Not All Relevance Scores are Equal: Efficient Uncertainty and Calibration Modeling for Deep Retrieval Models

Daniel Cohen
Brown University
Providence, R.I., USA
daniel_cohen@brown.edu

Bhaskar Mitra
Microsoft
Montreal, Canada
bmitra@microsoft.com

Oleg Lesota
Johannes Kepler University
Linz, Austria
oleg.lesota@jku.at

Navid Rekabsaz
Johannes Kepler University
Linz Institute of Technology, AI Lab
Linz, Austria
navid.rekabsaz@jku.at

Carsten Eickhoff
Brown University
Providence, R.I., USA
carsten@brown.edu

ABSTRACT

In any ranking system, the retrieval model outputs a single score for a document based on its belief on how relevant it is to a given search query. While retrieval models have continued to improve with the introduction of increasingly complex architectures, few works have investigated a retrieval model's belief in the score beyond the scope of a single value. We argue that capturing the model's uncertainty with respect to its own scoring of a document is a critical aspect of retrieval that allows for greater use of current models across new document distributions, collections, or even improving effectiveness for down-stream tasks. In this paper, we address this problem via an efficient Bayesian framework for retrieval models which captures the model's belief in the relevance score through a stochastic process while adding only negligible computational overhead. We evaluate this belief via a ranking based calibration metric showing that our approximate Bayesian framework significantly improves a retrieval model's ranking effectiveness through a risk aware reranking as well as its confidence calibration. Lastly, we demonstrate that this additional uncertainty information is actionable and reliable on down-stream tasks represented via cutoff prediction.

CCS CONCEPTS

• **Information systems** → **Information retrievals**; **Retrieval models and ranking**; • **Computer systems organization** → *Neural networks.*

KEYWORDS

uncertainty, neural networks, calibration, search

ACM Reference Format:
Daniel Cohen, Bhaskar Mitra, Oleg Lesota, Navid Rekabsaz, and Carsten Eickhoff. 2021. Not All Relevance Scores are Equal: Efficient Uncertainty and Calibration Modeling for Deep Retrieval Models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3404835.3462951

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '21, July 11–15, 2021, Virtual Event, Canada
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462951

1 INTRODUCTION

Recent work in neural information retrieval (IR) models have achieved impressive performance on a variety of retrieval tasks whether the models are based on pre-trained Transformer architectures [28, 40, 57, 58] or learned from scratch [11, 23, 35]. These state-of-the-art models treat their estimates of a document's relevance as a deterministic score. While effective, this deterministic perspective obfuscates a large amount of critical information that a user could use to determine whether their query is effective or when they have gone so far down a ranked list that the model is no longer sure of its scores. Ideally, an effective IR system should be able to gracefully convey when it is no longer effective or confident in its rankings, which is a substantial risk given neural models' vulnerability to out-of-distribution inputs from a different collection or even a new first stage ranker [6, 26, 30].

In order to fulfill the above criteria, a retrieval model should be capable of displaying this *uncertainty* over document relevance prediction through a distribution of possible scores as demonstrated in Figure 1. Furthermore, the retrieval model should be expressive enough such that the mean score of the document should convey the model's belief in the actual relevance while its corresponding variance captures the model's uncertainty – a high variance should indicate that the model is unsure of a document's relevance even if it is highly placed in a ranked list. Beyond this expression of uncertainty, calibration is another desirable quality as a comparatively low relevance score with respect to another candidate document should reflect a proportional likelihood that the lower scoring document is less relevant.

These concepts of calibration and uncertainty have been touched on in previous work in IR, most specifically query performance prediction [49, 50, 56] and query cutoff prediction [3, 8, 31]. In these settings, an external model attempts to capture a portion of the model's uncertainty as a function of the input data and its deterministic output. These post-retrieval approaches rely on scoring a large number of documents while simultaneously only

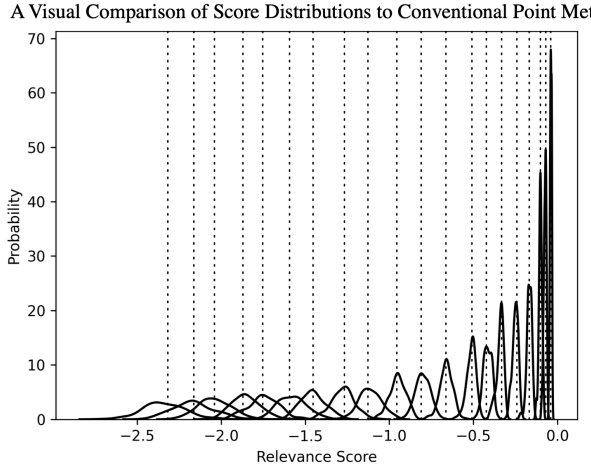

Figure 1: A visual comparison of the conventional interpretation of neural model rankings which provides a single score for each document (represented by the dashed vertical line), and a Bayesian perspective which captures the uncertainty over each score. Highly relevant documents have a narrow score distribution as the retrieval model is confident in itself. In the case of documents where it is uncertain, the model is able to convey its uncertainty by means of a wider distribution of possible relevance scores. The plot represents a single query and select candidate documents from TREC Deep Learning Track 2019 with multiple raw score estimates per document produced by a Bayesian mini BERT.

representing uncertainty over the documents rather than the model itself.

In contrast, ensemble methods such as models featuring on MS MARCO leaderboards[1] capture both types of uncertainty – aleatoric which is uncertainty over the input documents as mentioned above and epistemic uncertainty which is uncertainty over the parameters of a model. Unfortunately, ensemble methods become computationally expensive as retrieval models grow in size and complexity, and results in a substantial obstacle given the prevalence of BERT and other large transformer architectures [26, 28, 40, 41]. As the objective is to maximize performance while ranking as many top n documents as possible, running m models simultaneously results in $\frac{n}{m}$ fewer ranked documents. Lastly, in cases where these large models are used as pre-trained encoders, ensembles do not adequately capture uncertainty over these shared parameters.

As such, we approach capturing uncertainty from a Bayesian perspective where leveraging a convenient property of dropout [51] can be treated as a form of variational inference, referred to as Monte Carlo dropout [15]. In this setting, dropout induces a stochastic ranking model which creates a distribution of scores as the dropout mechanism outputs different values each time it is run over the same input candidate document. The characteristics of this distribution then allow us to capture both aleatoric and epistemic uncertainty. While MC sampling still relies on an infeasible number m of forward passes over the ranking model, we expand on this work with a theoretically motivated extension where only the last layer needs to be Bayesian to capture both epistemic and aleatoric uncertainty. By efficiently sampling from the last layers, we are able to not only attain a distribution of scores on par with standard deterministic models, but also leverage this uncertainty information to improve both final rankings and a downstream task of cut off prediction with a risk aware reranking.

Succinctly we summarize our research contributions as:

(1) An efficient approximate Bayesian framework for any neural model trained with pairwise cross entropy or pairwise hinge loss.
(2) A rank based calibration metric that facilitates uncertainty comparison across retrieval models.
(3) A risk aware reranking method that significantly improves reranking performance.
(4) Exposing the actionable information contained in uncertainty for downstream tasks via cut off prediction.

