Streamlining Conformal Information Retrieval via Score Refinement

Yotam Intrator* yotami@google.com
Regev Cohen* regevcohen@google.com
Ori Kelner* orikelner@google.com
Roman Goldenberg gnamor@gmail.com
Ehud Rivlin ehudr@cs.technion.ac.il
Daniel Freedman danielfreedman@gmail.com

Verily AI (Google Life Sciences), Israel.

Abstract

Information retrieval (IR) methods, like retrieval augmented generation, are fundamental to modern applications but often lack statistical guarantees. Conformal prediction addresses this by retrieving sets guaranteed to include relevant information, yet existing approaches produce large-sized sets, incurring high computational costs and slow response times. In this work, we introduce a score refinement method that applies a simple monotone transformation to retrieval scores, leading to significantly smaller conformal sets while maintaining their statistical guarantees. Experiments on various BEIR benchmarks validate the effectiveness of our approach in producing compact sets containing relevant information.

1 Introduction

Information retrieval (IR) methods lie at the heart of numerous modern applications, ranging from search engines and recommendation systems to question-answering platforms and decision support tools. These methods facilitate the identification and extraction of relevant information from vast collections of data, enabling users to access the knowledge they seek efficiently and effectively. A popular example of IR is Retrieval Augmented Generation (RAG), a technique for reducing hallucinations in large language models (LLMs) by grounding their responses on factual information retrieved from external sources.

While IR methods have been widely adopted, they traditionally lack statistical guarantees on the relevance of retrieved information. This limitation can lead to uncertainty regarding the reliability and correctness of the retrieved information. Conformal prediction (Angelopoulos and Bates, 2021; Angelopoulos et al., 2021) is an uncertainty quantification framework that can be used with any underlying model to construct sets that are statistically guaranteed to contain the ground truth with a user-specified probability. Conformal prediction has expanded far beyond its initial classification focus (Vovk et al., 2005; Angelopoulos and Bates, 2021; Ringel et al., 2024), now encompassing diverse applications like regression, image-to-image translation (Angelopoulos et al., 2022b; Kutiel et al., 2023), and foundation models (Gui et al., 2024), advancing to enable control of any monotone risk function (Angelopoulos et al., 2022a). In the context of IR, recent methods (Xu et al., 2024; Li et al., 2023; Angelopoulos et al., 2023) have incorporated conformal prediction into ranked retrieval systems to ensure the reliability and quality of retrieved items. However, existing conformal methods often produce excessively large retrieved sets, implying high computational costs and slower response times.

In this work, we address this limitation by introducing a novel score refinement method that employs a simple yet effective monotone transformation, inspired by ranking measures, to adjust the scores of any given information retrieval system. By applying standard conformal prediction methods to these refined scores, we deliver significantly smaller retrieved sets while preserving their statistical guarantees, striking a crucial balance between efficiency and accuracy. An illustration of the proposed pipeline is shown in Figure 1. We validate the effectiveness of our method through experiments on three of BEIR (Thakur et al., 2021) benchmark datasets, demonstrating its ability to outperform competing approaches in producing compact sets that contain the relevant information.

2 Background

To lay the groundwork for our work, we present a simplified description of the operation of information retrieval systems and how conformal inference

*Equal Contribution.

Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER), pages 186–191

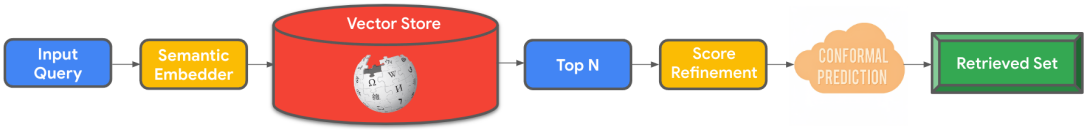

Figure 1: Retrieval Pipeline. The query is first embedded using a semantic embedder, and then the top N candidates are retrieved from a vector store. Crucially, their corresponding scores then undergo a refinement transformation before being passed through a conformal prediction method that outputs an adaptive set of documents.

can be seamlessly integrated within this context.

