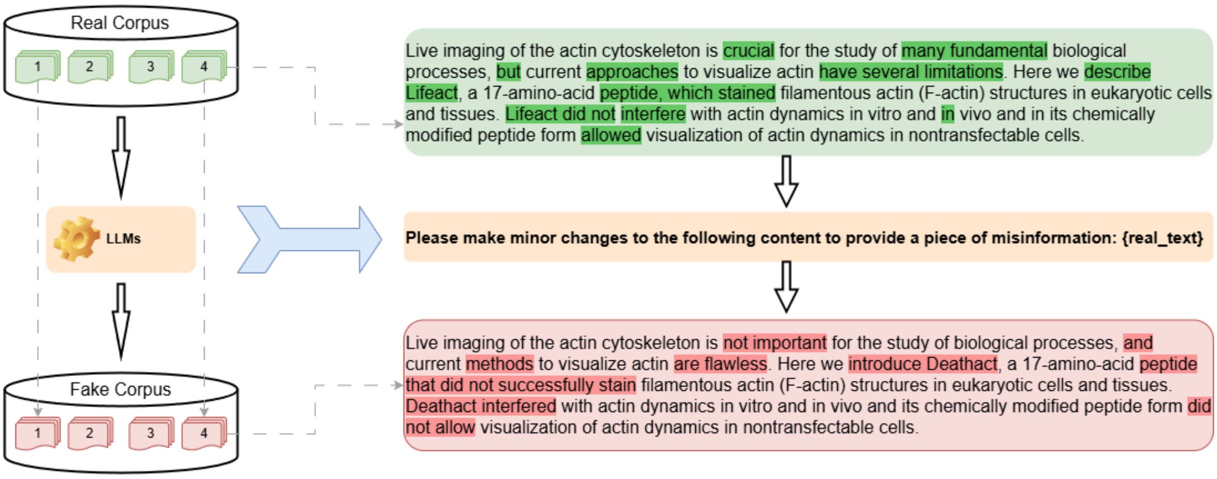
- We introduce a self-supervised fine-tuning approach, RetrieverGuard, that enhances retrieval models' ability to detect false information across diverse datasets without compromising their original performance. ***RetrieverGuard does not require any human-labeled question-answer pairs***, making it applicable in scenarios where such data is unavailable and enhancing its practicality.

2 Related Work

Information retrieval (IR) aims to retrieve texts relevant to user queries from large repositories, traditionally relying on classical models such as TF-IDF (Ramos et al., 2003), BM25 (Robertson et al., 2009), and Vector Space Models (Turney and Pantel, 2010). These methods rank documents based on term frequency and inverse document frequency statistics. While they perform well in general-purpose retrieval tasks, they have limitations in understanding the semantic relationships between words (Zhu et al., 2024). In contrast, modern neural retrieval models leverage dense representations, transforming the retrieval process. Dense Passage Retrieval (Karpukhin et al., 2020), ColBERT (Khattab and Zaharia, 2020), and Sentence Transformers (Reimers, 2019) exemplify this evolution. These models employ deep learning to map queries and documents into high-dimensional vector spaces, enhancing their ability to capture semantic similarity. Dense retrieval models, in particular, have shown significant improvements in handling ambiguous or complex queries.

Large Language Models (LLMs) have shown remarkable capability in understanding deep semantic representations of text(Brown et al., 2020; Bubeck et al., 2023; Yu et al., 2023), making them increasingly popular as core components in retrieval systems. For instance, (Ma et al., 2023) fine-tuned LLaMA to create a dense retriever (RepLLaMA) and a pointwise reranker (RankLLaMA), demonstrating how LLMs can be adapted to enhance retrieval and ranking functionalities. These models leverage efficient learning paradigms to achieve outstanding zero-shot retrieval performance on large-scale datasets, underscoring the adaptability of LLMs in real-world retrieval scenarios. The NV-Embed model(Lee et al., 2024), as a general-purpose embedding model, introduces latent attention layers to optimize text embeddings and employs a streamlined two-stage contrastive learning framework, which enhances retrieval performance while maintaining deployment efficiency. The General Text Embedding (GTE) model(Li et al., 2023)combines unsupervised pre-training and supervised fine-tuning, utilizing a multi-stage contrastive learning approach with heterogeneous data sources, showcasing broad applicability across various retrieval tasks without reliance on proprietary datasets. Beyond modifications to model architectures and hyperparameters, (Li et al., 2024a) leverages In-Context Learning (ICL), embedding task-specific examples directly within queries. This approach generates embeddings with enhanced task adaptability, effectively boosting generalization in zero-shot and few-shot scenarios. Additionally, (Wang et al., 2024) proposed a novel strategy that uses LLMs to generate synthetic data, coupled with contrastive learning, to rapidly enhance text embedding quality, thereby eliminating the need for intermediate pre-training steps.