2 RELATED WORK

Given the high prevalence of probabilistic ranking approaches in IR, the formal concept of uncertainty for retrieval was first discussed by Zhu et al. [60]. In their work, they treat the variance of a probabilistic language model [44] as a risk-based factor to improve its performance for retrieval. However, relying on a generative model's self reported uncertainty in modern models often results in high calibration errors [38]. Furthermore, they assume all document uncertainty to be normal in nature, whereas we demonstrate that a significant distribution shift occurs across rank positions. A modern perspective on this is through exploiting neural generative IR models. While the relevance estimation mechanism in current neural/deep ranking models are rooted in text-to-text matching principles, the probabilistic/generative paradigm views relevance estimation as the probability of generating query given document. This paradigm has a long history in IR research, initiating from Ponte and Croft [44]. In this regard, few recent works have provided a modern interpretation of generative IR models, primarily through exploiting sequence-to-sequence neural generative models. For instance, dos Santos et al. [14] exploit large-scale sequence-to-sequence Transformer-based models to rank answers according to their generation probability for given a question, while Nogueira et al. [42] use a sequence-to-sequence model to first generate queries conditioned on a document, and then use them to expand the document. The probabilistic nature of generative models makes them readily capable of estimating the aleatoric uncertainty. A natural way of estimating uncertainty in generative models is by exploiting the entropy of the next term prediction distributions, defined over the vocabulary set, as proposed by Izacard and Grave [24] in the context of abstractive summarization. In the light of studying uncertainty in IR, we also investigate the characteristics of this self-reported entropy-based uncertainty using a BERT-based deep generative IR model for the downstream task of cutoff prediction. Lastly, we highlight the simultaneous work of Penha and Hauff [43], where they explore the impact of Bayesian and ensemble uncertainty in the conversational response space under BERT.

[1] https://microsoft.github.io/msmarco/

In contrast to our work, their approach requires a full forward pass of BERT for every Monte Carlo sample.

In approaches that consider both probabilistic and discriminative models, a close parallel in spirit are the tasks of query performance prediction (QPP) and cutoff prediction. In QPP, the aim is to determine how difficult a query is given a collection. Within QPP, the concept of post-retrieval query prediction, which relies on the output of one or more retrieval models, is closest to our work. A variety of works have investigated this problem; however, no QPP method has directly incorporated a retrieval model's uncertainty.

For example, Cummins et al. [9] examine the concept of the distribution of scores by modeling the standard deviation of all candidate document scores to estimate the difficulty of the query. Aslam and Pavlu [2] use an ensemble of multiple ranked lists to predict the difficulty of a query by examining the diversity through Jensen-Shannon divergence. In subsequent work, Roitman et al. [48] attempt to achieve a mean document score for each query, introducing the notion of calibrated scores for retrieval. An alternative perspective that attempts to capture this uncertainty is via query perturbation approaches [59]. These methods inject noise into the initially ranked documents to determine the robustness of the ranked list which sheds some light on both aleatoric and epistemic uncertainty.

In a similar vein to QPP, cutoff prediction attempts to identify the optimal cutoff point to maximize some non-monotonic metric such as F_1 or to determine a set of candidate documents to pass on to the next stage for cascade based retrieval [3, 8, 31]. The motivating hypothesis is that retrieval models become increasingly volatile and unstable as documents drift further from the training distribution. As such, external models are trained to identify this instability and determine when the model is no longer reliable while attempting to learn the retrieval model's uncertainty through deterministic document scores.

Finally, in the context of classical IR models enhanced with word embeddings, Rekabsaz et al. [46] showcase the benefits of exploiting the uncertainty of word-to-word similarities for identifying a reliable threshold to filtering highly similar terms. In their work, the uncertainty is defined as the variance over the similarities achieved from an ensemble of neural word embeddings, all trained under the same learning configuration.

3 MEASURING RETRIEVAL UNCERTAINTY

In this section, we first define the problem and motivation prior to introducing the efficient Bayesian framework. As this framework relies on specific assumptions about the loss function used to train a retrieval model, we extend this work to not only work on pairwise cross entropy, but pairwise hinge loss to cover a wide spectrum of model architectures.

Subsequently, we discuss the issue of traditional calibration metrics when considering the ranking problem and posit an alternative metric. Finally, we discuss our risk aware reranking method that leverages the uncertainty information produced from the approximate Bayesian framework.

3.1 Problem Statement and Motivation

Let $Q = \{q_1, \ldots, q_n\}$ be the set of queries, and $C = \{d_1, \ldots, d_m\}$ be the collection of documents some retrieval model parameterized by $\theta \in \mathbb{R}^G$ is trained to retrieve over. Our dataset is then $\mathcal{D} = Q \times C$, such that $f_\theta : Q \times C \rightarrow \mathbb{R}$ produces a score for each query-document pair evaluated. The task is then to find an ordering of scores that maximizes an external metric such as user satisfaction or relevance and is approximated via mean average precision, nDCG, or MRR among other functions. The vast majority of retrieval models rely on stochastic weight optimization to achieve a maximum likelihood estimates (MLE) or maximum a posteriori (MAP) approximation for θ configuration. While effective, this setting produces point estimates for each candidate in $\mathcal{D}$, removing all uncertainty and confidence estimates, from the predictions. At this point, areas of research such as QPP and cut off prediction try to determine these quantities heuristically via score and document distributions [2, 8, 18, 31, 49, 50, 56]. This task has become increasingly challenging with the changing nature of neural retrieval models as previously established post-retrieval QPP methods are not as effective for neural models [19].

We therefore propose capturing this uncertainty information directly from the model by enforcing a Bayesian view on a portion of θ. In doing so, the distribution over our weights induces a distribution over our candidate scores, allowing for downstream uncertainty and confidence estimates for each document. The remainder of this section covers our efficient Bayesian retrieval framework, uncertainty calibration algorithm, and risk aware reranking.

3.2 Efficient Bayesian Neural Retrieval

We first introduce Bayesian inference, and then an efficient interpretation that allows for measuring uncertainty in real world environments. As discussed, the conventional process of training a retrieval model f_θ is through a form of stochastic gradient descent to achieve an estimate of the MLE over some dataset $\mathcal{D}$, $P(\mathcal{D}|\theta)$, or MAP if regularized, $P(\mathcal{D}|\theta)P(\theta)$, which minimizes some loss function,

$$\theta = \text{argmin}_\theta \mathcal{L}(\mathcal{D}, f, \theta).$$

This representation of f_θ unfortunately discards all other parametrizations of θ which are potentially just as capable of determining the relevance of a document. The hypothesis is that some parameterizations will be better for scoring some subset of query-document pairs while other parameterizations will excel on other areas of the dataset. The disparity between scores across the space of θ allows one to then capture the uncertainty of the model given a query-document pair [15, 27, 36, 54].

We therefore propose using a Bayesian approach to retain this uncertainty over our model by modeling the full posterior, $P(\theta|\mathcal{D})$ with

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} = \frac{P(\mathcal{D}|\theta)P(\theta)}{\int_\theta P(\mathcal{D}|\theta)P(\theta)d\theta}. \quad (1)$$

The advantage of modeling the full posterior is that we are then able to consider different parameterizations of the model weighted by how well the data supports such a weight configuration. At prediction, our posterior allows us to account for the likelihood of each paramaterization of our retrieval model through the marginalized predictive distribution:

$$P(y|x, \mathcal{D}) = \int_\theta P(y|x, \theta)P(\theta|\mathcal{D})d\theta. \quad (2)$$

We then exploit this to capture the retrieval model's uncertainty at retrieval time.

