# [EACL'21] On the Calibration and Uncertainty of Neural Learning to Rank Models

Gustavo Penha  Claudia Hauff

Delft University of Technology, Delft, Netherlands

2021.1.25

# Table of Contents


1. Motivation
2. Background and Related Work
3. Model
4. Experiments
5. Conclusion

# Table of Contents


1. Motivation
2. Background and Related Work
3. Model
4. Experiments
5. Conclusion

# Motivation

## Probability Ranking Principle(PRP) (Robertson, 1977)

ranking documents in decreasing order of their probability of relevance leads to an optimal document ranking for ad-hoc retrieval.

For the PNP to hold, ranking models must at least meet the following conditions (Gordan and Lenk, 1991):

- C1 assign **well calibrated** probabilities of relevance, i.e. if we gather all documents for which the model predicts relevance with a probability of e.g. 30%, the amount of relevant documents should be 30%
- C2 the probabilities of relevance are reported **with certainty**

# Motivation

C1 It has been shown that DNNs are not well calibrated in the context of computer vision (Guo et al., 2017).

C2 There are a number of sources of uncertainty in the training process of neural networks that make it unreasonable to assume that neural ranking models fulfill [C2]:

- *parameter uncertainty*
- *structural uncertainty*
- *aleatoric uncertainty*

Given these sources of uncertainty, using point estimate predictions and ranking according to the PRP might not achieve the optimal ranking for retrieval.

# What's Next?

1. Analyze the calibration of neural rankers, specially BERT-based rankers.
2. To model the uncertainty of BERT-based rankers, we propose *stochastic neural ranking models* by applying different techniques to model uncertainty of DNNs:
   - MC Dropout
   - Deep Ensembles

# Table of Contents


1. Motivation
2. Background and Related Work
3. Model
4. Experiments
5. Conclusion

# Calibration and Uncertainty in IR

1. Calibration:
   - The calibration of ranking models has received little attention in IR.
   - Important in automated medical domain due to model interpretability
2. Uncertainty:
   - Treating **variance** as a measure of uncertainty inspired by economics theory.
   - Applications:
     - improve the ranking effectiveness
     - in conversational search, decide between asking clarifying questions and providing a potential answer
     - perform dynamic query reformulation for queries where the intent is uncertain
     - predict questions with no correct answers

# Table of Contents


1. Motivation
2. Background and Related Work
3. Model
4. Experiments
5. Conclusion

# Research Questions

We introduce the models for answering the following research questions:

RQ1 How calibrated are deterministic and stochastic BERT-based rankers?

RQ2 Are the uncertainty estimates from stochastic BERT-based rankers useful for risk-aware ranking?

RQ3 Are the uncertainty estimates obtained from stochastic BERT-based rankers useful for identifying unanswerable queries?

# Measuring Calibration(RQ1)

## Empicical Calibration Error(ECE)

ECE is an intuitive way of measuring to what extent the confidence scores from neural networks align with the true correctness likelihood. It measures the difference between the observed reliability curve (DeGroot and Fienberg, 1983) and the ideal one.

We sort the predictions of the model, divide them into $c$ buckets $\{B_0, ..., B_c\}$, and take the weighted average between the average predicted probability of relevance $avg(B_i)$ and the fraction of relevant documents $\frac{rel(B_i)}{|B_i|}$ in the bucket:

$$ECE = \sum_{i=0}^{c} \frac{|B_i|}{n} \left| avg(B_i) - \frac{rel(B_i)}{|B_i|} \right| \quad (1)$$

where n is the total number of test examples.

# Modeling Uncertainty

1. Define the ranking problem we focus on
2. Deterministic BERT-based ranker baseline model(**BERT**)
3. Our model to answer RQ2 and RQ3:
   - a stochatic BERT-based ranker to model uncertainty(**S-BERT**)
   - a risk-aware BERT-based ranker to take into account uncertainty provided by S-BERT when ranking(**RA-BERT**)

# Conversation Response Ranking

Let $\mathcal{D} = \{(\mathcal{U}_i, \mathcal{R}_i, \mathcal{Y}_i)\}_{i=1}^{N}$ be a dataset consisting of N triplets: dialogue context, response candidates and response relevance labels.

