

# On the Calibration and Uncertainty with Pólya-Gamma Augmentation for Dialog Retrieval Models

Tong Ye<sup>1,2</sup>, Shijing Si<sup>1</sup>, Jianzong Wang<sup>1, \*</sup>, Ning Cheng<sup>1</sup>, Zhitao Li<sup>1</sup>, Jing Xiao<sup>1</sup>

<sup>1</sup> Ping An Technology (Shenzhen) Co., Ltd.

<sup>2</sup> University of Science and Technology of China  
jzwang@188.com

## Abstract

Deep neural retrieval models have amply demonstrated their power but estimating the reliability of their predictions remains challenging. Most dialog response retrieval models output a single score for a response on how relevant it is to a given question. However, the bad calibration of deep neural network results in various uncertainty for the single score such that the unreliable predictions always misinform user decisions. To investigate these issues, we present an efficient calibration and uncertainty estimation framework PG-DRR for dialog response retrieval models which adds a Gaussian Process layer to a deterministic deep neural network and recovers conjugacy for tractable posterior inference by Pólya-Gamma augmentation. Finally, PG-DRR achieves the lowest empirical calibration error (ECE) in the in-domain datasets and the distributional shift task while keeping R10@1 and MAP performance.

## Introduction

Dialog response retrieval models based on deep neural networks have shown impressive results on multiple benchmarks (Gu et al. 2020; Lu et al. 2020; Whang et al. 2021). However, the predictions from these models always fail to provide appropriate answers when deploying into real-world applications. For example, popular dialog agents always show users with incorrect predictions if questions fall outside of the training distribution, which could mislead their decisions. Therefore, an ideal model should abstain when they are likely to be error. The simplest solution is to provide the corresponding confidence estimation so that predictions with low confidence can be abstained. This problem is also defined as **Model Calibration**: making sure the confidence of the prediction is well correlated with the actual probabilities of correctness.

Generally, retrieval-based dialog models consider their estimation of a response’s relevance as a deterministic score, which can be subject to over confidence issues, i.e. are badly calibrated (Guo et al. 2017). Existing works usually quantify this uncertainty over predictions through a distribution of possible scores to achieve calibration (Cohen et al. 2021). Specifically, the mean of the distribution represents the model’s pre-

diction while its corresponding variance captures the model’s uncertainty. Consequently, a high variance could imply that the model is not sure of the prediction and should abstain, even if it is rated among the top hits.

Two of the principled methods to calculate the predicted uncertainty for a dialog model are Deep Ensemble and Bayesian methods. Bayesian methods (Cohen et al. 2021) place a prior distribution over model parameters and Deep Ensemble (Penha and Hauff 2021) usually trains independently multiple models. They are challenging to implement at the industrial scale due to their high inference cost and huge memory requirements. This inspires us to investigate principled approaches that only need a determinate deep neural network for high quality uncertainty estimation.

Gaussian processes (GPs) (Rasmussen 2003) are flexible models that perform well in varieties of tasks. Different from existing works, GPs belong to non-parametric Bayesian approaches, which only need to learn a few hyperparameters. When combined with Gaussian likelihoods, GPs obtain closed form expressions for the predictive and posterior distributions (Snell and Zemel 2021), which alleviate the computational defects of Bayesian methods with cubic scaled examples. Moreover, GPs are easily combined with a single deep neural network without training independently multiple models. However, the GPs are challenging to scale to large datasets for classification, this is partially due to the fact that the target variable’s categorical distribution leads to a non-Gaussian posterior and we can not obtain the marginal likelihood in a closed form. An especially intriguing family of methods is adding additional Pólya-Gamma variables to the GPs model (Polson, Scott, and Windle 2013) to recover it when the original model is marginalized out.

In this study, we are committed to investigating a simple and efficient approach PG-DRR by combining Pólya-Gamma Augmentation for calibrating Dialog Response Retrieval models. Specifically, we add a neural Gaussian process layer to a deterministic deep neural network to achieve better calibration. Importantly, we use the Pólya-Gamma (PG) augmentation to recover conjugacy for tractable posterior inference and use Gibbs sampling to collect samples from the posterior in order to improve the parameters of the mean and covariance functions. Besides, we theoretically verified why PG-DRR can be calibrated. The significant contributions of this paper are as follows:

\*Corresponding author: Jianzong Wang.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- We propose an efficient framework PG-DRR for a deterministic dialog response ranking model to estimate uncertainty. And we yield the lowest ECE in two in-domain datasets and the distributional shift task while keeping  $R_{10}@1$  and MAP performance.
- We innovatively estimate uncertainty in dialog retrieval tasks with a Gaussian Process layer with the Pólya-Gamma augmentation. In addition, we theoretically analyze that PG-DRR can achieve calibration.
- We conduct extensive experiments to verify that PG-DRR significantly calibrates well while maintaining performance. Besides, ablation study analyzes the relative contributions of the kernel function and the model architecture of PG-DRR to the effectiveness improvement.

## Related Works

### Calibration in Dialog Retrieval

Most recent approaches for ranking tasks in dialog system have focused primarily on discriminative methods using neural networks that learn a similarity function to compare questions and candidate answers, which has been shown that DNNs result in high calibration errors. Therefore, there has been growing research interest in quantifying predictive uncertainty in deep dialog retrieval networks. (Zhu et al. 2009) first investigates the retrieval uncertainty and considers the probabilistic language model’s variance is a risk factor to significantly boost performance. However, the majority of recent studies on dialog uncertainty put more emphasis on the dialog management aspect (Tegho, Budzianowski, and Gašić 2017; Gašić and Young 2013; Roy, Pineau, and Thrun 2000; van Niekerc et al. 2020) and few approaches have directly incorporated a retrieval model’s uncertainty.