Computing the posterior, specifically $\int_\theta P(\mathcal{D}|\theta)P(\theta)d\theta$, in neural architectures is intractable, and so we use an approximation scheme $q(\theta) \approx P(\theta|\mathcal{D})$, called variational inference [21]. The objective of variational inference is to find some $q(\theta)$ that sufficiently fits the data while minimizing the KL divergence to the true posterior $P(\theta|\mathcal{D})$ through the evidence lower bound (ELBO),

$$\text{ELBO}(q) = \mathbb{E}[\log P(\mathcal{D}|\theta)] - KL(q(\theta)||P(\theta)). \tag{3}$$

3.2.1 Monte Carlo Dropout. We approximate the posterior distribution via Monte Carlo (MC) sampling based on dropout (MC-Dropout) which is a stochastic regularization technique [15]. We then leverage a recent result by Kristiadi et al. [30] which bounds the confidence of the model in ReLU networks [37] while simultaneously reducing the computation cost of conventional Bayesian uncertainty estimation.

MC-Dropout approximates the posterior by inducing a distribution for each weight as a mixture of two simpler distributions. Letting $\theta = \mathbf{W_i}_1^L$, $\mathbf{W}_i \in \mathbb{R}^{K_i-1 \times K_i}$ for an L layer neural architecture, MC-Dropout models the variational distribution q via

$$o_{i,j} \sim \text{Bernoulli}(p_i) \text{ for } i = 1, \ldots, L, j = 1, \ldots K_i \tag{4}$$

$$\mathbf{W}_i = \mathbf{M}_i \cdot diag([o_{i,j}])_{j=1}^{K_i}. \tag{5}$$

Here, $o_{i,j}$ are Bernoulli random variables governed by p_i, and $\mathbf{M}_i$ are variational parameters to be optimized. $\mathbf{M}$ can be thought of as the mean of θ under a Gaussian with almost 0 variance (essentially delta peaks). The combination of the Bernoulli dropout creates a Gaussian distribution with non-zero variance to form on $\mathbf{W}$ and allows us to model a non trivial approximate posterior probability. We further adopt a concrete perspective on MC-Dropout [16], and include $\mathbf{p} = {p_i}_1^L$ as another variational parameter to allow the dropout rate to be a learnable parameter.

At this point, Monte Carlo sampling is used to approximate Equation 3 to get an unbiased estimate over N draws of θ from our variational distribution:

$$\int_\theta p(y|x, \theta, f)q(\theta)d\theta \approx \frac{1}{N}\sum_{t=1}^{N} p(y|x, \hat{\theta}_t, f). \tag{6}$$

The real strength of MC-dropout is the minimal change to standard training procedures. If we assume a standard neural network loss for regression and compare it to the ELBO used for variational inference, we observe a close parallel to standard MLE training with regularization [52]:

$$\frac{1}{2}||y - f_\theta(x)||^2 + ||\theta||^2 \approx -\frac{1}{\tau}p(y|f_\theta, x) + \text{KL}(q(\theta)||p(\theta)). \tag{7}$$

As the variational parameters are a delta distribution with the mean at θ, we have the property that $W_{MLE} = \mathbf{M}$ as long as we constrain L_2 regularization to equal the KL divergence. Therefore training a standard neural network with dropout is often equivalent to performing variational inference over the weights of the retrieval model. This feature has been used with success in a variety of computer vision tasks to capture model uncertainty, but the unique nature of pairwise loss functions used in IR, specifically the popular pairwise hinge loss, prevents the direct use of MC-dropout for ranking.

In order to apply MC-Dropout without violating Equation 7, we relax the conventional pairwise hinge loss to facilitate a Gaussian interpretation. In the case of IR, we can view pairwise hinge loss as minimizing the Euclidean distance between the regression goal and a random point within our collection $\mathcal{D}$ conditioned on some initial ranking r (BM25 sampling, random, kNN, etc):

$$\mathbb{E}[||1 - f_\theta(x_+) + f_\theta(x_-)||^2] = \mathbb{E}[||y - f_\theta(x_+) + f_\theta(X)||^2] \tag{8}$$

$$\text{s.t. } X \sim p(\mathcal{D}|r). \tag{9}$$

If we fix $X = x_-$ and define a new function $g_\theta(x_+, x_-) = f_\theta(x_+) + f_\theta(x_-)$, we can treat pairwise hinge loss as a standard regression optimization over g. This allows us to still optimize the negative log likelihood of a Gaussian, satisfying the distribution requirements for MC-Dropout:

$$\frac{1}{2}||y - g_\theta(x_+, x_-)||^2 = -\frac{1}{\tau}p(y|f_\theta, x_-, x_+) + c. \tag{10}$$

At evaluation time, setting $g_\theta(x, 0)$ where we have a deterministic output of 0 for x_-, we can uncover the uncertainty of $f_\theta(x_+)$. We also consider pairwise cross entropy loss where the direct parallel to its probabilistic interpretation allows for a standard application of MC-Dropout.

3.3 Efficient Ranking Uncertainty

As retrieval models have become increasingly computationally demanding [12, 32, 58], the multiple passes required for MC-dropout are not always feasible. In order to retain the effective uncertainty information, we leverage a recent theorem from Kristiadi et al. [30] which demonstrates that in the case of binary classification, only the last layer of a model needs to be Bayesian to capture uncertainty information and correct overconfidence. When this occurs, as the test data becomes increasingly distant from the well fit training data, the estimates approach a distribution determined only by the mean and largest eigenvalue of θ:

Theorem 3.1. *Let $g : \mathbb{R}^d \to \mathbb{R}$ be a binary linear classifier defined by $g(\phi(x)) := \theta^\intercal \phi(x)$, where $\phi : \mathbb{R}^n \to \mathbb{R}^d$ is a fixed ReLU network and let $\mathcal{N}(\theta|\mu, \Sigma)$ be the Gaussian approximation over the last layer's weights with eigenvalues of Σ as $\lambda_1 \leq \ldots \leq \lambda_r$. Then for any input $\mathbf{x} \in \mathbb{R}^n$ and $\delta > 0$,*

$$\lim_{\delta \to \infty} \sigma(|z(\delta \mathbf{x})|) \leq \sigma\Big(\frac{||\theta_\mu||}{\sqrt{\pi/8\lambda_1}}\Big)$$

While the theoretical result is proven only for the binary classification case, the authors demonstrate its application to softmax with multiple classes. In this work, we demonstrate its empirical validity to the case of ranking with pointwise evaluation where we are able to capture uncertainty information with minimal computation cost.