- $\mathcal{U}_i = \{u^1, u^2, ..., u^\tau\}$
- $\mathcal{R}_i = \{r^1, r^2, ..., r^K\}$
- $\mathcal{Y}_i = \{y^1, y^2, ..., y^k\}$

The task is to learn a ranking function $f(.)$ that is able to generate a ranked list for the set of candidate responses $\mathcal{R}_i$ based on their predicted relevance scores $f(\mathcal{U}_i, r)$

# Deterministic BERT Ranker

Pointwise BERT

# Stochastic S-BERT Ranker

We want to obtain a predictive distribution which allows us to extract uncertainty estimates

$$R_r = \{f(\mathcal{U}_i, r)^0, f(\mathcal{U}_i, r)^1, ..., f(\mathcal{U}_i, r)^n\} \tag{2}$$

Two techniques:

1. Deep Ensembles(**S-BERT$^{E}$**)
2. MC Dropout(**S-BERT$^{D}$**)

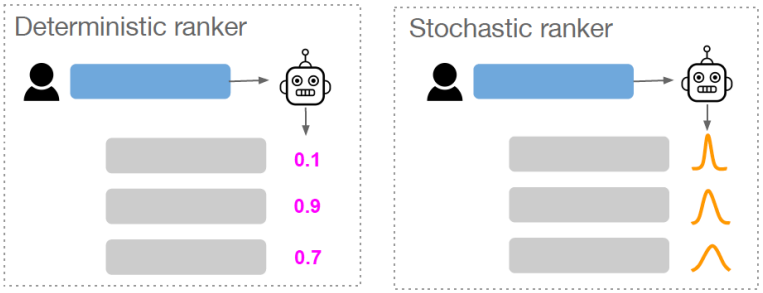


Figure 1: While deterministic neural rankers output a point estimate probability (magenta values) of relevance for a combination of query (blue bars) and document (grey bars), stochastic neural rankers output a predictive distribution (orange curves). The dispersion of the predictive distribution provides an estimation of the model uncertainty.

# Deep Ensembles(**S-BERT$^{\mathbf{E}}$**)

We train $M$ models using different random seeds and make predictions with each one of them to generate M predicted values:

$$R_r^E = \{f(\mathcal{U}_i, r)^0, f(\mathcal{U}_i, r)^1, ..., f(\mathcal{U}_i, r)^M\} \tag{3}$$

The mean of the predicted values is used as the predicted probability of relevance:

$$\textbf{S-BERT}^{\mathbf{E}}(\mathcal{U}_i, r) = E[R_r^E] \tag{4}$$

The variance $var[R_r^E]$ gives us a measure of the uncertainty in the prediction.

# MC Dropout(**S-BERT$^{\mathbf{D}}$**)

We train a single model and employ dropout at test time and generate stochastic predictions of relevance by conducting T forward passes:

$$R_r^D = \{f(\mathcal{U}_i, r)^0, f(\mathcal{U}_i, r)^1, ..., f(\mathcal{U}_i, r)^T\} \tag{5}$$

The mean of the predicted values is used as the predicted probability of relevance:

$$\textbf{S-BERT}^{\mathbf{D}}(\mathcal{U}_i, r) = E[R_r^D] \tag{6}$$

The variance $var[R_r^D]$ gives us a measure of the uncertainty in the prediction.

# Risk-Aware RA-BERT Ranker

Given the predictive distribution $R_r$, obtained either by **Ensemble** or **Dropout**, we use the following function to rank responses with riskawareness:

$$\textbf{RA-BERT}(\mathcal{U}_i, r) = E[R_r] - b * var[R_r] - 2b \sum_{i}^{n-1} cov[R_r, R_{r_i}] \quad (7)$$

where b is a hyperparameter that controls the aversion or predilection towards risk.