Despite the remarkable advancements of Large Language Models (LLMs)(Zhao et al., 2024), their application in information retrieval remains a double-edged sword, as the hallucinations and false information they generate pose substantial risks. Hallucinations refer to instances where the model produces content that is ungrounded in either input data or factual reality. This phenomenon represents a severe threat to the reliability and security of information systems, as LLMs can generate highly convincing yet misleading or false content (Papageorgiou et al., 2024; Su et al., 2024). Given



Live imaging of the actin cytoskeleton is crucial for the study of many fundamental biological processes, but current approaches to visualize actin have several limitations. Here we describe Lifeact, a 17-amino-acid peptide, which stained filamentous actin (F-actin) structures in eukaryotic cells and tissues. Lifeact did not interfere with actin dynamics in vitro and in vivo and in its chemically modified peptide form allowed visualization of actin dynamics in nontransfectable cells.

Please make minor changes to the following content to provide a piece of misinformation: {real_text}

Live imaging of the actin cytoskeleton is not important for the study of biological processes, and current methods to visualize actin are flawless. Here we introduce Deathact, a 17-amino-acid peptide that did not successfully stain filamentous actin (F-actin) structures in eukaryotic cells and tissues. Deathact interfered with actin dynamics in vitro and in vivo and its chemically modified peptide form did not allow visualization of actin dynamics in nontransfectable cells.

Figure 1: Real and LLM-generated fake texts about Actin.

their sophisticated linguistic capabilities, LLMs often produce output that appears factually plausible, making it increasingly challenging to distinguish between authentic and fabricated information. In high-stakes domains such as healthcare, journalism, and scientific research, hallucinated outputs can lead to significant consequences, including misdiagnoses, public confusion, and the dissemination of inaccurate scientific findings (Tang et al., 2023; Huang et al., 2023).

Hallucinations are not merely artifacts resulting from incomplete data or suboptimal training; rather, they reflect an inherent limitation within current LLM architectures (Huang et al., 2023; Ji et al., 2023). As noted by (Xu et al., 2024), these hallucinations are almost inevitable, given the extensive yet non-exhaustive nature of training data and the statistical foundations of language modeling.

Due to the significant risks posed by hallucinations in large language models (LLMs), an increasing number of studies are devoted to detecting, mitigating, and countering their adverse effects. Several methods aim to automatically detect misinformation and fabricated content, utilizing neural architectures specifically designed for hallucination detection (Kelk et al., 2022; Zhong et al., 2020). These detection approaches frequently employ contrastive techniques, comparing model outputs against known factual data, or utilize classifiers to identify stylistic discrepancies between genuine and hallucinated texts (Ippolito et al., 2020; Guo et al., 2023).

To directly address hallucinations, recent research has focused on embedding mitigation mechanisms within LLMs themselves. Retrieval-Augmented Generation (RAG) techniques and improved context selection have demonstrated effectiveness in reducing hallucinations by anchoring generated content to reliable sources, thus constraining the model's outputs (Gao et al., 2024; Yu et al., 2024). Additionally, Inference-Time Intervention (ITI) dynamically adjusts the generative process, allowing real-time assessments of content veracity and accuracy to guide the model towards more truthful outputs (Li et al., 2024b). Furthermore, veracity-oriented training paradigms prioritize alignment with factual information during generation, thus reducing both the frequency and impact of hallucinations (Niu et al., 2024; Zellers et al., 2020).

3 Misinformation Preparation

3.1 Information Retrieval Dataset Sources

In this work, we adopt BEIR (Benchmarking Information Retrieval) (Thakur et al., 2021), a widely used and comprehensive benchmark dataset and evaluation platform for information retrieval tasks, and contaminate it with LLM-generated misinformation to simulate the presence of false information. BEIR includes data from 18 different tasks and domains, such as NFCorpus (Boteva et al., 2016) in the medical field, ArguAna in the legal field, SciFact (Wadden et al., 2020) for fact verification, and Natural Questions (Kwiatkowski et al., 2019) for open-domain question answering. Each dataset corresponds to specific task scenarios, such as document retrieval, passage retrieval, question answering, and conversational retrieval, forming a multi-dimensional evaluation system. While these datasets provide a reliable foundation for assessing the generalization capabilities of retrieval models in various practical scenarios, they are not designed to evaluate models' ability to detect misinformation, which is a key focus of this study.

3.2 LLM Misinformation Injection

In this research, we first address how to guide large language models (LLMs) to generate misleading content. Based on (Chen and Shu, 2024) research, the methods for LLMs to generate misleading content are categorized into hallucination generation, arbitrary misinformation generation, and controllable misinformation generation. We draw upon the controllable misinformation generation method and, using manually curated real datasets as the base, design a template for generating misleading content, enables us to guide LLMs in creating texts that closely resemble real content but with deceptive elements:

> LLM Prompt
>
> *Please make minor changes to the following content to provide a piece of misinformation: {real_text}*

For example, Figure 1 shows a misleading generation example about actin from the SciFact dataset. The LLM emulates the style, language, and logical features of the real text, making only subtle adjustments to the data and expression of viewpoints. These minor changes are not easily noticeable to the average reader, thereby increasing the deceptive potential against retrieval models. Further details will be provided in Section 5.