3.4 Ranked List Uncertainty Calibration

Having introduced an approximate Bayesian retrieval framework using only two dropout layers, the usefulness of this uncertainty information partially depends on how well *calibrated* these document score estimates are with respect to the actual accuracy [17]. Referred to as expected calibration error (ECE), a neural model's calibration is often modeled via binning estimates from [0,1] into M equally distributed buckets, B_m, and measuring how much the

model's confidence in this estimate deviates from the accuracy. I.e. all predictions with a confidence of $p = 0.3$ should have a mean accuracy of 30%. This evaluation is represented as

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \frac{1}{B_m} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i) - \frac{1}{B_m} \sum_{i \in B_m} \hat{p}_i \right| \tag{11}$$

for n samples. However, ECE does not effectively measure a ranking model's calibration for a number of reasons. Neural relevance scores are not calibrated across queries, so document relevance is not distributed on a scale from 0 to 1 where each interval within this range is just as important. This is partially why pairwise or listwise training is substantially more effective than pointwise relevance classification as it is the relative comparison across documents that is most effective, which ECE does not capture. A possible remedy to this issue is to take the softmax over all document scores to force a distribution in $[0, 1]$. However, this is again inconsistent as one can reduce the confidence in any individual document by increasing the total number of documents when taking the softmax. We therefore model uncertainty in a pairwise fashion, where calibration is measured between scored documents from the same query via our proposed expected ranking calibration error (ERCE):

$$ERCE = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \frac{1}{|B_m|} \sum_{(D_i, D_j) \in B_m} P(D_i > D_j) - \frac{1}{|B_m|} \sum_{(D_i, D_j) \in B_m} \mathbf{1}(D_i > D_j) \right|. \tag{12}$$

This allows for a consistent calibration error while still accounting for relevance score distributions being conditioned on queries. The indicator function is defined as

$$\mathbf{1}(D_i > D_j) = \begin{cases} 1 & \text{if ranking } D_i \text{ above } D_j \text{ increases MAP} \\ 0 & \text{if ranking } D_i \text{ above } D_j \text{ decreases MAP,} \end{cases}$$

where MAP is mean average precision. This formulation removes all comparisons between pairs of relevant documents, pairs of non-relevant documents, and documents scored from different queries, which allows for measuring only the calibration between relevant and non-relevant pairs conditioned in the same query. In the case of deterministic models which do not have a probabilistic perspective on relevance, we use the pairwise softmax function to calculate $P(D_i > D_j)$.

3.5 Risk Aware Rerankings

As each document now has a predictive distribution, we are able to rerank a set of candidates based on a user defined allotted risk. One difficulty of this task is that variance can substantially differ across queries which leaves a linear combination of the type $\lambda\mu + (1 - \lambda)\sigma^2,\ \lambda \in [0, 1]$ ill-suited for robust probabilistic rankings. In order to normalize across all query types and outputs, we approach this problem using the cumulative distribution function (CDF) over scores s, $F_S : \mathcal{R} \rightarrow [0, 1]$, which maps a score s to the probability of the document achieving a score less than or equal to s, i.e., $F_S(s) = P(S \leq s)$. This representation has the advantage of normalizing scales to a range of $[0, 1]$ across multiple queries and facilitates query agnostic measures of uncertainty [25]. We then rank candidate documents using Conditional Value at Risk (CVaR),

$$\text{CVaR}_\alpha(S) = \mathbb{E}[S | S \geq F_S^{-1}(\alpha)],$$

where F_S^{-1} is the inverse CDF of sampled scores. This takes the expected score of a document from the top or bottom $(1 - \alpha)\%$ of samples for an optimistic or pessimistic view of outcomes and is often used for risk aware planning or decision making [4, 47]. As CVaR is a coherent risk measure, it satisfies monotonicity and sub-additivity properties and allows us to extend individual document risk to bound total risk for the entire ranked list:

$$\text{CVaR}_\alpha(S_1, \ldots, S_n) \leq \text{CVaR}_\alpha(S_1) + \ldots + \text{CVaR}_\alpha(S_n).$$

Table 1: Training, validation, and test statistics for the collections used. No training, fine-tuning, or validation is performed for any retrieval models on Robust04 thus use no validation set is used.

Collection	Documents	Validation	Test
MS MARCO	9M	6,980	48,598
TREC 2019 DLT	9M	6,980	43
Robust04	0.5M	-	250

4 EXPERIMENTS

We examine four attributes with respect to efficient Bayesian retrieval. First, we examine the hypothesis that the mean weight of the dropout model is equivalent to its deterministic variant. Second, we study the ranking calibration error in the same manner. Third, we investigate the impact of CVaR_α in both optimistic and pessimistic settings for risk-aware rankings. And finally, we evaluate the usefulness of this uncertainty information under the downstream task of cutoff prediction [3, 31].

4.1 Data

We utilize three collections for our experiments. Evaluating retrieval and cutoff performance, we use the MS MARCO [39] passage dataset, the TREC 2019 Deep Learning Track (DLT) [7] based on the original MS MARCO dataset, and Robust04 with its corresponding title queries. While MS MARCO is commonly used to evaluate performance of retrieval models, the one hot relevance judgements limit the investigation of uncertainty in ranking. We therefore use this dataset as a general training collection and use both the TREC 2019 Deep Learning Track and Robust04 collections with their more fine grained relevance judgements to examine uncertainty properties with risk-aware re-rankings, calibration analysis, and cutoff prediction.

With respect to training, validation, and test splits, in the case of MS MARCO, we evaluate with the same validation and test splits as Hofstätter et al. [22]. We do not fine tune or validate on TREC 2019 DLT nor Robust04, and therefore all queries are used to test the models. As demonstrated by Yilmaz et al. [1], an MS MARCO trained BERT architecture is an effective retrieval model for Robust04 which results in a non-trivial evaluation. We transform each full document into the first 512 tokens as its representative passage in a similar fashion to Dai and Callan [10]. Statistics of the collections are provided in Table 1.

4.2 Models

We examine two representative retrieval architectures using our uncertainty framework: BERT [40] and Conv-KNRM [11]. BERT represents the current trend towards large pre-trained transformer architectures, and Conv-KNRM provides insight into how well last layer MC-Dropout works on retrieval models which rely on hand crafted similarity functions as input into the Bayesian layer. Given hardware constraints, we evaluate on tiny and mini BERT (BERT-L2, BERT-L4) versions which achieve close to full BERT performance [53].

4.3 Baselines

For each Bayesian model, we evaluate with respect to the deterministic version of the architecture denoted by subscript D. While the work by Zhu et al. [60] discusses the concept of risk aware reranking by the variance of a language model, this view does not directly apply to the case of modern deep retrieval architectures. In the case cutoff prediction, we include a modern interpretation of their work with a generative BERT-to-Transformer architecture to provide context

For the application task of cutoff prediction, we use the architecture proposed by Liu and Lapata [34], where the document is encoded by the mini/tiny BERT and the query is decoded term-by-term through a 4-layer Transformer decoder conditioned on the encoded document. We train the model following the same procedure as in Nogueira et al. [42], and similar to Izacard and Grave [24] consider relevance as $\log P(Q|D)$, namely the sum of the logarithms of the next term generation probabilities. An uncertainty interpretation for these generative architectures produces entropy values over $P(Q|D)$, and provides useful information as we can compare self reported aleatoric uncertainty to both aleatoric and epistemic uncertainty from our Bayesian approximation.

4.4 Evaluation

4.4.1 Retrieval and Reranking: We evaluate each model using mean reciprocal rank (MRR) and normalized discounted cumulative gain (nDCG@n). MRR was selected due to the single judged relevant passage per query in MS MARCO, while nDCG@20 was used for Robust04 and nDCG@200 for TREC 2019 DLT. We use a large cutoff for TREC to better capture less confident document relevance judgements, and use the traditional lower cutoff for Robust04 as it is already out of distribution.