- **RA-BERT$^{\mathbf{D}}$** when using **S-BERT$^{\mathbf{D}}$**'s predictive distribution
- **RA-BERT$^{\mathbf{E}}$** when using **S-BERT$^{\mathbf{D}}$**'s predictive distribution

# Robustness to Distributional Shift

In order to evaluate whether we can trust the model's calibration and uncertainty estimates, we evaluate how robust the models are to different types of shift in the test data.
For all three research questions, we test the models under the following two settings to measure whether the model *"know what it knows"*:

1. Cross Domain
2. Cross Negative Sampling: test models on negative documents that were sampled using a different NS strategy than during training.
   - $\textbf{NS}_{random}$
   - $\textbf{NS}_{classic}$
   - $\textbf{NS}_{sentenceEmb}$

# Table of Contents


1. Motivation
2. Background and Related Work
3. Model
4. Experiments
5. Conclusion

# Experimental Setup

1. Dataset
   - MSDialog (246k context-response pairs, from the MS QA forum for several Microsoft products)
   - MANTis (1.3 million context-response pairs, from 14 Stack Exchange sites, such as *askubuntu* and *travel*
   - $UDC_{DSTC8}$ (184k context-response pairs of disentangled Ubuntu IRC dialogues)
2. Implementation Details
   - Training BERT using sample method of BM25
   - $\textbf{NS}_{sentenceEmb}$ use **sentenceBERT** (Reimers and Gurevych, 2019)
   - The baseline BERT-based ranker setup yields comparable effectiveness with SOTA methods
3. Evaluation
   - effectiveness: recall at positiion K with n candidates: $R_n@K$
   - evaluate quality of the uncertainty estimation with two downstream tasks:
     - improve conversation response ranking itself via Risk-Aware ranking
     - predict unanswerable conversational contexts

# Results: Calibration of Neural Rankers(RQ1)

Table 1: Calibration (ECE, lower is better) and effectiveness ($R_{10}@1$, higher is better) of BERT for conversation response ranking in cross-domain, and cross-NS conditions. All models were trained using $NS_{BM25}$. ECE is calculated using a balanced number of relevant and non relevant documents. Underlined values indicate no distributional shift ($\mathcal{D}_S = \mathcal{D}_T$ and train NS = test NS).

| | cross-domain | | | | | | cross-NS | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Test on → | MANTiS | | MSDialog | | $UDC_{DSTC8}$ | | $NS_{random}$ | | $NS_{sentenceBERT}$ | |
| Train on ↓ ($NS_{BM25}$) | $R_{10}@1$ | ECE | $R_{10}@1$ | ECE | $R_{10}@1$ | ECE | $R_{10}@1$ | ECE | $R_{10}@1$ | ECE |
| MANTiS | 0.615 | 0.003 | 0.653 | 0.010 | 0.422 | 0.028 | 0.263 | 0.011 | 0.310 | 0.009 |
| MSDialog | 0.398 | 0.009 | 0.652 | 0.006 | 0.495 | 0.014 | 0.298 | 0.029 | 0.239 | 0.027 |
| $UDC_{DSTC8}$ | 0.349 | 0.016 | 0.306 | 0.023 | 0.834 | 0.002 | 0.318 | 0.050 | 0.182 | 0.045 |

- **BERT** is both effective and calibrated under no distributional shift conditions. However, the calibration error increases significantly in cross-domain and cross-NS settings, **indicating that they do not have robust calibrated predictions**.