Deep Ensemble and Monte Carlo (MC) Dropout have emerged as two of the most prominent and practical uncertainty estimation methods for dialog retrieval networks. (Feng et al. 2020) covers the value of identifying uncertainty in end-to-end dialog jobs and using Dropout to identify questions that cannot be answered. (Penha and Hauff 2021) investigates the effectiveness of MC Dropout and Deep Ensemble for calibration in the conversational response space under BERT. (Cohen et al. 2021) suggests an effective Bayesian framework using a stochastic process to convey model’s confidence. Unfortunately, as retrieval models expand in complexity and size, MC Dropout and Deep Ensemble become computationally expensive, posing a significant challenge given the prevalence of the pre-trained architectures like BERT (Durasov et al. 2021). In addition, (Pei, Wang, and Szarvas 2022) constructs a stochastic self-attention mechanism to capture uncertainty, which means that it needs retraining the models for the different downstream tasks. Therefore, it is urgent to explore an efficient but straightforward technique to quantify the uncertainty in the determinate neural retrieval models.

### Pólya-Gamma Augmentation

It is obvious that the classification likelihood is non-Gaussian. Sequentially, the predictive distributions are also not Gaussian anymore and its closed-form solution is not available.

There are several methods offered to overcome this limitation. The classic methods mainly consist of least squares classification (Rifkin and Klautau 2004), Laplace approximation (Williams and Barber 1998), Variational approaches (de G. Matthews et al. 2016) and expectation propagation (Minka 2001). For a thorough introduction of GPs, we refer readers to (Rasmussen 2003).

The Pólya-Gamma augmentation approach (Polson, Scott, and Windle 2013), which introduces auxiliary random variables to recover it when the original model is marginalized out, can model the discrete likelihood in Gaussian Processes and has attracted extensive attention. (Girolami and Rogers 2006) investigates a Gaussian augmentation for an accurate Bayesian examination of multinomial probit regression models. According to (Linderman, Johnson, and Adams 2015), a logistic stick-breaking representation and Pólya-Gamma augmentation are used to convert a multinomial distribution into a product of binomials. (Wenzel et al. 2019) represents a scalable stochastic variational method for the Gaussian process classification based on the Pólya-Gamma data augmentation. (Galy-Fajou et al. 2019) introduces a logistic-softmax likelihood for multi-class classification and employs Gamma, Poisson and Pólya-Gamma augmentation in order to obtain a conditionally conjugate model. In addition, (Snell and Zemel 2021) combines Pólya-Gamma augmentation with the one-vs-each softmax approximation in a novel way and presents a Gaussian process classifier, which demonstrated well calibration. (Achituv et al. 2021) proposes a GP-Tree framework for multi-class classification based on Pólya-Gamma augmentation and allows the posterior inference via the variational inference approach or the Gibbs sampling technique. Therefore, we wish to leverage the conjugacy of the Pólya-Gamma augmentation to yield better calibrated and more accurate models in dialog tasks.

## Methodology

In this section, we first recall key properties of Pólya-Gamma distributions before deriving the augmented model likelihood. We next introduce our method PG-DRR for PG-based Gaussian process classification in detail. All the framework is visually represent in Fig 1.

### Preliminaries

The Pólya-Gamma augmentation method was developed to overcome the Bayesian inference issue in logistic models. (Polson, Scott, and Windle 2013). The Pólya-Gamma distribution  $w \sim PG(b, c)$  with parameters  $b > 0$  and  $c \in R$  can be expressed as a form of Gamma distributions that are infinitely convolved:

$$w \stackrel{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{\text{Gamma}(b, 1)}{(k - 1/2)^2 + c^2/(4\pi^2)}, \quad (1)$$

where  $\text{Gamma}(b, 1)$  indicates Gamma distributions and  $\stackrel{D}{=}$  stands for equality of distribution. When  $w$  follows the Pólya-Gamma distribution  $w \sim PG(b, 0)$ , the fundamental integral

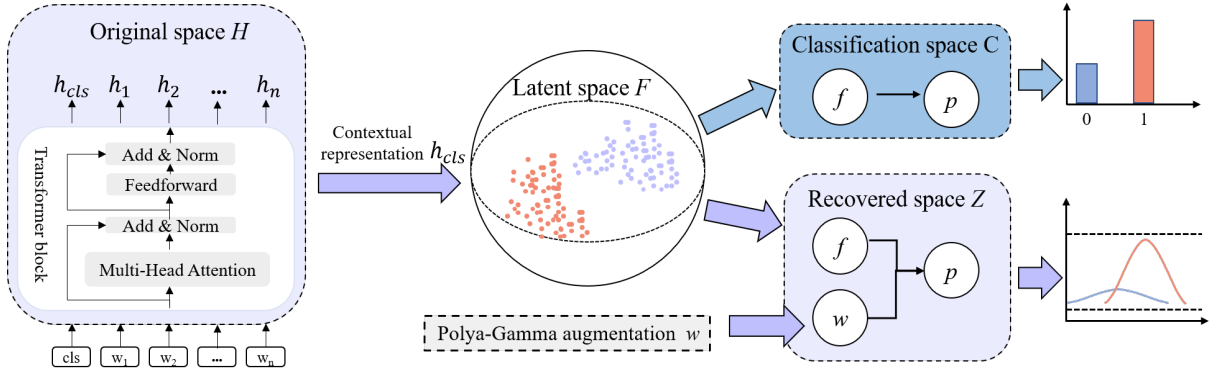


Figure 1: An illustration of PG-DRR prediction models for dialog response retrieval.

identity holds for  $b > 0$ :

$$\frac{(e^\psi)^a}{(1 + e^\psi)^b} = 2^{-b} e^{\lambda\psi} E_w[e^{-w\psi^2/2}], \quad (2)$$

where  $\lambda = a - b/2$ .