4.5 Calibration:

We use an adaptive binning scheme to dynamically create 10 equally filled buckets B.

4.5.1 Cutoff Prediction: We use a state-of-the-art transformer based prediction model, Choppy, to determine where to cut a ranked list to maximize F_1 [3]. As the performance is a function of the oracle, we report

$$\frac{F1_{Choppy}}{F1_{Oracle}}$$

where $F1_{Choppy}$ and $F1_{Oracle}$ are the F_1 scores produced by the cutoff identified by Choppy and the oracle cutoff, respectively. We report our results using a ranked list of the top 200 documents for each model. In case of the Bayesian models, each document is represented as $< \mu, \sigma, \delta, H(S) >$, with δ as the skew and $H(S)$ as the entropy over the score distribution S.

4.6 Hyperparameters and Training

For each model architecture, we use the same hyperparameters used for the deterministic variant. BERT models were trained with a learning rate of 1×10^{-4} while the Conv-KNRM used 1×10^{-3} using Adam optimization [29]. We use two additional feed-forward layers after the main model to facilitate last layer MC-Dropout of sizes $[K, K]$, $[K, f]$ where K is the output size of the kernel features in Conv-KNRM or the hidden dimension of BERT, and f is the final output dimension of the architecture. In case of the stochastic models, a single sample was used during training and 150 samples were used at inference time for each query-document pair.

5 RESULTS

In this section, we first compare our Bayesian interpretation against baseline deterministic models to ensure the stochastic nature does not significantly degrade actual retrieval performance. We then study the behaviour of uncertainty across relevance scores and evaluate risk aware re-ranking. We further report ERCE measures prior to discussing downstream usefulness with cutoff prediction.

5.1 Comparison to Deterministic Models

While we have discussed the many benefits of well calibrated measures of uncertainty, the introduction of stochastic scoring should not harm actual retrieval performance. As such, we address our first hypothesis: *Does efficient MC-Dropout at inference time result in the same mean performance as a deterministic retrieval model?* As Table 2 demonstrates, mean MC-Dropout document scoring achieves parity with its deterministic variants across collections and architecture types. While there is some discrepancy with BERT-L2$_B$ slightly under performing and Conv-KNRM$_B$ and BERT-L4$_B$ outperforming the deterministic versions, the small differences paired with the nature of stochastic gradient update demonstrates that this framework is a safe way to include uncertainty information into a variety of architectures without the risk of deteriorating performance. This parity also confirms that using MC-Dropout empirically satisfies Equation 7.

On Robust04, we observe the largest difference in performance of approximately -2.6% for BERT-L4. This result highlights that uncertainty aware models do not inherently perform better on out-of-distribution data; however, we discuss how the models are able to convey their uncertainty and how to leverage this information to improve performance.

5.2 Risk-Aware Ranking

Incorporating the CVaR based ranking (Table 3) allows us to leverage the model's reported uncertainty, consistently improving performance across all collections. After accounting for the variance and skew of each document's score distribution, MC-Dropout based models significantly improve performance both with respect to their mean performance, but also their deterministic baseline models by 3-5%. This performance increase is present in both the optimistic (CVaR$_+$) and pessimistic (CVaR$_-$) setup, suggesting that focusing

Table 2: Comparison of MRR and nDCG@n performance between deterministic ($_D$) and mean Bayesian MC-Dropout ($_B$) model performance. Document scores were calculated as mean performance over 150 samples for the stochastic models. n was 200 for TREC 2019 DLT and 20 for Robust04. * represents statistical significance of $p < 0.05$ using paired t test.

Collection	Models					
	MRR					
	BERT-L2$_D$	BERT-L2$_B$	BERT-L4$_D$	BERT-L4$_B$	Conv-KNRM$_D$	Conv-KNRM$_B$
MS MARCO	0.305	0.301 (-1.3%)	0.308	0.308 (+0.1%)	0.279	0.280 (+0.4%)
TREC 2019 DLT	0.912	0.916 (+0.5%)	0.929	0.936 (+0.8%)	0.900	0.901 (0.0%)
MS MARCO → Robust04	0.617	0.628* (+1.8%)	0.657	0.640* (-2.6%)	0.591	0.598 (+1.2%)
	nDCG@200,20					
	BERT-L2$_D$	BERT-L2$_B$	BERT-L4$_D$	BERT-L4$_B$	Conv-KNRM$_D$	Conv-KNRM$_B$
MS MARCO	0.398	0.395 (-0.8%)	0.401	0.401 (-0.1%)	0.380	0.377 (-0.8%)
TREC 2019 DLT	0.582	0.576 (-1.0%)	0.582	0.581 (-0.2%)	0.565	0.567 (+0.4%)
MS MARCO → Robust04	0.431	0.433 (+0.7%)	0.434	0.431 (-0.7%)	0.425	0.426 (+0.2%)

Table 3: Risk-aware rerankings of the top 200 candidate documents for each query using CVaR. CVaR$_1$ indicates standard mean score, CVaR$_+$ and CVaR$_-$ represents taking the optimistic and pessimistic perspectives above and below α accordingly. We report nDCG@200 for TREC DLT and nDCG@20 for Robust04. * denotes statistical significance with respect to baselines with $p < 0.05$ using paired t test.

Collection	Model	CVaR$_+$	CVaR$_-$	CVaR$_1$
TREC 2019 DL	BERT-L2$_D$	0.582	0.582	0.582
	BERT-L2$_B$	0.597*	**0.598***	0.576
	BERT-L4$_D$	0.582	0.582	0.582
	BERT-L4$_B$	**0.606***	0.605*	0.581
	Conv-KNRM$_D$	0.565	0.565	0.565
	Conv-KNRM$_B$	0.584*	**0.585***	0.567
Robust04	BERT-L2$_D$	0.398	0.398	0.398
	BERT-L2$_B$	0.411*	**0.412***	0.402
	BERT-L4$_D$	0.400	0.400	0.400
	BERT-L4$_B$	**0.407***	**0.407***	0.395
	Conv-KNRM$_D$	0.382	0.382	0.382
	Conv-KNRM$_B$	**0.404***	0.403*	0.386

on the tails of either end of the distributions provides pertinent uncertainty information with respect to the model outputs. Interestingly, we see a similar change in performance for the Robust04 collection, suggesting that MC-Dropout models are equally capable of expressing risk on the data used to train the model in addition to collections where all documents are out of distribution.

To provide additional insight into how this risk aware re-ranking functions over different candidate ranking positions, we plot the relationship between μ to σ^2 and μ to skew δ in Figure 2. It is in these figures that we notice the impact that the neural architecture has on the uncertainty of the documents. Both BERTs have a direct relationship with predicted relevance such that the model is most certain about highly relevant documents, and a non-linear increase in uncertainty as documents move further away from the query. With this increase we also see the scores converge to a normal distribution. The top most relevant documents are highly skewed with low variance within the range of 0 to -0.1 relevance score with the majority of its mass on the right of the distribution as indicated in Figure 3. As the variance increases and relevance scores drop, the score distribution follows Theorem 3.1 and approaches a normal distribution.