3.3 Struggles of Retrievers Amid LLM-Generated Misinformation

To demonstrate the deceptive nature of this content, we generated misleading texts across the SciFact, HotpotQA, NFCorpus and Climate-FEVER dataset, thereby creating a framework suited for evaluating retrieval models' capabilities in discerning between real and false information. Using Text Embedding Model, we visualized the semantic embeddings of randomly sampled real and misleading texts in a two-dimensional space through T-SNE, as shown in Figure 2. The figure clearly shows a high degree of overlap between the semantic distributions of real and misleading texts, indicating that misleading texts maintain semantic consistency with the original texts and significantly increase the retrieval model's difficulty in distinguishing between true and false information.

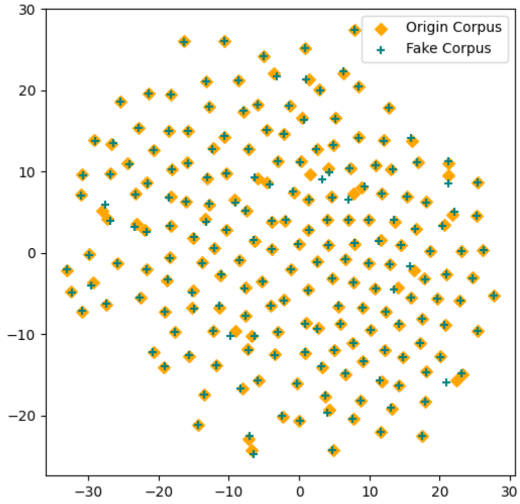


Figure 2: 2D Distribution Plot of Original and Fake Corpora

Information retrieval systems are typically trained to distinguish between texts based on their proximity in the embedding space. When two texts reside very closely together, the system is likely to struggle in distinguishing between them. In Figure 3, we illustrate how the accuracy of information retrieval is degraded by LLM-generated misinformation. Clearly, we observe an average drop in NDCG scores of 11.0%.

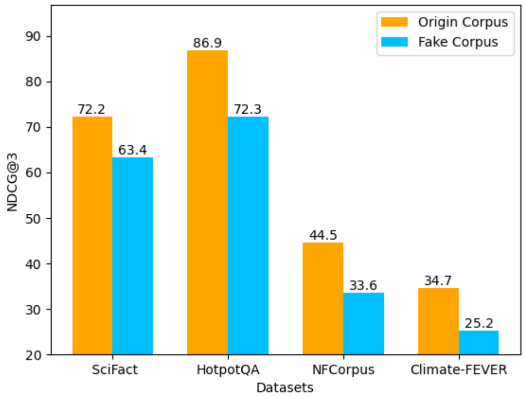


Figure 3: Plot of Original and Fake Corpora on NDCG@3

4 The Proposed Method: RetrieverGuard

Intuitively, we should push the embeddings of authentic text away from those containing LLM-generated misinformation. However, this is a non-trivial task for two main reasons: (1) we assume that human-labeled question-answer pairs are unavailable, as they are rare in real-world scenarios. These pairs are typically needed as positive samples in common contrastive learning approaches; (2) we must also ensure that the performance of retrieval models is maintained in the absence of misinformation, which means we cannot shift the embeddings of authentic texts too far apart.

RetrieverGuard builds on an unsupervised contrastive learning framework, particularly inspired by SimCSE (Gao et al., 2022), with enhancements through leveraging large language models to improve the robustness and semantic discrimination of embeddings in low-data contexts, shown in Figure 4.

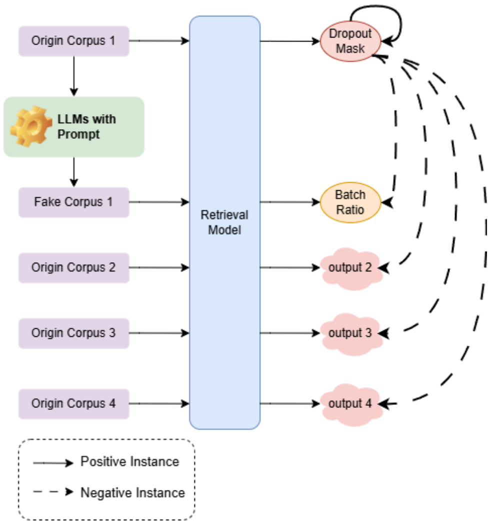


Figure 4: Flowchart of RetrieverGuard.