For the binary classification problems  $y \in \{0, 1\}^N$ , a vector of logits is  $\psi \in R^N$ . When  $a = y$ ,  $b = 1$  and  $\lambda = y - 1/2$  in the Eq. 2, the likelihood can be written as:

$$p(y|\psi) = \sigma(\psi)^y (1 - \sigma(\psi))^{1-y} = \frac{(e^\psi)^y}{1 + e^\psi} \quad (3)$$

$$= \frac{1}{2} e^{\lambda\psi} \int_0^\infty e^{-w\psi^2/2} p(w) dw,$$

where the  $\sigma(\cdot)$  denotes the sigmoid function. The Eq. 3 will be augmented with  $w$  such that marginlizing can be used to recover original likelihood.

### Contextual Encoder

Given a training set denoted as a triples  $D = \{(U_i, A_i, y_i)\}_{i=1}^M$ ,  $U_i = \{u_1, u_2, \dots, u_t\}$  is a dialog context consisting of  $t$  utterances and  $y_i$  is the response relevance labels  $y_i \in \{0, 1\}$ . Answer candidates is denoted as  $A_i = \{a_1, a_2, \dots, a_r\}$  including  $r$  responses. The input sentence is prepared as  $x_i = \{[CLS], u_1, \dots, u_t, [SEP], a_r\}$ . The special token  $[CLS]$  denotes the beginning of the sequence and  $[SEP]$  denotes the separate the response from the contexts. Following BERT (Devlin et al. 2019), we split the input into words by the same WordPiece tokenizer (Wu et al. 2016) and embeddings for sub-word was obtained by summing up position embedding, word embedding and segment embedding. Subsequently, BERT is utilized to learn the contextual representation  $f(u_i, a_i)$  based on the  $[CLS]$  token in a pointwise manner.

### A Gaussian Process Layer

The Gaussian process is made up of random variables and every finite subset would follow a multivariate normal distribution. We can denote a Gaussian process as  $g(x) \sim GP(m(x), K(x, x'))$ , where  $m(x) = E(g(x))$  and  $K(x, x') = Cov[g(x), g(x')]$  respectively represents

the mean function and the covariance kernel function (Rasmussen 2003).

Supposing the output layer  $g : F \rightarrow Z$ , PG-DRR replaces the typical dense output layer with a Gaussian process (GP). Specifically, according to the  $M$  training samples  $D = \{(x_i, y_i)\}_{i=1}^M$ , the samples will pass the contextual encoder and obtain a latent representation  $h_i = h_\theta(x_i)$ , where  $\theta$  is the weight of the contextual encoder. The Gaussian process output layer  $g = [g(h_1), g(h_2), \dots, g(h_N)]$  follows a multivariate normal distribution:  $g: g \sim MVN(0, K(h, h'))$ . We set radial basis function (RBF) as the kernel function in this paper. Therefore, the formal definition of a GP layer is denoted as follows:

$$g \sim MVN(0, K), K_{i,j} = \sigma^2 \exp(-\|h_i - h_j\|_2^2 / 2l^2) \quad (4)$$

where  $l$  and  $\sigma$  are the length and kernel scale parameters (Jankowiak and Pleiss 2021), respectively.

Therefore, the joint likelihood of the latent GP and the label are as follows:

$$p(y|x, g) = \prod_i p(y_i|x_i, g_i) \quad (5)$$

For a latent function  $g$ , the posterior distribution  $p(g|x, y)$  may be expressed as follows:

$$p(g|x, y) = \frac{p(g)p(y|x, g)}{p(y|x)} \quad (6)$$

Since the likelihood probability is non-Gaussian, the posterior  $p(g|x, y)$  can not be computed analytically. Following (Wenzel et al. 2019), we build an approximate gradient estimator upon Fisher's identity. We denote  $w$  as the Pólya-Gamma random variables. The full classification's marginal likelihood term is as follows:

$$p(y_i|x_i) = \int p_\theta(y_i|x_i, w_i) p(w_i) dw \quad (7)$$

According to the Pólya-Gamma augmentation and Eq. 3,  $p_\theta(y_i|x_i, w_i)$  follows a Gaussian distribution in terms of proportion.