With respect to Conv-KNRM$_B$, the same pattern is present but not as salient. The upper bound of variance continues to grow while a large portion of non-relevant documents still have very low relevance. However, the asymmetric nature across *tanh* used in ConvKNRM demonstrates that uncertainty is still being expressed over both the handmade kernel features and the condensing nature of *tanh*. This same asymmetry is found in the skew plot, with a greater number of documents expressing high positive skew than the fewer highly relevant documents with large negative skew. We hypothesize that the polarizing skew values is due to the high gradient of *tanh*.

The variance and skewness trends are reinforced as the risk-aware CVaR consistently increases metric performance as we increase the n cutoff for nDCG@n, suggesting that risk based re-ranking can be most utilized for high recall tasks or where effective performance is required outside of the top few documents. This result is unsurprising when considering the inverse CDF plots in Figure 3. The highly ranked documents have an almost point distribution while the lower ranked documents exhibit significant uncertainty.

5.3 Calibration

We now inspect the expressiveness of a stochastic retrieval model's uncertainty to answer *Does efficient MC-Dropout improve uncertainty calibration for IR models?* We record ERCE in Table 4 for all model and collection permutations and observe a substantial decrease in calibration error resulting in approximate Bayesian models being $\tilde{3}0\%$ more calibrated. This confirms our hypothesis that Bayesian retrieval models will have better expressiveness of their confidence. It follows from recent results in computer vision [17],

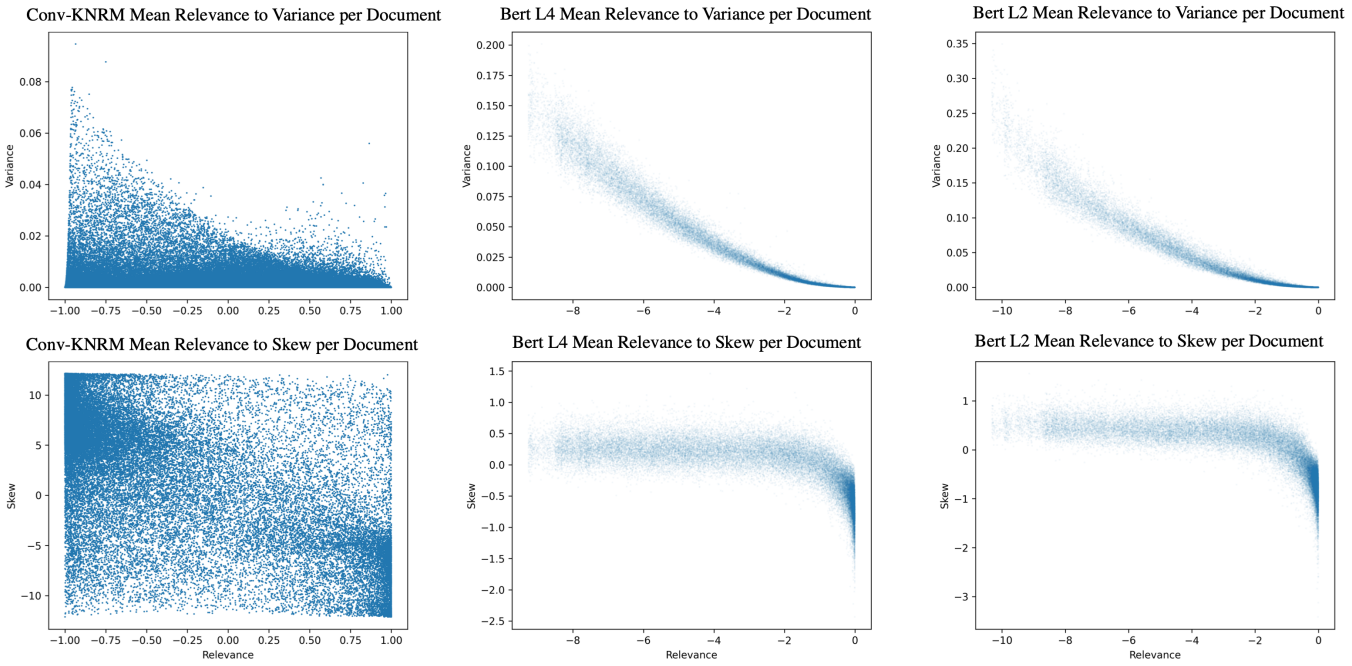

Figure 2: The mean to variance and mean to skew relationship for each document scored by Conv-KNRM$_B$ (top), BERT-L2$_B$ (middle), and BERT-L4$_B$ on TREC 2019 DLT.

Table 4: Expected ranking calibration error (ERCE) for BERT models with respect to their deterministic variants (lower is better).

Collection	Models					
	BERT-L2$_D$	BERT-L2$_B$	BERT-L4$_D$	BERT-L4$_B$	Conv-KNRM$_D$	Conv-KNRM$_B$
TREC 2019 DLT	0.703	**0.465**	0.700	**0.507**	0.519	**0.452**
MS MARCO → Robust04	0.493	**0.396**	0.477	**0.395**	**0.256**	0.264

where non Bayesian models are poorly calibrated. These results explain the natural separation of mean document scores in Figure 3 as the highly relevant documents are clustered together while the lower ranked documents show a greater spread. One perspective on this is that the increased spread of scores across the range of relevance expressed by the model can be viewed as the relative confidence that they are ranked in the right order.

5.4 Downstream Application: Cutoff Prediction

Having discussed the improved calibration and risk based re-ranking, we now address our question *Is uncertainty information actionable in the context of downstream tasks?* To do so, we use the cutoff prediction task where the objective is to find a cutoff point in a ranked list that maximizes some non-monotonic metric. The motivation, discussed by Lien et al. [31], is that at some point, a neural model loses effectiveness as documents move further away from the query. Using the Choppy cutoff predictor from Bahri et al. [3], we compare the cutoff performance between the information contained in a deterministic model, a modern version of Zhu et al. [60]'s risk based language model, and our Bayesian model's score distribution in Table 5. As shown, we observe a significant improvement to the upper performance bound (oracle) under the Bayesian framework when compared to the deterministic models. As indicated in Section 3.2.1, the mean weights of the MC-Dropout models closely approximates those of their deterministic cousins. This suggests, combined with the visual inspection of Figure 1, that the key determining factors are the variance, entropy, and skew values as a function of document relevance.

Examining the cutoff performance across TREC DLT 2019, we note an approximate 9% increase in cutoff accuracy when including the additional uncertainty information, confirming our hypothesis that the uncertainty information displayed by the MC-Dropout models can be used in downstream decision making. Moving to the Robust04 results, we see a greater increase in comparative performance. As the retrieval models are now out of distribution, there exists significant variance across all ranking positions which introduces noise into the additional dimensions. Following the same trends discussed in risk-aware re-ranking, we see the greatest improvement using an initial ranked list of the top 200 documents. As we decrease the set of candidate documents to be re-ranked, the

Table 5: Choppy [3] performance as a percentage of oracle cutoff under the F1 metric. G is the generative baseline using BERT-L2 and BERT-L4, and * denotes $p < .05$ significance using t test with respect to baseline variants of the same architecture.