# Results: Calibration of Neural Rankers(RQ1)

Table 2: Relative decreases of ECE (lower is better) of S-BERT$^E$ and S-BERT$^D$ over BERT. Superscript $^\dagger$ denote significant improvements (95% confidence interval) using Student's t-tests.

| | cross-domain | | | | | | cross-NS | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Test on → | MANTiS | | MSDialog | | $\text{UDC}_{\text{DSTC8}}$ | | $\text{NS}_{\text{random}}$ | | $\text{NS}_{\text{sentenceBERT}}$ | |
| Train on ↓ ($\text{NS}_{\text{BM25}}$) | S-BERT$^E$ | S-BERT$^D$ | S-BERT$^E$ | S-BERT$^D$ | S-BERT$^E$ | S-BERT$^D$ | S-BERT$^E$ | S-BERT$^D$ | S-BERT$^E$ | S-BERT$^D$ |
| MANTiS | -35.13%$^\dagger$ | -56.14%$^\dagger$ | -03.42% | -26.89%$^\dagger$ | -04.94% | -00.83% | -31.35% | -18.65%$^\dagger$ | -37.65%$^\dagger$ | -02.79% |
| MSDialog | +25.05% | +08.27% | -43.11% | -11.54% | +22.77% | +05.85% | -15.91% | -10.58% | -17.17% | -12.93% |
| $\text{UDC}_{\text{DSTC8}}$ | -54.95%$^\dagger$ | -09.98%$^\dagger$ | -25.78%$^\dagger$ | -09.15% | +24.77% | -01.84% | -08.05% | -01.78% | -04.81% | -01.28% |

- **stochastic BERT** displays the improvements (relative drop in ECE) over **BERT** in terms of calibration.
- **S-BERT$^E$** is on average **14%** better than **BERT**, while **S-BERT$^D$** is on average **10%** better than **BERT**.
- **answering our RQ1: stochastic BERT-based rankers have better calibration than deterministic BERT-based ranker**