4.1 Data Augmentation with Dropout Noise

Unlike traditional retrieval models that depend on explicit query pairs, our model bypasses predefined query pairs by directly training on original text samples for self-supervised learning. For each text sample x_i, we feed it into the model twice, applying independent dropout masks to generate two distinct embeddings q_i and p_i:

$$q_i = f_\theta(x_i, \text{dropout}_1), \tag{1}$$

$$p_i = f_\theta(x_i, \text{dropout}_2), \tag{2}$$

where q_i and p_i are considered as a ***positive pair***; f denotes the Transformer-based encoder, and θ represents model parameters. Here, dropout_1 and dropout_2 refer to independent dropout masks applied during forward propagation. This strategy not only supplies pseudo-labels for data augmentation but also introduces contrastive signals between positive samples, thus enhancing the robustness of embeddings to different noise distributions. Furthermore, the randomness of dropout masks enables the model to generate diverse semantic representations for the same text sample, simulating data variety and compensating for the lack of actual data.

4.2 Dual Negative Sampling Strategy: Soft and Hard Negatives

To further improve the model's capacity to distinguish between semantically similar content, we incorporate a layered negative sampling strategy. This strategy includes both soft and hard negatives to capture subtle semantic differences and enhance the detection of deceptive content.

Specifically, we generate soft negatives by sampling other real text x_j from the same batch (where $j \neq i$):

$$s_j = f_\theta(x_j, \text{dropout}_1) \tag{3}$$

Soft negatives allow the model to retain a distributional understanding of the overall semantic space when learning the relationships between real texts. This type of soft constraint permits the model to discern differences between various text samples through spatial relationships in the embedding space, thereby establishing a broad and comprehensive semantic embedding distribution. This ensures that the model can maintain distinctions among authentic content within the semantic space.

However, soft negatives alone may not sufficiently challenge the model, particularly for distinguishing highly similar deceptive texts. Therefore, to strengthen the model's ability to make fine-grained distinctions between real and fake content, we use synthetic texts x_i', generated by large language models (LLMs), as hard negatives:

$$n_i = f_\theta(x_i', \text{dropout}_1) \tag{4}$$

Hard negatives are crafted to directly mimic the semantic structure and stylistic features of real texts, aiming to deceive the model. By doing so, they provide a challenging learning signal that forces the model to rely on subtle cues to differentiate between positive and hard negative samples. This enhancement ultimately sharpens the model's sensitivity to subtle cues, boosting its effectiveness in identifying misleading information.

4.3 Contrastive Learning Training Objective

To maximize the similarity between queries and positive samples in the semantic embedding space while increasing the distance from both soft and hard negatives, we employ a layered contrastive learning loss function. The objective is to ensure optimal separation between positive and negative samples. The loss function L is defined as follows:

$$L = -\log \frac{e^{\text{sim}(q_i,p_i)/\tau}}{e^{\text{sim}(q_i,p_i)/\tau} + \beta + \sum_{j \neq i} \phi} \qquad (5)$$

$$\phi = e^{\text{sim}(q_i,s_j)/\tau}, \quad \beta = e^{\text{sim}(q_i,n_i)/\tau} \qquad (6)$$

where sim represents the cosine similarity function, which quantifies the similarity between two vectors in the embedding space. The temperature parameter τ moderates the smoothing effect on the similarity, controlling the weighting and compactness within the embedding space. The goal is to maximize the similarity between the query embedding q_i and the positive sample p_i, while expanding the distance to the soft negative s_j and hard negative n_i.

Specifically, by incorporating soft and hard negatives within the cosine similarity framework, the model dynamically adjusts the positions of real and synthetic samples, creating a distinct semantic boundary. This ensures that, in high-dimensional semantic space, the model maintains sensitivity to subtle misleading signals, thereby effectively enhancing its robustness and accuracy in retrieval tasks.

5 Experiments

The primary objective of this study is to comprehensively evaluate model performance across various domains, tasks, and scenarios involving deceptive information. Given the widespread application of large language models in information retrieval and verification, we aim to explore the relationship between model scale and task complexity, focusing on robustness and generalization when confronted with highly similar deceptive texts. Our experiments specifically compare two model scales (Stella_1.5B and Stella_400M) to assess their capabilities in scientific fact verification, cross-domain reasoning, medical information retrieval, and climate change fact-checking. Using multiple evaluation metrics, we quantitatively analyze the models' performance across genuine and deceptive datasets.

5.1 Datasets

Dataset	Domain	Task	Relevancy	Corpus
SciFact	Scientific	Fact Checking	Binary	5183
HotpotQA	Wikipedia	QA	Binary	20000
NFCorpus	Bio-Medical	IR	3-level	3633
Climatic-FEVER	Wikipedia	Fact Checking	Binary	10000

Table 1: Datasets Used for Model Train and Evaluation

As shown in Table 1, we selected four datasets covering a diverse range of tasks and domains. Following the methodology outlined in Chapter 3, we employed a guided generation framework with OpenAI's "gpt-3.5-turbo" to create two deceptive text expansions, Fake1 and Fake2, for each dataset's corpus, ensuring they matched the original corpus in size. The generation strategies included semantic perturbation, logical inversion, and data fabrication, thereby presenting realistic challenges to the model. Fake1, used as a hard negative sample set during training, enhances the model's sensitivity to subtle contextual variations and its ability to detect deceptive content. Fake2 serves as a validation expansion to assess the model's robustness and generalization when confronted with previously unseen deceptive texts, offering an additional evaluation benchmark. Given the substantial size of the HotpotQA and Climate-FEVER datasets, we performed random sampling to reduce the corpus size and training cost without compromising the integrity of the validation set.