2.1 Information Retrieval: Overview

Consider a large information database $\mathcal{D} = \{d_1, d_2, ..., d_N\}$. At inference time, an IR model $R : \mathcal{Q} \to \mathcal{D}$ accepts a query $q \in \mathcal{Q}$ as input and returns a subset of candidates $\mathcal{S} \subset \mathcal{D}$. To do this, the IR model computes a semantic embedding $e_q = E(q)$ for the query and compares it to precomputed embeddings $e_i = E(d_i)$ for each item in the database using a similarity metric:

$$s_i = \texttt{sim}(e_q, e_i), \tag{1}$$

where E is the chosen representation model (e.g., a neural network encoder) and `sim` is a similarity metric, such as cosine similarity. Subsequently, the items are typically ranked based on their similarity scores, and the top ranked items are retrieved, forming the following set

$$\mathcal{S}_K \triangleq \left\{ d_i \in \mathcal{D} : s_i \geq s_{(K)} \right\} \tag{2}$$

where $s_{(K)}$ denotes the Kth largest similarity score, for a predefined $K > 0$ constant across all queries.

The approach above suffers from two key limitations. First, using a fixed K can be problematic: it might be too restrictive for some queries, leading to the omission of relevant information, while for others, it might be too permissive, resulting in the retrieval of numerous redundant or irrelevant items. The latter scenario significantly impacts efficiency and prolongs response times. Second, this approach lacks guarantees that truly relevant information, such as a specific item d^* within the database $\mathcal{D}$, will be included in the retrieved set S.

2.2 Conformal Prediction for Retrieval

Conformal prediction can be seamlessly integrated into IR systems by constructing calibrated prediction sets designed to include, on average, the desired information with a user-specified high probability. Formally, given a query q and its corresponding similarity scores s_i, we construct a prediction set parameterized by $\tau > 0$ as follows:

$$\mathcal{C}_\tau(q) \triangleq \{d_i \in \mathcal{D} : c_i \leq \tau\}, \tag{3}$$

where $c_i \triangleq -s_i$ represents a *non-conformity* score. To appropriately set the value of τ, we utilize a held-out calibration dataset $\mathcal{D}_C$ consisting of n samples $(q_i, d_i) \in \mathcal{Q} \times \mathcal{D}$ drawn exchangeably from an underlying distribution $\mathcal{P}$. Here, q_i represents a query whose most relevant information is assumed to be a single item d_i from the database, for simplicity. Given a user-chosen error rate $\alpha \in [0, 1]$, we set τ as the $\frac{(n+1)(1-\alpha)}{n}$-th quantile of the calibration non-conformity scores. This ensures that for a new exchangeable test sample (q_{n+1}, d_{n+1}), we have the following marginal coverage guarantee:

$$\mathbb{P}\big(d_{n+1} \in \mathcal{C}_\tau(q_{n+1})\big) \geq 1 - \alpha \tag{4}$$

for any distribution P. The probability above is marginal (averaged) over all $n + 1$ calibration and test samples. This ensures that the IR model retrieves sets of adaptive size, guaranteed to contain the relevant information at least α-fraction of the time, thereby overcoming the limitation above.

While the conformal sets above use a calibrated threshold, other parameterizations are possible, such as setting the calibration parameter to the set size K, as in (2). Furthermore, it is important to note that the description above merely presents conformal prediction in its simplest, most common form. However, there have been significant advancements in the field in recent years, leading to the development of more involved and efficient conformal methods (Romano et al., 2020; Angelopoulos et al., 2020) and to extensions that provide guarantees beyond marginal coverage (Angelopoulos et al., 2022a; Fisch et al., 2020; Li et al., 2023).

3 Method

Integrating conformal prediction to IR systems enhances their reliability by providing statistical guarantees. However, CP methods prioritize trustworthiness and are not optimized for efficiency, thus they often produce excessively large retrieval sets.

Following the above, our goal is to improve the predictive efficiency of CP methods by reducing the average size of the retrieved sets $\mathbb{E}_q\big[|\mathcal{C}_\tau(q)|\big]$, while maintaining their coverage guarantees. In contrast to approaches that focus on improving the IR model or developing more efficient conformal methods, we propose an alternative approach that introduces an intermediate step of score refinement. Specifically, given a query q and its scores $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$, we adjust them prior to employing conformal prediction $T(\mathcal{S}) = \{t_1, t_2, \ldots, t_N\}$.

In designing the transformation T, we identify that scores from different queries can vary significantly in scale. This can cause the calibration threshold τ to be dominated by queries with small scores, leading to excessively large prediction sets. To mitigate this issue, we first normalize the retrieval scores by dividing them by their maximum, ensuring that scores across all queries are comparable in scale. We remark that the maximum score $s_{\max}$ can be interpreted as the IR model's confidence. When this value is small, it suggests a lack of relevant information for the given query, suggesting that ideally no items should be retrieved. Thus, normalization in such scenarios may be counterproductive, resulting in irrelevant items. However, we assume the corpus is sufficiently extensive to contain at least one relevant item for any query, an assumption particularly valid for the calibration.