# Results: Uncertainty Estimates for Risk-Aware Neural Ranking(RQ2)

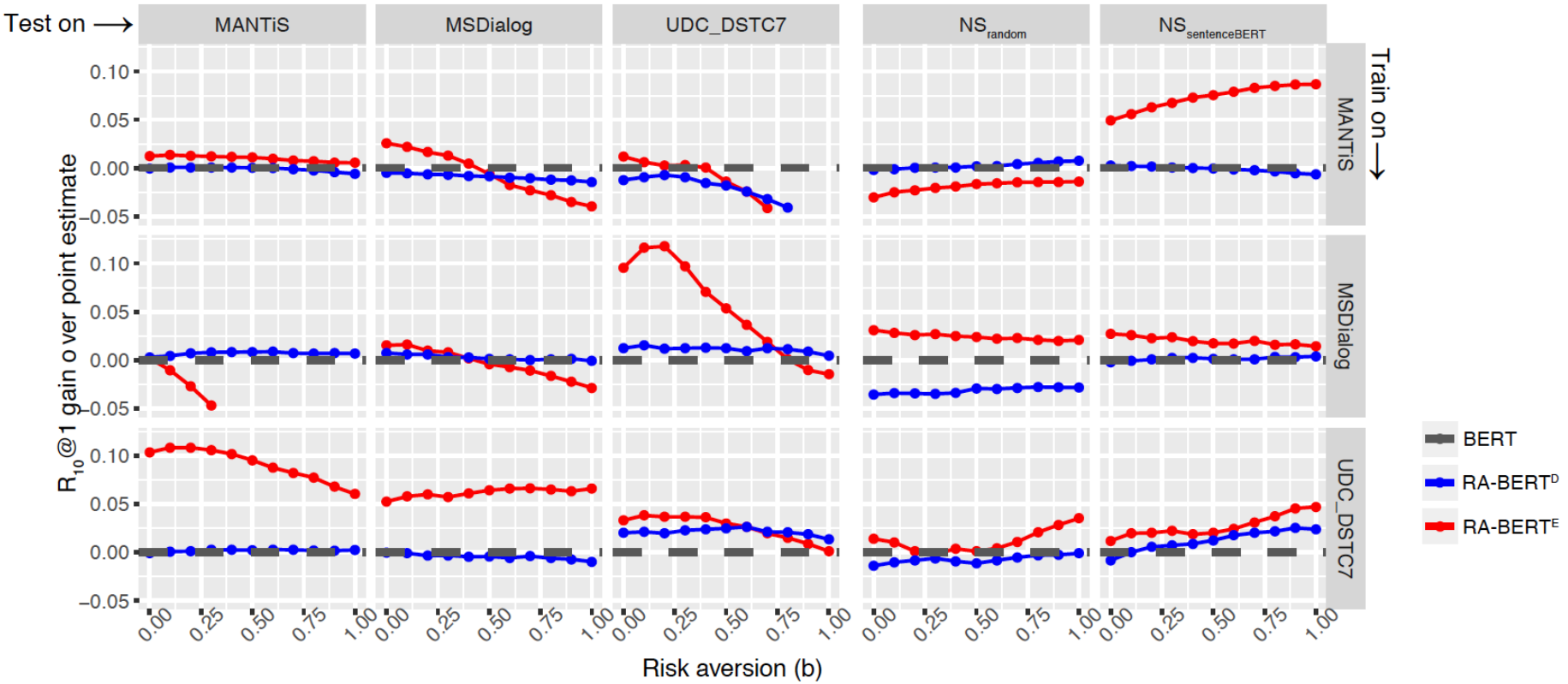


Figure 3: Gains of the Risk-Aware BERT-ranker for different values of risk aversion $b$.

# Results: Uncertainty Estimates for Risk-Aware Neural Ranking(RQ2)

- When $b = 0$, we are using the mean of the predictive distribution and disregard the risk, which is relevant to **S-BERT**
- When $b < 0$, we are ranking with risk predilection. In all conditions the effectiveness was significantly worse than when $b = 0$
- When increasing the risk aversion ($b > 0$), it has different effects depending on the combination of domain and NS.

# Results: Uncertainty Estimates for Risk-Aware Neural Ranking(RQ2)

In order to investigate whether ranking with risk aversion is more effective than using the predictive distribution mean, we select $b$ based on the best value observed on the validation set.

Table 3: Relative improvements (higher is better) of $R_{10}$@1 of RA-BERT$^E$ and RA-BERT$^D$ over the mean of stochastic BERT predictions (S-BERT$^E$ and S-BERT$^D$). Superscript $^\dagger$ denote statistically significant improvements over the S-BERT ranker at 95% confidence interval using Student's t-tests.

| | cross-domain | | | | | | cross-NS | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Test on → | MANTiS | | MSDialog | | $\text{UDC}_{\text{DSTC8}}$ | | $\text{NS}_{\text{random}}$ | | $\text{NS}_{\text{sentenceBERT}}$ | |
| Train on ↓ ($\text{NS}_{\text{BM25}}$) | RA-BERT$^E$ | RA-BERT$^D$ | RA-BERT$^E$ | RA-BERT$^D$ | RA-BERT$^E$ | RA-BERT$^D$ | RA-BERT$^E$ | RA-BERT$^D$ | RA-BERT$^E$ | RA-BERT$^D$ |
| MANTiS | -0.14% | +0.16%$^\dagger$ | +0.00% | +0.00% | +0.00% | +0.00% | +4.73%$^\dagger$ | +4.58%$^\dagger$ | +9.68%$^\dagger$ | -2.68% |
| MSDialog | -2.74% | +0.39% | -1.05% | -0.66% | +5.08%$^\dagger$ | -0.10% | -7.61% | +3.29% | -0.61% | +0.63% |
| $\text{UDC}_{\text{DSTC8}}$ | +0.00% | +0.00% | +0.00% | +0.00% | +0.42% | -0.06% | +6.32%$^\dagger$ | +3.83%$^\dagger$ | +16.39%$^\dagger$ | +17.18%$^\dagger$ |