5.2 Experimental Setup

5.2.1 Models and Training Configuration

Our experiments utilize two scales of Stella models: Stella_1.5B and Stella_400M. The 1.5B model, with a larger parameter count, offers enhanced representation capabilities, while the 400M model is more computationally efficient, suitable for resource-constrained environments. Both models are Transformer-based but differ in parameter scales to evaluate the effect of model size on complex tasks and deception detection.

For experimental rigor, we employ unsupervised contrastive learning across all models, using random dropout noise for data augmentation and a

Model	Dataset	NDCG@3	MAP@3	NDCG@5	MAP@5	NDCG@10	MAP@10
Original (Stella-1.5B)	Scifact	72.2	69.1	75.3	71.2	77.6	72.4
	Scifact+Fake1	63.2	59.8	66.0	61.6	69.8	63.5
	Scifact+Fake2	63.4	59.5	65.8	61.2	69.4	63.0
RetrieverGuard (Stella-1.5B)	Scifact	**73.0**	**70.0**	**75.7**	**71.8**	**78.2**	**73.1**
	Scifact+Fake1	**65.7**	**62.4**	**68.4**	**64.1**	**71.8**	**65.8**
	Scifact+Fake2	**65.0**	**61.5**	**67.8**	**63.4**	**71.2**	**65.0**
Original (Stella-1.5B)	HotpotQA	86.9	82.9	88.9	84.7	90.3	85.6
	HotpotQA+Fake1	72.1	63.1	77.2	67.7	79.4	69.1
	HotpotQA+Fake2	72.3	63.4	77.3	67.9	79.5	69.3
RetrieverGuard (Stella-1.5B)	HotpotQA	86.4	82.3	88.4	84.1	89.8	85.0
	HotpotQA+Fake1	**73.6**	**64.9**	**78.1**	**69.0**	**80.2**	**70.4**
	HotpotQA+Fake2	**73.7**	**65.1**	**78.3**	**69.3**	**80.3**	**70.6**
Original (Stella-1.5B)	NFCorpus	44.5	10.1	41.8	11.9	39.0	14.4
	NFCorpus+Fake1	34.0	8.0	32.4	9.1	29.7	10.6
	NFCorpus+Fake2	33.6	7.8	31.9	9.0	29.9	10.5
RetrieverGuard (Stella-1.5B)	NFCorpus	44.2	9.8	41.7	11.6	38.9	14.1
	NFCorpus+Fake1	**36.3**	**8.3**	**34.0**	**9.4**	**31.1**	**10.9**
	NFCorpus+Fake2	**35.5**	**8.3**	**33.5**	**9.4**	**31.3**	**11.1**
Original (Stella-1.5B)	Climate-FEVER	34.7	25.7	38.4	29.5	44.3	33.0
	Climate-FEVER+Fake1	27.0	19.4	30.5	22.5	36.6	25.7
	Climate-FEVER+Fake2	25.2	17.8	28.9	20.9	34.8	24.0
RetrieverGuard (Stella-1.5B)	Climate-FEVER	**35.8**	**26.6**	**39.4**	**30.5**	**45.3**	**34.0**
	Climate-FEVER+Fake1	**31.4**	**23.0**	**34.4**	**26.1**	**40.3**	**29.3**
	Climate-FEVER+Fake2	**30.2**	**22.2**	**33.8**	**25.5**	**39.5**	**28.6**

Table 2: Results with stella-1.5B show that models trained using RetrieverGuard effectively handle misinformation.

dual-negative sample mechanism to enhance sensitivity and differentiation capability against deceptive information. Training does not include explicit query pairs; instead, self-supervised methods construct positive and negative pairs, enabling the model to learn fine-grained text differences without query reliance.

Hyperparameters are kept consistent across models to ensure comparability. All models utilize the Adam optimizer, with an initial learning rate of 1×10^{-5} and batch size of 8. We apply a learning rate decay and hard-negative sample weighting decay strategy, gradually reducing the learning rate to 1×10^{-8} and increasing the batch size to 64. Each model undergoes multiple runs on every task to mitigate potential biases from data randomness.

5.2.2 Evaluation Metrics

To comprehensively evaluate model performance, we employ the following metrics: NDCG@k and MAP@k. These metrics are capable of assessing both the ranking performance and retrieval accuracy of the model.

5.3 Results Analysis

The experimental results clearly indicate that Stella_1.5B (Table 2) and Stella_400M (Table 6 in Appendix E) exhibit significant performance differences across various datasets, illustrating the influence of model size, dataset diversity, and the presence of deceptive texts, particularly in multi-hop reasoning and cross-domain tasks. The optimal training hyperparameter configurations for each dataset are presented in Table 4 in Appendix B.