Collection	Models							
	BERT-L2$_D$	BERT-L2$_B$	BERT-L2$_G$	BERT-L4$_D$	BERT-L4$_B$	BERT-L4$_G$	Conv-KNRM$_D$	Conv-KNRM$_B$
TREC 2019 DLT	77.3%	**80.6%***	74.5%	73.6%	**79.9%***	73.8%	75.3%	**86.4%***
Robust04	74.1%	**78.9%***	62.4%	75.4%	**78.1%***	63.3%	66.2%	**77.7%***

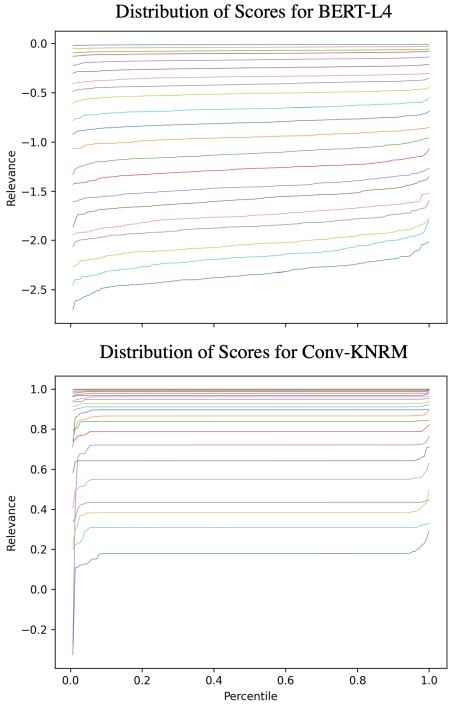

Figure 3: Empirical CDFs of a subset of documents for a single query on BERT-L4$_B$ (top), Conv-KNRM$_B$ (bottom) for the TREC 2019 DLT dataset. Documents were selected for every 10^{th} rank position $(0, 10, 20, 30 \ldots, 200)$.

performance difference between deterministic and Bayesian variants closes. At the top 50 candidates we record only a 2% difference in performance across model frameworks.

Remarking on the related work of capturing uncertainty through generative retrieval models, we further highlight the performance gap between the generative BERT and the Bayesian BERT models [24, 60]. Using the generative framework introduced by Liu and Lapata [34] where the BERT component uses BERT-L2 or BERT-L4, its scores and entropy values representing uncertainty over relevance fail to achieve the same level of calibrated uncertainty as the approximate Bayesian approach. This highlights the finding by Nalisnick et al. [38] that generative models, while capable of expressing uncertainty, are often overconfident in their own self-estimates, resulting in uncertainty measures which are not as robust when compared to those made by Bayesian models. Further, the uncertainty values self-reported by the model are substantially worse when out of distribution on Robust04 with close to a 15% degradation in cutoff performance, which demonstrates the robust uncertainty present in the Bayesian distributions.

5.5 Efficiency

As one of the primary contributions of this work is the efficient modeling of uncertainty, one of the most significant obstacles to be addressed is the computational cost of scoring each query-document pair n times. We benchmark the additional compute cost for our last-layer MC-Dropout on a GTX 1080ti. While not completely free, the additional cost of running 100 additional samples is 0.326 ± 0.012 μs for Conv-KNRM4$_B$ and 0.368 ± 0.016 μs for BERT-L4$_B$, L2$_B$, or for any other large transformer architecture as the additional cost is a function of the final output dimension, not of the retrieval model itself.

6 CONCLUSION

In this paper, we introduced an efficient Bayesian framework to estimate epistemic and aleatoric uncertainty in retrieval models. We demonstrated that query-document uncertainty can be modeled using only the last two layers of a neural model, allowing for its use in state-of-the-art retrieval models relying on BERT – be it pre-trained or as part of the actual retrieval architecture. The performance of these stochastic models stays reasonably close to their deterministic versions while offering substantially more information per document score. Furthermore, the actual scores themselves are better calibrated with each other allowing for a more accurate comparison between documents. These properties enable improved performance on ranking via risk-aware reranking in addition to the downstream task of cutoff prediction when compared to the deterministic versions.

As this approximate Bayesian inference is efficient and conveys useful information for both fully distributed and handcrafted models, there exists a promising body of future work incorporating stochastic models for fairness [13, 33], diversity [5], transparent search [45], dialogue agents [20] and improved sample efficiency when training neural retrieval models [55]. Lastly, we hope to explore the impact of uncertainty modeling in situations where the retrieval model acts as an information gathering agent in larger systems.

7 ACKNOWLEDGEMENTS

This research is supported in part by the NSF (IIS-1956221), ODNI and IARPA via the BETTER program (2019-19051600004). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, ODNI, IARPA or the U.S. Government.

REFERENCES

[1] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3490–3496. https://doi.org/10.18653/v1/D19-1352

[2] Javed A. Aslam and Virgil Pavlu. 2007. Query Hardness Estimation Using Jensen-Shannon Divergence among Multiple Scoring Functions. In *Proceedings of the 29th European Conference on IR Research* (Rome, Italy) *(ECIR'07)*. Springer-Verlag, Berlin, Heidelberg, 198–209.

[3] Dara Bahri, Yi Tay, Che Zheng, Donald Metzler, and Andrew Tomkins. 2020. *Choppy: Cut Transformer for Ranked List Truncation.* Association for Computing Machinery, New York, NY, USA, 1513–1516. https://doi.org/10.1145/3397271.3401188

[4] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. 2015. Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc., 1522–1530. https://proceedings.neurips.cc/paper/2015/file/64223ccf70bbb65a3a4aceac37e21016-Paper.pdf

[5] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446

[6] Daniel Cohen, Bhaskar Mitra, Katja Hofmann, and W. Bruce Croft. 2018. Cross Domain Regularization for Neural Ranking Models Using Adversarial Learning. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 1025–1028. https://doi.org/10.1145/3209978.3210141

[7] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2019. Overview of the TREC 2019 deep learning track.

[8] J. Shane Culpepper, Charles L. A. Clarke, and Jimmy J. Lin. 2016. Dynamic Cutoff Prediction in Multi-Stage Retrieval Systems. In *Proceedings of the 21st Australasian Document Computing Symposium, ADCS 2016, Caulfield, VIC, Australia, December 5-7, 2016*, Sarvnaz Karimi and Mark James Carman (Eds.). ACM, 17–24. http://dl.acm.org/citation.cfm?id=3015026

[9] Ronan Cummins, Joemon Jose, and Colm O'Riordan. 2011. Improved Query Performance Prediction Using Standard Deviation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 1089–1090. https://doi.org/10.1145/2009916.2010063

[10] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 985–988. https://doi.org/10.1145/3331184.3331303

[11] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-Hoc Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 126–134. https://doi.org/10.1145/3159652.3159659

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[13] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. *Evaluating Stochastic Rankings with Expected Exposure.* Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3340531.3411962

[14] Cicero dos Santos, Xiaofei Ma, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Beyond [CLS] through Ranking by Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1722–1727.

[15] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 1050–1059. http://proceedings.mlr.press/v48/gal16.html

[16] Yarin Gal, Jiri Hron, and Alex Kendall. 2017. Concrete Dropout. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 3581–3590. https://proceedings.neurips.cc/paper/2017/hash/84ddfb34126fc3a48ee38d7044e87276-Abstract.html

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 1321–1330.