**This answers our RQ2, indicating that the uncertainties obtained from stochastic neural rankers are useful for risk-aware ranking, specially in the cross-NS setting where the baseline model is quite ineffective.**

# Results: Uncertainty Estimates for NOTA prediction(RQ3)

Table 4: Results of the *cross-domain* condition for the NOTA prediction task, using a Random Forest classifier and different input spaces. The F1-Macro and standard deviation over the 5 folds of the cross validation are displayed. Superscript † denote statistically significant improvements over $E[R^D]$ at 95% confidence interval using Student's t-tests. Bold indicates the most effective approach.

| | cross-domain | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Test on → | MANTiS | | | MSDialog | | | $\text{UDC}_{\text{DSTC8}}$ | | |
| Train on ↓ ($\text{NS}_{\text{BM25}}$) | $E[R^D]$ | $+var[R^E]$ | $+var[R^D]$ | $E[R^D]$ | $+var[R^E]$ | $+var[R^D]$ | $E[R^D]$ | $+var[R^E]$ | $+var[R^D]$ |
| MANTiS | 0.635 (.02) | 0.686 (.01)† | **0.792 (.02)**† | 0.669 (.03) | 0.731 (.04) | **0.855 (.02)**† | 0.571 (.04) | 0.590 (.08)† | **0.621 (.04)**† |
| MSDialog | 0.561 (.02) | 0.598 (.02)† | **0.633 (.02)**† | 0.662 (.04) | **0.702 (.01)**† | 0.699 (.06)† | 0.596 (.04) | 0.566 (.06)† | **0.655 (.06)**† |
| $\text{UDC}_{\text{DSTC8}}$ | 0.527 (.04) | 0.665 (.02)† | **0.738 (.03)**† | 0.523 (.05) | 0.691 (.03)† | **0.757 (.04)**† | 0.787 (.01) | **0.829 (.03)**† | 0.807 (.01)† |

Table 5: Results of the *cross-NS* condition for the NOTA prediction task.

| | cross-NS | | | | | |
|---|---|---|---|---|---|---|
| Test on → | $\text{NS}_{\text{random}}$ | | | $\text{NS}_{\text{sentenceBERT}}$ | | |
| Train on ↓ ($\text{NS}_{\text{BM25}}$) | $E[R^D]$ | $+var[R^E]$ | $+var[R^D]$ | $E[R^D]$ | $+var[R^E]$ | $+var[R^D]$ |
| MANTiS | 0.557 (.01) | 0.604 (.02)† | **0.698 (.02)**† | 0.534 (.03) | 0.587 (.02)† | **0.647 (.05)**† |
| MSDialog | 0.505 (.02) | 0.606 (.02)† | **0.702 (.05)**† | 0.522 (.03) | 0.611 (.07)† | **0.653 (.04)**† |
| $\text{UDC}_{\text{DSTC8}}$ | 0.565 (.03) | 0.800 (.02)† | **0.942 (.04)**† | 0.506 (.05) | 0.755 (.05)† | **0.821 (.05)**† |

# Results: Uncertainty Estimates for NOTA prediction(RQ3)

The uncertainties from **S-BERT**$^{D}$ and **S-BERT**$^{E}$ significantly improve the F1 for NOTA prediction for both cross-domain and cross-NS settings

- cross-domain: improvement of 24% on average
- cross-NS: improvement of 46% on average

**This answers our RQ3 that the uncertainty estimates from stochastic neural rankers do improve the effectiveness of the NOTA prediction task (by an average of 33% for all conditions considered)**

# Table of Contents


1. Motivation
2. Background and Related Work
3. Model
4. Experiments
5. Conclusion

# Conclusion

1. Show that deterministic BERT-based ranker is not robustly calibrated for the task of conversation response ranking
2. Improve BERT-based ranker with two techniques to estimate uncertainty through *stochastic neural ranking*
3. Benefits of estimating uncertainty using risk-aware neural ranking and for predicting unanswerable conversational contexts