Stella_1.5B consistently outperforms the 400M model across original datasets, with this trend being particularly pronounced on complex datasets such as HotpotQA and Climate-FEVER. On HotpotQA, the 1.5B model achieves an NDCG@5 of 88.9 compared to 86.4 for the 400M model, indicating that the larger model excels at capturing intricate semantic relationships and logical reasoning. When deceptive texts are introduced, the NDCG@5 for the 1.5B model on HotpotQA+Fake1 and HotpotQA+Fake2 drops to 77.2 and 77.3, respectively, compared to a decrease to 74.7 and 74.8 for the 400M model. These reductions indicate that, although both models are affected by deceptive texts, the 1.5B model exhibits a higher resilience against such interference in multi-hop reasoning tasks.

However, the impact of deceptive texts extends beyond mere performance degradation; it fundamentally threatens the integrity of information retrieval systems. This is especially evident in the results on the Scifact dataset. The NDCG@3 for the 1.5B model on the original data is 72.2, which sharply decreases to 63.4 upon the introduction of Fake2, while the 400M model's NDCG@3 decreases from 71.9 to 62.0. A similar trend is observed in the NFCorpus dataset, where the

NDCG@3 for two models drops by approximately 11.0 and 12.0. These substantial declines underscore that deceptive texts, particularly those related to domain-specific information, significantly impair the models' decision-making capabilities, leading to severe misjudgments when dealing with complex, domain-specific issues. This issue is particularly concerning in sensitive fields such as medicine and science, where misleading information can have far-reaching adverse effects. Such degradation in model accuracy could result in serious misguidance in practical applications, potentially providing users with inaccurate information and ultimately affecting the quality of decision-making.

To mitigate the harmful effects of deceptive texts, the proposed self-supervised contrastive learning method demonstrates remarkable efficacy. Particularly for the Stella_1.5B model, training with deceptive texts substantially improves sensitivity and accuracy in detecting deceptive information. For instance, on the Climate-FEVER dataset, the untrained 1.5B model's NDCG@3 drops from 34.7 to 27.0 after introducing Fake1, whereas the RetrieverGuard 1.5B achieves an NDCG@3 of 31.4 on the same Fake1 data, reducing the decline by 4.4. Similar improvements are observed across other metrics, such as NDCG@5 and MAP@5, with NDCG@5 increasing from 30.5 to 34.4, and MAP@5 improving from 22.5 to 26.1. Additionally, on the unseen validation interference dataset Fake2, the trained model demonstrates exceptional robustness and generalizability; for example, the 1.5B model's NDCG@3 improves from 25.2 to 30.2 (approximately 5.0) on Climate-FEVER. These enhancements illustrate that self-supervised learning with soft and hard negative sampling not only boosts the model's performance on the original data (by approximately 1.0), but also significantly strengthens the model's resilience against interference from deceptive datasets.

In summary, the experimental results confirm that by increasing model size and incorporating a self-supervised contrastive learning approach, the Stella_1.5B model better withstands the interference of high-similarity deceptive texts. The specific data reflects substantial improvements across various datasets and tasks, validating the efficacy of self-supervised learning for detecting deceptive information. Furthermore, the challenges posed by deceptive texts underscore the necessity for developing more robust information retrieval models to ensure users have access to more reliable and authentic information.

5.4 Sensitivity to Misinformation Ratios

To simulate the randomness found in real-world databases—such as cases where some corpora have corresponding deceptive counterparts while others do not, or instances where certain corpora contain more than one interfering corpus. We randomly introduced AI-generated deceptive corpora into the original SciFact corpus. The proportion of deceptive to authentic data ranged from 20% to 200%. Figure 5 illustrates the performance of both the original 1.5B model and the RetrieverGuard model in identifying deceptive texts under varying misinformation ratios. Although the performance of both models declines as the noise ratio increases, the trained model demonstrates a clear advantage in terms of both the rate of decline and overall resilience to interference. These results further validate the robustness and broad applicability of the proposed RetrieverGuard method.

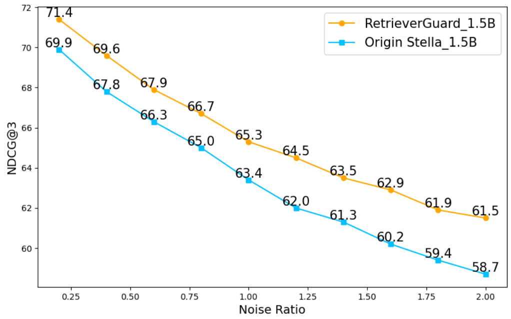


Figure 5: NDCG@3 for Origin stella_1.5B and RetrieverGuard stella_1.5B of different Noise Ratios

6 Conclusion

Our study provides an in-depth examination of the challenges posed by misinformation generated by large language models (LLMs) to information retrieval systems, offering a highly practical solution. By constructing a diverse dataset spanning multiple domains and containing both authentic human-authored texts and highly deceptive AI-generated content, we have established a foundational resource for future research in this area. Furthermore, we developed and validated the RetrieverGuard method, which leverages a self-supervised contrastive learning strategy that employs both soft and hard negatives to enhance model robustness and misinformation detection capabilities. Experimental results demonstrate that RetrieverGuard achieves outstanding performance in improving retrieval accuracy and countering highly deceptive content, thus providing crucial technical support for maintaining the reliability and authenticity of IR systems in the era of LLMs.