[18] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2019. Performance Prediction for Non-Factoid Question Answering. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval* (Santa Clara, CA, USA) *(ICTIR '19)*. Association for Computing Machinery, New York, NY, USA, 55–58. https://doi.org/10.1145/3341981.3344249

[19] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2019. Performance Prediction for Non-Factoid Question Answering. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval* (Santa Clara, CA, USA) *(ICTIR '19)*. Association for Computing Machinery, New York, NY, USA, 55–58. https://doi.org/10.1145/3341981.3344249

[20] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2020. *Guided Transformer: Leveraging Multiple External Sources for Representation Learning in Conversational Search.* Association for Computing Machinery, New York, NY, USA, 1131–1140. https://doi.org/10.1145/3397271.3401061

[21] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. 2013. Stochastic Variational Inference. *J. Mach. Learn. Res.* 14, 1 (May 2013), 1303–1347.

[22] Sebastian Hofstätter, Navid Rekabsaz, Carsten Eickhoff, and Allan Hanbury. 2019. On the Effect of Low-Frequency Terms on Neural-IR Models. In *Proc. of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

[23] Sebastian Hofstätter, Hamed Zamani, Bhaskar Mitra, Nick Craswell, and Allan Hanbury. 2020. Local Self-Attention over Long Text for Efficient Document Retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 2021–2024. https://doi.org/10.1145/3397271.3401224

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

[25] Scott M. Jordan, Yash Chandak, Daniel Cohen, Mengxue Zhang, and Philip S. Thomas. 2020. Evaluating the Performance of Reinforcement Learning Algorithms. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 4962–4973. http://proceedings.mlr.press/v119/jordan20a.html

[26] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[27] Alex Kendall and Yarin Gal. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 5580–5590.

[28] Omar Khattab and Matei Zaharia. 2020. *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT.* Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3397271.3401075

[29] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[30] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. 2020. Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, Virtual, 5436–5446. http://proceedings.mlr.press/v119/kristiadi20a.html

[31] Yen-Chieh Lien, Daniel Cohen, and W. Bruce Croft. 2019. An Assumption-Free Approach to the Dynamic Truncation of Ranked Lists. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval* (Santa Clara, CA, USA) *(ICTIR '19)*. Association for Computing Machinery, New York, NY, USA, 79–82. https://doi.org/10.1145/3341981.3344234

[32] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained Transformers for Text Ranking: BERT and Beyond. arXiv:2010.06467 [cs.IR]

[33] Aldo Lipani. 2016. Fairness in Information Retrieval. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 1171. https://doi.org/10.1145/2911451.2911473

[34] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3721–3731.

[35] Bhaskar Mitra, Sebastian Hofstätter, Hamed Zamani, and Nick Craswell. 2020. Conformer-Kernel with Query Term Independence at TREC 2020 Deep Learning Track.

[36] John Mitros and Brian Mac Namee. 2019. On the Validity of Bayesian Neural Networks for Uncertainty Estimation. In *Proceedings for the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, Galway, Ireland, December 5-6, 2019 (CEUR Workshop Proceedings, Vol. 2563)*, Edward Curry, Mark T. Keane, Adegboyega Ojo, and Dhaval Salwala (Eds.). CEUR-WS.org, 140–151. http://ceur-ws.org/Vol-2563/aics_15.pdf

[37] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning* (Haifa, Israel) *(ICML'10)*. Omnipress, Madison, WI, USA, 807–814.

[38] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. 2019. Do Deep Generative Models Know What They Don't Know?. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1xwNhCcYm

[39] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR* abs/1611.09268 (2016). arXiv:1611.09268 http://arxiv.org/abs/1611.09268

[40] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *CoRR* abs/1901.04085 (2019). arXiv:1901.04085 http://arxiv.org/abs/1901.04085

[41] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[42] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. *arXiv preprint arXiv:1904.08375* (2019).

[43] Gustavo Penha and Claudia Hauff. 2021. On the Calibration and Uncertainty of Neural Learning to Rank Models. arXiv:2101.04356 [cs.IR]

[44] Jay M Ponte and W Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. of the 21st annual ACM SIGIR conference on Research and development in information retrieval*. ACM, 275–281.

[45] Jerome Ramos and Carsten Eickhoff. 2020. *Search Result Explanations Improve Efficiency and Trust*. Association for Computing Machinery, New York, NY, USA, 1597–1600. https://doi.org/10.1145/3397271.3401279

[46] Navid Rekabsaz, Mihai Lupu, and Allan Hanbury. 2017. Exploration of a threshold for similarity based on uncertainty in word embedding. In *Proc. of the European Conference on Information Retrieval*.

[47] R. Tyrrell Rockafellar and Stanislav Uryasev. 2000. Optimization of Conditional Value-at-Risk. *Journal of Risk* 2 (2000), 21–41.

[48] Haggai Roitman, Shai Erera, Oren Sar-Shalom, and Bar Weiner. 2017. Enhanced Mean Retrieval Score Estimation for Query Performance Prediction. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands) *(ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 35–42. https://doi.org/10.1145/3121050.3121051

[49] Haggai Roitman, Shai Erera, and Bar Weiner. 2017. Robust Standard Deviation Estimation for Query Performance Prediction. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands) *(ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 245–248. https://doi.org/10.1145/3121050.3121087

[50] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting Query Performance by Query-Drift Estimation. *ACM Trans. Inf. Syst.* 30, 2, Article 11 (May 2012), 35 pages. https://doi.org/10.1145/2180868.2180873

[51] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15, 1 (Jan. 2014), 1929–1958.

[52] Naftali Tishby, E. Levin, and S. Solla. 1989. Consistent inference of probabilities in layered networks: predictions and generalizations. *International 1989 Joint Conference on Neural Networks* (1989), 403–409 vol.2.

[53] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-Read Students Learn Better: The Impact of Student Initialization on Knowledge Distillation. *CoRR* abs/1908.08962 (2019). arXiv:1908.08962 http://arxiv.org/abs/1908.08962

[54] Andrew Gordon Wilson and Pavel Izmailov. 2020. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/322f62469c5e3c7dc3e58f5a4d1ea399-Abstract.html

[55] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. arXiv:2007.00808 [cs.IR]

[56] Hamed Zamani, W. Bruce Croft, and J. Shane Culpepper. 2018. Neural Query Performance Prediction Using Weak Supervision from Multiple Signals. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 105–114. https://doi.org/10.1145/3209978.3210041

[57] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Learning To Retrieve: How to Train a Dense Retrieval Model Effectively and Efficiently. arXiv:2010.10469 [cs.IR]

[58] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. RepBERT: Contextualized Text Embeddings for First-Stage Retrieval. arXiv:2006.15498 [cs.IR]

[59] Yun Zhou and W. Bruce Croft. 2006. Ranking Robustness: A Novel Framework to Predict Query Performance. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management* (Arlington, Virginia, USA) *(CIKM '06)*. Association for Computing Machinery, New York, NY, USA, 567–574. https://doi.org/10.1145/1183614.1183696

[60] Jianhan Zhu, Jun Wang, Michael Taylor, and Ingemar J. Cox. 2009. Risk-Aware Information Retrieval. In *Advances in Information Retrieval*, Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 17–28.