Next, assume without loss of generality that the scores are sorted in decreasing order: $\mathcal{S} = \{s_{(1)}, s_{(2)}, \ldots, s_{(N)}\}$, where $s_{(r)}$ is the rth largest score and $r \geq 1$ represents its rank. Inspired by ranking measures (Yining et al., 2013), we define our transformation as follows

$$T(s_{(r)}, r) \triangleq \frac{s_{(r)}}{s_{\max}} W(r) \tag{5}$$

where $W(r) \in [0, 1]$ is a discount function that penalizes scores based on their rank. We specifically employ the inverse logarithm decay $W(r) = \frac{1}{\log(1+r)}$, which offers a balance between emphasis on top items and exploration of lower-ranked items. To offer additional flexibility, we introduce a hyperparameter $\lambda \in [0, 1]$:

$$T(s_{(r)}, r) \triangleq \frac{s_{(r)}}{s_{\max}} \frac{1}{\log(1 + r^\lambda)}. \tag{6}$$

We tune λ by performing a search over a sequence of values to minimize the set size on a validation set. Note the transformation is monotone, preserving the IR model's induced order and maintaining its core functionality. Furthermore, it is simple to implement, computationally efficient, and easily integrated into existing systems. As demonstrated in the following section, the proposed transformation is highly effective in reducing the size of the conformal retrieved sets.

4 Experiments

4.1 Setup

Datasets For our evaluation, we utilized BEIR (Thakur et al., 2021), a large collection of information retrieval benchmarks. Specifically, we focus on the following datasets: FEVER (Thorne et al., 2018), SCIFACT (Wadden et al., 2020), and FIQA (Maia et al., 2018). Data statistics are presented at Table 1. It is important to note that each query within these datasets may have multiple relevant documents within the corpus. For this study, we adopted a pragmatic approach, considering the document with the highest score among the relevant documents as the ground truth. This ensures that a successful retrieval implies at least one relevant document is present in the inference set.

To simulate real-world production environments, we employ a vector store, specifically FAISS-GPU (Johnson et al., 2019) for its efficiency and performance in handling large-scale databases. We retrieve the top $2,000$ documents for each query and apply our refinement process exclusively to these initially retrieved documents.

Dataset	#Corpus	#Calibration	#Test
FEVER	5,416,568	6,666	6,666
SCIFACT	5,183	150*	150*
FIQA	57,638	500	648

Table 1: Data Summary. #Corpus indicates the number of documents, while #Calibration and #Test indicate the number of queries. As SCIFACT lacks a calibration set, we randomly split its test set into calibration and test subsets.

Embedders Initial semantic scores were derived using deep sentence embedders, which encode textual input into a fixed-dimensional latent space where semantic similarity is represented by vector proximity. We employ two models: BGE-large-1.5 (Xiao et al., 2023) (326M parameters) and E5-Mistral-7b model (Wang et al., 2023) (7B parameters). BGE-large-1.5 is a smaller model with a latent representation dimension of 1024, whereas E5-Mistral, a finetuned encoder version of the mistral-7b model, has a latent representation dimension

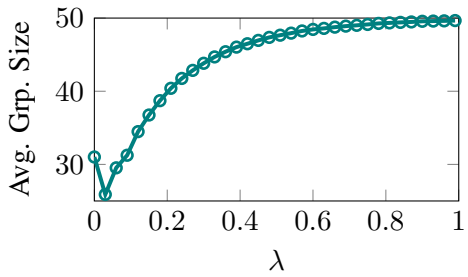

Figure 2: Impact of λ value on average group size using BGE-large-1.5 on SCIFACT with $\alpha = 0.05$.