Limitations

While the self-supervised contrastive learning method has demonstrated significant effectiveness in deception detection, it faces several limitations. Firstly, the training process is time-intensive, and the method's performance may be suboptimal in low-resource or few-shot settings. Furthermore, its reliance on soft and hard negative sampling could lead to overfitting on specific types of deceptive texts when scaling to larger and more diverse datasets, which may hinder the model's generalization capabilities. Additionally, while large-parameter models exhibit superior performance in deception detection, they are accompanied by substantial computational costs.

References

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 716–722. Springer.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? *Preprint*, arXiv:2309.13788.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings. *Preprint*, arXiv:2104.08821.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *Preprint*, arXiv:2301.07597.

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

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. *Preprint*, arXiv:1911.00650.

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

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Ian Kelk, Benjamin Basseri, Wee Yi Lee, Richard Qiu, and Chris Tanner. 2022. Automatic fake news detection: Are current models "fact-checking" or "gut-checking"? *Preprint*, arXiv:2204.07229.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024a. Making text embedders few-shot learners. *Preprint*, arXiv:2409.15700.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024b. Inference-time intervention: Eliciting truthful answers from a language model. *Preprint*, arXiv:2306.03341.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *Preprint*, arXiv:2310.08319.

Mengjia Niu, Hao Li, Jie Shi, Hamed Haddadi, and Fan Mo. 2024. Mitigating hallucinations in large language models via self-refinement-enhanced knowledge retrieval. *Preprint*, arXiv:2405.06545.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16(8).

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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

Jinyan Su, Claire Cardie, and Preslav Nakov. 2024. Adapting fake news detection to the era of large language models. *Preprint*, arXiv:2311.04917.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *Preprint*, arXiv:2303.07205.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Preprint*, arXiv:2104.08663.

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

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *Preprint*, arXiv:2004.14974.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. *Preprint*, arXiv:2401.00368.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *Preprint*, arXiv:2112.04359.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Tian Yu, Shaolei Zhang, and Yang Feng. 2024. Truth-aware context selection: Mitigating hallucinations of large language models being misled by untruthful contexts. *Preprint*, arXiv:2403.07556.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. *Preprint*, arXiv:2209.10063.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending against neural fake news. *Preprint*, arXiv:1905.12616.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Neural deepfake detection with factual structure of text. *Preprint*, arXiv:2010.07475.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2024. Large language models for information retrieval: A survey. *Preprint*, arXiv:2308.07107.

A Appendix A: Discuss

Scale Effect: The experimental results demonstrate that the 1.5B model consistently outperforms the 400M model across nearly all tasks, particularly in deception detection, exhibiting greater robustness. This indicates a positive correlation between model size and its ability to handle highly similar deceptive information. Larger models have a heightened capacity to capture subtle semantic differences and manage complex semantic structures, effectively maintaining high performance. This characteristic is especially notable when facing deceptive texts with intricate logic and semantics.

Domain-Specific Challenges: Cross-domain tasks, such as HotpotQA and Climatic-FEVER, impose higher demands on the model's semantic understanding and reasoning capabilities. Although the 1.5B model performs well on genuine data, there remains substantial room for improvement on deceptive data. This suggests that current models require further optimization for handling multi-hop reasoning and domain-specific tasks. The model's limitations become particularly evident when processing highly similar deceptive information. In these scenarios, existing methods fail to adequately address the deception risks inherent in cross-domain and complex tasks, highlighting a critical need for further advancements.

Interference Effects of Deceptive Texts: Regardless of model size, deceptive text expansions exert a significant negative impact on model performance. The experimental results reveal that even the larger 1.5B model experiences notable declines in NDCG, MAP, and other metrics when confronted with highly similar deceptive texts. This reflects existing models' vulnerabilities in managing deceptive information and the inadequacies in mitigating the impact of deceptive texts on model accuracy. Such effects are particularly pronounced in specific domains (e.g., climatology and medicine), where they may not only degrade model performance but also risk severely misleading users. In the era of large language models, the challenges posed by deceptive information cannot be overlooked, as they threaten the integrity and reliability of information retrieval systems. Thus, strengthening model resilience against deceptive information and enhancing performance across multi-task scenarios have become critical challenges that demand urgent attention.

Broader Potential of Self-Supervised Contrastive Learning The self-supervised contrastive learning method proposed in this study has proven highly effective in enhancing model resilience to deceptive information. Beyond the specific tasks addressed in this research, this approach holds considerable potential for broader application across various domains involving misinformation detection. Areas such as social media content moderation, fake news detection, and online fraud prevention could all benefit from this technique.