$$p_\theta(y_i|x_i, w_i) \propto e^{-w_i g_i^2 / 2} e^{(y_i - 1/2) g_i} \quad (8)$$

$$\propto N(diag(w)^{-1} (y_i - 1/2) | g, diag(w)^{-1})$$

Consequently, the conditional likelihood  $p(y|g, w)$  and the Gaussian prior  $p(g)$  are conjugate, which leads to the posterior  $p(g|y, w)$  also following a Gaussian distribution. In order to generate samples from this distribution ( $g_{(t)}, w_{(t)}$ ), we can readily apply Gibbs sampling (Douc, Moulines, and Stoffer 2014) and calculate the conditional posteriors  $p(g|y, w)$  and  $p(w|g)$ . According to (Wenzel et al. 2019), When the auxiliary latent variables  $w$  are taken into account, the posterior over the latent function values  $g$  is calculated as follows:

$$p(w|y, g) = PG(1, g), \quad (9)$$

$$p(g|y, w) = N(g|(K^{-1}\mu + k)\Sigma, \Sigma), \quad (10)$$

$$\Sigma = (K^{-1} + \text{diag}(w))^{-1} \quad (11)$$

## Training Loss

The following is a possible way to express the log marginal likelihood:

$$\begin{aligned} L &= \log p_\theta(y|x) = \log \int p(w)p_\theta(y|x, w) dw \\ &= \log \int p(w) \int p_\theta(g|x)p(g|y, w) dg dw \end{aligned} \quad (12)$$

By utilizing posterior samples  $w \sim p_\theta(w|X, Y)$ , it is possible to measure the gradient of the log marginal likelihood. The samples of  $w$  from Gibbs chains serve as the foundation for the stochastic training target in actual practice. The following describes the gradient estimator upon Fisher's identity.

$$\begin{aligned} \nabla_\theta L &= \int p_\theta(w|x, y) \nabla_\theta \log p_\theta(y|x, w) dw \\ &\approx \frac{1}{N} \sum_{n=1}^N \nabla_\theta \log p_\theta(y|w_n, x) \end{aligned} \quad (13)$$

where  $w_1, w_2, \dots, w_N$  represent the samples generated by the posterior Gibbs chains.

When predicting, we designate the input and label with  $x^*$  and  $y^*$ , respectively.

$$\begin{aligned} p(g^*|x, y, w, x^*) &= N(g^*|\mu^*, \Sigma^*), \text{ and} \\ \mu^* &= (K^*)^T (K + \text{diag}(w)^{-1})^{-1} \text{diag}(w)^{-1} k, \\ \Sigma^* &= K^{**} - (K^*)^T (K + \text{diag}(w)^{-1})^{-1} K^* \end{aligned} \quad (14)$$

$$p(y^*|x, y, w, x^*) = \int p(g^*|x, y, w, x^*) p(y^*|g^*) dg^* \quad (15)$$

$K^{**}$  indicates the test point's kernel value and  $K^* = K(x, x^*)$ . We use 1D Gaussian-Hermite quadrature to calculate the intractable integral in Eq.15.

## Theoretical Analysis

Existing works tend to let the weights Bayesian to capture uncertainty. GPs belong to non-parametric Bayesian approaches and only need to learn a few hyperparameters. Practical interest lies in the following question: Is PG-DRR still Bayesian enough to correct overconfidence? No surprisingly, the answer is yes. According to (Hein, Andriushchenko, and Bitterwolf 2019), when the test data is far away from the training

data, ReLU networks exhibit arbitrarily high confidence. In order to achieve our goal, we just need to show that as the gap between test and training data grows, the model prediction approaches zero.

Following (Hein, Andriushchenko, and Bitterwolf 2019), let  $g : R^d \rightarrow R$  be a binary Gaussian process classifier defined by  $g(h(x))$ , where  $h : R^n \rightarrow R^d$  is a fixed ReLU network denoted as  $h(x) = W^T x + b$ . Let  $\sigma(\cdot)$  be the sigmoid function and  $N(g|\mu, \Sigma)$  be the Gaussian approximation of the last layer's outputs with eigenvalues of  $\Sigma^*$  as  $\lambda_1 \leq \dots \leq \lambda_r$ . Then for any input  $x^* \in R^n$ ,

$$\begin{aligned} p(y^*|x, y, w, x^*) &= \int p(g^*|x, y, w, x^*) p(y^*|g^*) dg^* \\ &\approx \int \Phi(\sqrt{\pi/8}g^*) N(g^*|\mu^*, \Sigma^*) dg^* \\ &= \Phi\left(\frac{\mu^*}{\sqrt{8/\pi + \Sigma^*}}\right) \approx \sigma(z(x^*)) \end{aligned} \quad (16)$$

where  $\Phi$  is denoted as the standard Gaussian distribution function and  $\Phi(\sqrt{\pi/8}g^*) \approx \sigma(g^*)$ . Following the theorem of (Kristiadi, Hein, and Hennig 2020), when  $\delta > 0$  and as  $\delta \rightarrow \infty$ , we found:

$$\begin{aligned} \lim_{\delta \rightarrow \infty} \sigma(|z(\delta x^*)|) &= \lim_{\delta \rightarrow \infty} \Phi\left(\frac{\delta \mu^*}{\sqrt{8/\pi + \delta^2 \Sigma^*}}\right) \\ &= \lim_{\delta \rightarrow \infty} \sigma\left(\frac{\delta \mu^*}{\sqrt{1 + \delta^2 \pi/8 \Sigma^*}}\right) = \sigma\left(\frac{\|\mu^*\|}{\sqrt{\pi/8 \Sigma^*}}\right) \end{aligned} \quad (17)$$

The value goes to a quantity that only depends on the covariance and mean of the Gaussian process. This finding suggests that by manipulating the Gaussian, it is possible to move the confidence closer to the uniform further from the training locations. We can conclude that The PG-DRR is Bayesian enough to calibrate models and collect uncertainty information.

## Learning Algorithm

Algorithm 1 summarizes our learning algorithm for marginal likelihood.