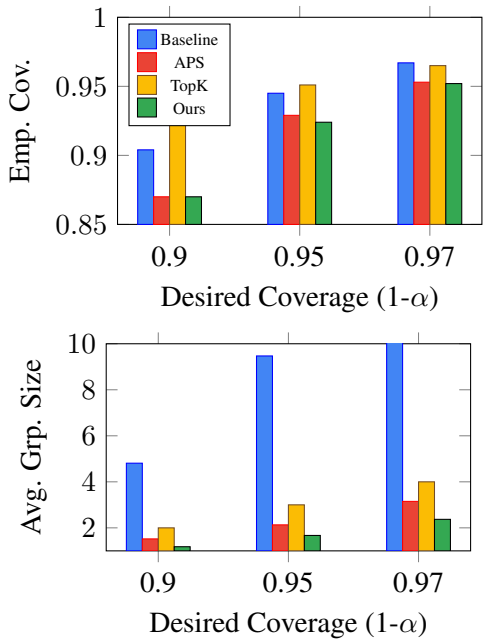

Figure 3: Performance comparison using BGE-large-1.5 on FEVER dataset across various values of α.

of 4096. The semantic score between a query q and a candidate document d is the cosine similarity between their respective latent representations.

Competitors For our method, we employ the Vanilla CP method (Vovk et al., 2005), applying it to the refined retrieval scores. We compared our approach to three established approaches: *Baseline*, which applies Vanilla CP directly to the retrieval scores without modification; *TopK*, which utilizes Vanilla CP but calibrates to a fixed set size K for all queries; *APS* (Romano et al., 2020) and *RAPS* (Angelopoulos et al., 2020), which introduce novel conformity scores.

4.2 Results

We first conduct experiments on the smaller SCIFACT dataset to optimize the hyperparameter λ. The results, shown in Figure 2, reveal a favorable value for λ, prompting us to set $\alpha = 0.05$.

Dataset	α	Method	Emp. Cov.	Avg. Grp. Size
FIQA	0.1	Baseline	0.89	417.77
		APS	0.89	119.76
		TopK	0.87	90.0
		Ours	0.86	**56.72**
	0.05	Baseline	0.94	846.0
		APS	0.94	477.27
		TopK	0.92	259.0
		Ours	0.92	**190.5**
	0.03	Baseline	0.96	1206.93
		APS	0.98	1393.96
		TopK	0.94	480.0
		Ours	0.95	**347.16**
SCIFACT	0.1	Baseline	0.91	231.17
		APS	0.91	30.82
		TopK	0.91	31.0
		Ours	0.85	**14.07**
	0.05	Baseline	0.97	760.75
		APS	0.92	91.23
		TopK	0.92	91.0
		Ours	0.89	**29.59**
	0.03	Baseline	0.98	1211.11
		APS	0.95	276.15
		TopK	0.95	279.0
		Ours	0.97	**160.66**

Table 2: Performance comparison using BGE-large-1.5 on FIQA and SCIFACT across various values of α.

Next, we conduct experiments on the large-scale FEVER dataset. As illustrated in Figure 3, our score refinement method consistently outperforms other approaches by producing significantly smaller retrieved sets in experiments with BGE-large-1.5 across various values of α, while maintaining comparable, albeit slightly lower, empirical coverage. Results for the other datasets are summarized in Table 2, consistent with our previous findings. We note that RAPS produced comparable results to APS, so we omit them for brevity. Additional results using E5-Mistral, which exhibit similar trends, are presented in Table 3 of the appendix, along with an ablation study comparing other simple transformations.

5 Conclusion

We addressed the challenge of large prediction sets in conformal prediction for IR by introducing a novel score refinement method. Our experiments on the BEIR benchmark demonstrated its effectiveness in generating compact, statistically reliable prediction sets, enabling the deployment of conformal prediction in real-world IR systems without sacrificing performance.

6 Limitations

The conclusions of this study could be further strengthened by evaluating the method on a wider range of datasets and employing diverse embedding models. Currently, our method does not han-

dle cases where no relevant information exists in the database, potentially limiting its applicability. Additionally, while we introduced a simple transformation, exploring more involved or even parameterized functions, e.g. neural networks, could further enhance efficiency and statistical guarantees.

References

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Anastasios N Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. 2021. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2022a. Conformal risk control. *arXiv preprint arXiv:2208.02814*.

Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. 2022b. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pages 717–730. PMLR.

Anastasios N Angelopoulos, Karl Krauth, Stephen Bates, Yixin Wang, and Michael I Jordan. 2023. Recommendation systems with distribution-free reliability guarantees. In *Conformal and Probabilistic Prediction with Applications*, pages 175–193. PMLR.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. 2020. Efficient conformal prediction via cascaded inference with expanded admission. *arXiv preprint arXiv:2007.03114*.