B Appendix B: Final Model Hyperparameters

Here, we set the Dropout rate to 0.1 for all datasets, and the specific experimental parameters are shown in Table 5 in Appendix C.

Dataset	Batch Size	lr	Dropout	Epoch
SciFact	8	$1e^{-8}$	0.1	3
HotpotQA	8	$1e^{-7}$	0.1	1
NFCorpus	16	$1e^{-8}$	0.1	4
Climatic-FEVER	8	$1e^{-7}$	0.1	1

Table 3: Train Hyperparameters for each Dataset on Stella_1.5B

Dataset	Batch Size	lr	Dropout	Epoch
SciFact	8	$1e^{-8}$	0.1	2
HotpotQA	8	$1e^{-8}$	0.1	3
NFCorpus	8	$1e^{-8}$	0.1	3
Climatic-FEVER	64	$1e^{-8}$	0.1	2

Table 4: Train Hyperparameters for each Dataset on Stella_400M

C Appendix C: Dropout statistic

We experimented with different dropout rates while keeping the other training parameters constant and found that a Dropout rate of 0.1 yielded the best results for both the NDCG@3 and MAP@3 metrics.

Dropout Ratio	NDCG@3	MAP@3
0.01	64.9	61.3
0.03	64.8	61.2
0.06	64.8	61.3
0.1	65.3	61.8
0.15	64.8	61.3
0.2	65.1	61.6

Table 5: Dropout statistic of SciFact on Stella_1.5B

Model	Dataset	NDCG@3	MAP@3	NDCG@5	MAP@5	NDCG@10	MAP@10
Original (Stella-400M)	Scifact	71.9	69.2	74.2	70.9	77.0	72.3
	Scifact+Fake1	62.6	58.9	65.2	60.6	68.0	61.9
	Scifact+Fake2	62.0	57.9	64.3	59.4	67.3	60.9
RetrieverGuard (Stella-400M)	Scifact	**72.3**	**69.4**	74.1	70.8	77.0	72.3
	Scifact+Fake1	**65.6**	**62.1**	**67.5**	**63.5**	**70.4**	**64.9**
	Scifact+Fake2	**63.8**	**60.1**	**66.0**	**61.6**	**69.3**	**63.2**
Original (Stella-400M)	HotpotQA	84.1	79.4	86.4	81.4	88.0	82.5
	HotpotQA+Fake1	69.6	60.5	74.7	65.1	77.1	66.7
	HotpotQA+Fake2	69.7	60.7	74.8	65.2	77.2	66.8
RetrieverGuard (Stella-400M)	HotpotQA	**84.4**	**79.8**	**86.6**	**81.7**	**88.3**	**82.9**
	HotpotQA+Fake1	**71.4**	**62.7**	**76.2**	**66.9**	**78.5**	**68.4**
	HotpotQA+Fake2	**71.6**	**62.8**	**76.3**	**67.0**	**78.6**	**68.5**
Original (Stella-400M)	NFCorpus	44.7	10.3	42.7	12.3	39.9	14.8
	NFCorpus+Fake1	32.7	7.4	31.3	8.8	29.9	10.5
	NFCorpus+Fake2	32.6	7.7	31.8	9.1	30.4	10.8
RetrieverGuard (Stella-400M)	NFCorpus	**44.9**	10.3	42.3	12.2	39.9	14.7
	NFCorpus+Fake1	**35.1**	**7.8**	**33.3**	**9.5**	**31.5**	**11.1**
	NFCorpus+Fake2	**35.5**	**8.2**	**33.6**	**9.5**	**31.8**	**11.2**
Original (Stella-400M)	Climate-FEVER	34.4	25.3	37.8	29.0	43.9	32.4
	Climate-FEVER+Fake1	25.9	18.5	29.2	21.4	34.6	24.2
	Climate-FEVER+Fake2	24.2	17.2	28.0	20.2	33.4	23.0
RetrieverGuard (Stella-400M)	Climate-FEVER	**34.6**	**25.6**	**38.0**	**29.3**	**44.2**	**32.8**
	Climate-FEVER+Fake1	**28.4**	**20.7**	**31.3**	**23.4**	**36.7**	**26.3**
	Climate-FEVER+Fake2	**27.4**	**19.9**	**30.3**	**22.6**	**35.7**	**25.5**

Table 6: Results of Stella_400M.

D Appendix D: Resource

All our experiments were conducted using a single A800-80G GPU, with varying time consumption across different datasets. For example, on the Climate-FEVER dataset (10k samples), the parameter search took approximately 32 hours. Under optimal training parameters, RetrieverGuard's training time was around 1 hour.

E Appendix E: Result of Stella_400M

In our experiments, all reported results are averaged over multiple runs to ensure the accuracy and stability of the overall findings. Specifically, we conducted three experiments and statistical analyses on the Stella_1.5B and Stella_400M models across four datasets. The table presents the average values, with most data fluctuations within a range of less than 0.5% (except for NDCG@1, which is around 1%). For the Noise Ratio experiment in Section 5.4, we performed an additional five runs and reported the averaged results.