---

### Algorithm 1: PG-DRR Learning

---

**Input:** Training data  $D$ , the weight of networks  $\theta$ , learning rate  $\eta$ , number of steps  $T$ , number of parallel Gibbs chains  $N$

- 1: Initialization parameters  $\theta$  randomly
  - 2: **repeat**
  - 3:   batch samples  $(X, Y) \sim D$
  - 4:   **for**  $n = 1$  to  $N$  **do**
  - 5:      $w_n^0 \sim PG(1, 0)$ ,  $g_n^0 \sim p_\theta(g|X)$
  - 6:     **for**  $t = 1$  to  $T$  **do**
  - 7:        $w_n^t \sim PG(1, g_n^{t-1})$
  - 8:        $g_n^t \sim p_\theta(g|X, Y, w_n^t)$
  - 9:     **end for**
  - 10:   **end for**
  - 11:   Update  $\theta \leftarrow \theta + \frac{\eta}{N} \sum_{n=1}^N \nabla_\theta \log p_\theta(Y|X, w_n^T)$
  - 12: **until** convergence
-

## Experiment

In this section, we first introduce the datasets and the evaluation metrics which we use to quantitatively compare PG-DRR against MC Dropout and Deep Ensemble. We then compare our method with both in-domain and cross-domain datasets to verify the calibration, efficiency and robustness of PG-DRR. Finally, we discuss the computational and memory complexity.

### Dataset

We utilize two large-scale conversational response ranking datasets for our in-domain experiments:

- **MS Dialog**(Qu et al. 2018), crawled from over 35K information seeking conversations from the Microsoft Answer community, contains 246,000 context-response pairs. For each utterance, it is accompanied with a list of 10 potential responses, only one of which is accepted as the right response.
- **MANtIS**(Penha, Balan, and Hauff 2019): based on the Stack Exchange community question-answering website and has 1.3 million context-response pairs with 14 distinct domains. For each context, We create 10 negative sampled instances by replacing the provider’s most recent statement with a negative sample drawn from BM25 using the correct response as the query.

Besides using the uncertainty estimation for in-domain datasets, we also train the model using the training set from one dataset, i.e. train set and predict it on the test set of a different dataset, which is also know as domain generalization or distributional shift tasks.

- **MS Dialog**→**MANtIS** This distributional shift scenario takes the training set of MS Dialog and to validate and test on the MANtIS.
- **MANtIS**→**MS Dialog** This distributional shift scenario takes the training set of MANtIS and to validate and test on the MS Dialog.

### Metrics

**Recall@1:** For retrieval performance evaluation, we employ the recall at 1 out of 10 candidates named  $R_{10}@1$ , which consists of 1 ground-truth response and 9 candidates randomly selected from the test set.

**Mean Average Precision (MAP):** To evaluate average precision across multiple queries, we use the MAP. Simply put, it is the mean of the accuracy average over all queries.

**Empirical Calibration Error (ECE):** To evaluate the calibration efficiency of dialog retrieval tasks, we utilize the Empirical Calibration Error (ECE)(Naeni, Cooper, and Hauskrecht 2015). It can assess the relationship between expected probability and accuracy. We partition the range  $[0, 1]$  into  $C$  bins that are evenly spaced apart. By calculating the weighted average of the absolute difference between the accuracy  $A_i$  and confidence  $B_i$  of each bin, the ECE may be roughly calculated as:  $ECE = \sum_{i=1}^C \frac{|B_i|}{N} |A_i - \text{mean}(B_i)|$ . We set  $C = 10$ .

## Baselines

We compare several recent methods for qualifying uncertainty in conversation response ranking tasks.

**BERT**(Devlin et al. 2019): a pre-trained BERT (bert-base-cased) that fine-tuned on the dialog response retrieval task.

**MC Dropout**(Penha and Hauff 2021): a BERT-based model employing dropout at both train and test time to approximate Bayesian inference and generating a predictive distribution after 10 forward passes.

**Deep Ensemble**(Penha and Hauff 2021): integrating predictions of any 5 BERT models with dense output layers. We denote as Ensemble for simplification.

**SNGP**(Liu et al. 2020): A modified version of BERT that the output layer was replaced by a Gaussian Process layer and weighting normalization step was add to training loop.

### Implementation Details

We employ *BERT*, which has a hidden state dimension of 768 and consists of 12 Transformer blocks with 12 attention heads, as the backbone. We utilize the Adam optimizer and set dropout probability to 0.1. Following recent research (Gu et al. 2020) that employed finetuned BERT for dialog response ranking, we choose a response at random from the whole collection of responses as the negative samples when training.

In our experiments, we set the learning rates to 1e-5 for baseline methods, and personal learning rates to 3e-3 for GP layer (GP layer is not fine-tuned and therefore need bigger rates). RBF kernel function is configured to use fix length scale of 1 and 8 output scale. During training and evaluation, we both use 10 steps of Gibbs and 30 parallel Gibbs chains to reduce variance. We use the same set of hyper parameters across different model architecture for deterministic variant. In the MS Dialog dataset, we train our model with a batchsize of 16 for 3 epochs while 1 epoch for the MANtIS dataset. We used Pytorch and models were trained on 1 Tesla V100 with 16G memory <sup>1</sup>.

## Results

In this section, we present our proposed PG-DRR in terms of both performance and calibration. Our aim is to compare the calibration among a variety of methods. Therefore, we need to develop new benchmark evaluations: retrieval performance, the calibration, and the robustness. In order to achieve our goal, we could not just take the accuracy findings from previous studies, but instead, each of these baselines requires being trained from scratch.