Yu Gui, Ying Jin, and Zhimei Ren. 2024. Conformal alignment: Knowing when to trust foundation models with guarantees. *arXiv preprint arXiv:2405.10301*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Gilad Kutiel, Regev Cohen, Michael Elad, Daniel Freedman, and Ehud Rivlin. 2023. Conformal prediction masks: Visualizing uncertainty in medical imaging. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pages 163–176. Springer.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. 2023. Trac: Trustworthy retrieval augmented chatbot. *arXiv preprint arXiv:2307.04642*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Liran Ringel, Regev Cohen, Daniel Freedman, Michael Elad, and Yaniv Romano. 2024. Early time classification with accumulated accuracy gap control. *arXiv preprint arXiv:2402.00857*.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Yunpeng Xu, Wenge Guo, and Zhi Wei. 2024. Conformal ranked retrieval. *arXiv preprint arXiv:2404.17769*.

Wang Yining, Wang Liwei, Li Yuanzhi, He Di, Chen Wei, and Liu Tie-Yan. 2013. A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th annual conference on learning theory*.

A Additional Experiments

Here evaluate our method with the E5-Mistral embedder on SCIFACT and FIQA datasets. Results, presented in Table 3, show our method consistently outperforms competitors. Moreover, using E5-Mistral leads to improved performance in both empirical coverage and average group size compared to BGE-large-1.5.

In addition to the aforementioned experiments, we compared our method against alternative transformations: *Max Score*, where scores are normalized by dividing each by the maximum score, and *Z-Score*, which standardizes the initial retrieved scores. The results, summarized in Table 4, show that our score refinement transformation outperforms these other refinement methods in the vast majority of cases.

Dataset	α	Method	Emp. Cov.	Avg. Grp. Size
SCIFACT	0.10	Baseline	0.91	68.91
		APS	0.94	17.46
		TopK	0.95	19.0
		Ours	0.93	**15.09**
	0.05	Baseline	0.96	311.73
		APS	0.98	139.36
		TopK	0.99	150.0
		Ours	0.97	**48.71**
	0.03	Baseline	0.99	1093.85
		APS	1.0	324.09
		TopK	1.0	368.0
		Ours	1.0	**127.29**
FIQA	0.10	Baseline	0.91	144.31
		APS	0.9	46.48
		TopK	0.89	38.0
		Ours	0.9	**33.35**
	0.05	Baseline	0.96	458.79
		APS	0.95	123.09
		TopK	0.94	108.0
		Ours	0.94	**67.21**
	0.03	Baseline	0.98	710.86
		APS	0.97	439.64
		TopK	0.96	193.0
		Ours	0.96	**143.76**

Table 3: Empirical coverage and average group size for FIQA and SCIFACT, alpha values, and methods using the e5-mistral-7b-instruct.

Dataset	α	Method	Emp. Cov.	Avg. Grp. Size
FEVER	0.10	Baseline	0.90	4.81
		Max Score	0.87	1.19
		Z-Score	0.85	1.63
		Ours	0.87	**1.18**
	0.05	Baseline	0.95	9.47
		Max Score	0.93	1.89
		Z-Score	0.92	2.44
		Ours	0.93	**1.67**
	0.03	Baseline	0.97	15.63
		Max Score	0.96	2.88
		Z-Score	0.95	3.28
		Ours	0.95	**2.37**
SCIFACT	0.10	Baseline	0.91	231.17
		Max Score	0.83	20.68
		Z-Score	0.88	22.01
		Ours	0.85	**14.07**
	0.05	Baseline	0.97	760.75
		Max Score	0.86	31.01
		Z-Score	0.91	52.91
		Ours	0.89	**29.59**
	0.03	Baseline	0.98	1211.11
		Max Score	0.94	**132.31**
		Z-Score	0.93	197.77
		Ours	0.97	160.66
FIQA	0.10	Baseline	0.89	417.77
		Max Score	0.87	83.23
		Z-Score	0.87	78.02
		Ours	0.86	**56.72**
	0.05	Baseline	0.94	846.0
		Max Score	0.92	254.8
		Z-Score	0.92	217.79
		Ours	0.92	**190.5**
	0.03	Baseline	0.96	1206.93
		Max Score	0.94	380.62
		Z-Score	0.94	437.01
		Ours	0.95	**347.16**

Table 4: Ablation study comparing different score refinement methods with BGE-large-v1.5 encodings. The table shows empirical coverage and average group size for different datasets and methods. Bold values indicate the best performance for each α.