### Performance, Calibration and Efficiency

**Performance.** We first report the performance of PG-DRR in comparison with other baselines on the MS dialog and MANtIS datasets by evaluating the metrics of  $R_{10}@1$  and MAP. The results are represented in the Table 1.

As shown, PG-DRR is competitive with that of a deterministic network BERT, and outperforms the other single-model

<sup>1</sup>Detailed experimental codes can be found at <https://github.com/pingantechnlp/PG-DRR>.

	Metric	BERT	MC Dropout	Ensemble	SNGP	PG-DRR
MS Dialog	$R_{10}@1$	0.658 $\pm$ 0.010	0.644 $\pm$ 0.009	<b>0.678<math>\pm</math>0.004</b>	0.600 $\pm$ 0.036	0.656 $\pm$ 0.046
	MAP	0.782 $\pm$ 0.007	0.776 $\pm$ 0.005	<b>0.798<math>\pm</math>0.003</b>	0.744 $\pm$ 0.022	0.784 $\pm$ 0.032
MANtIS	$R_{10}@1$	0.556 $\pm$ 0.028	0.562 $\pm$ 0.026	0.553 $\pm$ 0.012	0.545 $\pm$ 0.050	<b>0.606<math>\pm</math>0.016</b>
	MAP	0.683 $\pm$ 0.021	0.688 $\pm$ 0.019	0.681 $\pm$ 0.008	0.680 $\pm$ 0.031	<b>0.731<math>\pm</math>0.011</b>

Table 1: Average performance ( $R_{10}@1$ , MAP) and standard deviation over 5 seeds for dialog response retrieval tasks. The best results are highlighted in bold.

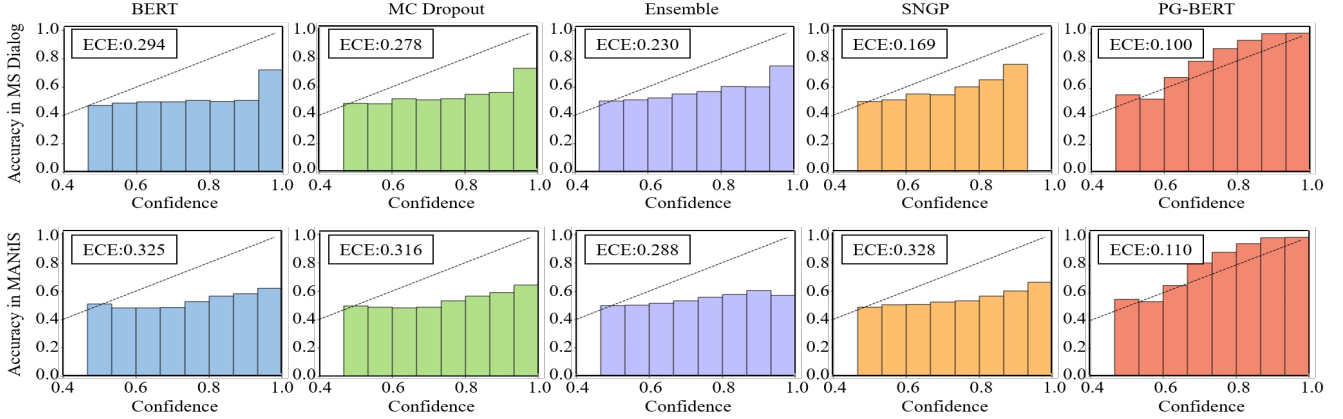


Figure 2: Reliability diagrams and expected calibration error (ECE) on the MS Dialog and MANtIS dataset.

	Parameters (M)	Inference time (ms)
BERT	108.31	11.48
MC Dropout	108.31	111.28
Ensemble	514.56	57.39
SNGP	118.87	12.82
PG-DRR	108.90	17.09

Table 2: Parameters and wall-clock time (in millisecond) of inference over test sets.

calibration approaches like MC Dropout and SNGP on the MS Dialog dataset. In addition, PG-DRR is competitive with the Ensemble, which is reduced by less than 2 %. For MANtIS, PG-DRR yields strong retrieval results, outperforming all the deterministic and calibration approaches. This indicates that PG-DRR generally achieves competitive prediction performance under in-domain conditions.

**Calibration.** We turn to the calibration for dialog retrieval tasks, which is crucial for a dialog system to delay to make a wise choice when there is insufficient information. We chose the most commonly used metric ECE for evaluating calibration. The results can be found in the Figure 2.

Obviously, The confidence of BERT, MC Dropout, Ensemble and SNGP is much higher than their accuracy, which means that they are deeply overconfident. Specifically, BERT, which is a vanilla model, usually performs well but is not well calibrated. The calibration of MC Dropout and Deep Ensemble outperform BERT, which verifies that Bayesian models will have better confidence expressiveness, but unfortunately

still obtain poor calibration. Compared to the prior methods, the ECE of PG-DRR is almost lowest, which is reduced by almost 6% and 16% respectively in two datasets, while the  $R_{10}@1$  and MAP are better or less than 1% decrease. Namely, PG-DRR includes uncertainty information while keeping  $R_{10}@1$  and MAP performance in in-domain datasets.

**Efficiency.** Computational cost is One of the most significant obstacles when employing Bayesian to capture uncertainty. We analyze the efficiency of PG-DRR on the number of parameters and the inference time in the MS Dialog dataset in Table 2. Compared to MC Dropout and Ensemble, PG-DRR has a significant decrease (at least 8 times) on inference time. While not completely free, PG-DRR only adds negligible computational cost to BERT, but it greatly improves the calibration, which facilitates adaptation to other models. In addition, non-parameters Gaussian process significantly reduces the number of parameters compared with Ensemble.

We believe that GPs belong to non-parametric Bayesian approaches and only need to learn a few hyperparameters, which enables PG-DRR improve uncertainty calibration for dialog response retrieval models. This happens to be coincident with (Kristiadi, Hein, and Hennig 2020) that a little Bayesian is enough to capture uncertainty information and correct the overconfidence even if it is only placed at the last layer of model when handling binary classification.

## Robustness and Generalization

To evaluate the model’s robustness and generalization, we consider the distributional shift task.



	MS Dialog→MANtIS			MANtIS→MS Dialog		
	$R_{10}@1\uparrow$	MAP $\uparrow$	ECE $\downarrow$	$R_{10}@1\uparrow$	MAP $\uparrow$	ECE $\downarrow$
BERT	0.378 $\pm$ 0.015	0.537 $\pm$ 0.012	0.343 $\pm$ 0.035	0.424 $\pm$ 0.047	0.599 $\pm$ 0.035	0.514 $\pm$ 0.037
MC Dropout	0.358 $\pm$ 0.032	0.523 $\pm$ 0.023	0.328 $\pm$ 0.045	0.409 $\pm$ 0.023	0.590 $\pm$ 0.020	0.497 $\pm$ 0.040
Ensemble	0.405 $\pm$ 0.007	0.558 $\pm$ 0.005	0.331 $\pm$ 0.037	<b>0.457<math>\pm</math>0.019</b>	<b>0.625<math>\pm</math>0.013</b>	0.503 $\pm$ 0.037
SNGP	0.340 $\pm$ 0.042	0.514 $\pm$ 0.032	0.307 $\pm$ 0.011	0.333 $\pm$ 0.126	0.520 $\pm$ 0.107	0.485 $\pm$ 0.017
PG-DRR	<b>0.530<math>\pm</math>0.049</b>	<b>0.673<math>\pm</math>0.038</b>	<b>0.101<math>\pm</math>0.003</b>	0.330 $\pm$ 0.014	0.553 $\pm$ 0.009	<b>0.091<math>\pm</math>0.002</b>

Table 3: Average calibration (ECE) and performance ( $R_{10}@1$ , MAP) over 5 seeds for distributional shift tasks. ” $\uparrow$ ” represents the higher the better while ” $\downarrow$ ” means the opposite.

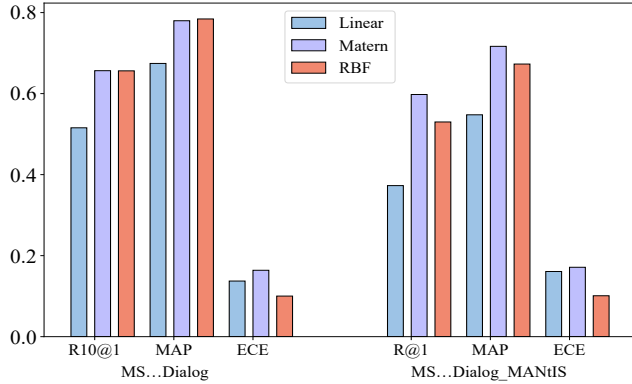


Figure 3: Performance and calibration for response retrieval tasks with different kernel functions.

As for the distributional shift task, we develop the model using the test set from a separate dataset after training it on the training set from the first dataset, which is also known as domain generalization tasks. We record all the retrieval performance and calibration in Table 3. As shown, we observe that SNGP has a lower ECE value than BERT, MC Dropout, and Ensemble, indicating better calibration. Additionally, SNGP achieves competitive performance on the  $R_{10}@1$  and MAP metrics. Under the SNGP architecture, the PG-DRR achieves a significant drop of at least 20% to the upper calibration bound, and its  $R_{10}@1$  and MAP are significantly higher or competitive in distribution shift tasks. This demonstrates that GP-based retrieval models will convey their confidence in distributional shift tasks with more robust expressiveness. According to the results of SNGP and PG-DRR, we find that performance, calibration, and resilience are strongly influenced by the posterior.

### Ablation Study

**Kernel function.** The performance of Gaussian process is associated with kernel function, so we frequently discuss how choosing a different kernel affects dialog retrieval for our PG-DRR. According to (Achituv et al. 2021), we compare the Linear, Matern and RBF kernel. The results can be found in the Figure 3. It is clear that the RBF kernel performs better in our scenario. We hypothesis that the representation in the embedding space is more mixed. The RBF kernel tend to generate non-linear decision boundaries and thus achieves stronger and better performance.

	$R_{10}@1\uparrow$	MAP $\uparrow$	ECE $\downarrow$
RoBERTa	0.632 $\pm$ 0.005	0.767 $\pm$ 0.003	0.208 $\pm$ 0.029
MC Dropout	0.615 $\pm$ 0.021	0.756 $\pm$ 0.014	0.198 $\pm$ 0.059
Ensemble	0.666 $\pm$ 0.006	0.790 $\pm$ 0.003	0.165 $\pm$ 0.018
SNGP	0.488 $\pm$ 0.072	0.669 $\pm$ 0.053	0.194 $\pm$ 0.021
PG-DRR	0.626 $\pm$ 0.024	0.764 $\pm$ 0.013	0.110 $\pm$ 0.006
	$R_{10}@1\uparrow$	MAP $\uparrow$	ECE $\downarrow$
RoBERTa	0.485 $\pm$ 0.025	0.627 $\pm$ 0.019	0.385 $\pm$ 0.024
MC Dropout	0.441 $\pm$ 0.021	0.594 $\pm$ 0.015	0.440 $\pm$ 0.024
Ensemble	0.532 $\pm$ 0.007	0.662 $\pm$ 0.005	0.330 $\pm$ 0.019
SNGP	0.384 $\pm$ 0.072	0.554 $\pm$ 0.060	0.366 $\pm$ 0.052
PG-DRR	0.484 $\pm$ 0.057	0.632 $\pm$ 0.044	0.166 $\pm$ 0.018

Table 4: Effectiveness ( $R_{10}@1$ , MAP) for dialog response retrieval. ” $\uparrow$ ” represents the higher the better while ” $\downarrow$ ” means the opposite.

**Model architecture.** We next attempt to disentangle how the differences of the model architecture affect their calibration properties. In this study, we choose the RoBERTa, another famous models in nature language processing, to verify the generalization performance of the PG-DRR method in the in-domain and cross-domain scenario. As shown in Table 4, We discovered that the RoBERTa-based approaches outperform BERT-based methods in terms of retrieval performance and calibration error. We can assume that model architecture and pretraining may affect model calibration and performance deeply. Regardless of model architectures, PG-DRR can always remain state-of-the-art calibration error, which indicates that PG-DRR can be applied to various model architectures.

### Conclusion

In this paper, we present an efficient uncertainty estimation architecture PG-DRR for reliable dialog response retrieval tasks. PG-DRR only adds a neural GP layer to a deterministic DNN to improve the ability of calibration and use the Pólya-Gamma augmentation to recover conjugacy for tractable posterior inference and optimize the parameters. In addition, Gibbs sampling is used to collect samples from the posterior. We conduct extensive experiment to verify the effectiveness between parameter and inference time. Besides, we explore the relative contributions of the kernel function and the model architecture of PG-DRR to the effectiveness improvement in the ablation study. We believe that Gaussian process modeling is an effective technique for managing calibration and hope that our work will shed more light on efficient Bayesian inference in dialog retrieval scenario.

## Acknowledgments

This paper is supported by the Key Research and Development Program of Guangdong Province under grant No. 2021B0101400003. Corresponding author is Jianzong Wang from Ping An Technology (Shenzhen) Co., Ltd (jzwang@188.com).

## References

- Achituve, I.; Navon, A.; Yemini, Y.; Chechik, G.; and Fetaya, E. 2021. GP-Tree: A Gaussian Process Classifier for Few-Shot Incremental Learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, 54–65. PMLR.
- Cohen, D.; Mitra, B.; Lesota, O.; Rekabsaz, N.; and Eickhoff, C. 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*, 654–664.
- de G. Matthews, A. G.; Hensman, J.; Turner, R. E.; and Ghahramani, Z. 2016. On Sparse Variational Methods and the Kullback-Leibler Divergence between Stochastic Processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 51 of *JMLR Workshop and Conference Proceedings*, 231–239.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.
- Douc, R.; Moulines, E.; and Stoffer, D. 2014. *Nonlinear time series: Theory, methods and applications with R examples*. CRC press.
- Durasov, N.; Bagautdinov, T.; Baque, P.; and Fua, P. 2021. Masksembles for uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13539–13548.
- Feng, Y.; Mehri, S.; Eskénazi, M.; and Zhao, T. 2020. "None of the Above": Measure Uncertainty in Dialog Response Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2013–2020.
- Galy-Fajou, T.; Wenzel, F.; Donner, C.; and Oppel, M. 2019. Multi-Class Gaussian Process Classification Made Conjugate: Efficient Inference via Data Augmentation. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI*, volume 115 of *Proceedings of Machine Learning Research*, 755–765. AUAI Press.
- Gašić, M.; and Young, S. 2013. Gaussian processes for pomdp-based dialogue manager optimization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1): 28–40.
- Girolami, M. A.; and Rogers, S. 2006. Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors. *Neural Computation*, 18(8): 1790–1817.
- Gu, J.-C.; Li, T.; Liu, Q.; Ling, Z.-H.; Su, Z.; Wei, S.; and Zhu, X. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2041–2044.
- Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330. PMLR.
- Hein, M.; Andriushchenko, M.; and Bitterwolf, J. 2019. Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 41–50. Computer Vision Foundation / IEEE.
- Jankowiak, M.; and Pleiss, G. 2021. Scalable Cross Validation Losses for Gaussian Process Models. *arXiv preprint arXiv:2105.11535*.
- Kristiadi, A.; Hein, M.; and Hennig, P. 2020. Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, volume 119 of *Proceedings of Machine Learning Research*, 5436–5446. PMLR.
- Linderman, S. W.; Johnson, M. J.; and Adams, R. P. 2015. Dependent Multinomial Models Made Easy: Stick-Breaking with the Polya-gamma Augmentation. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, 3456–3464.
- Liu, J.; Lin, Z.; Padhy, S.; Tran, D.; Bedrax Weiss, T.; and Lakshminarayanan, B. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33: 7498–7512.
- Lu, J.; Ren, X.; Ren, Y.; Liu, A.; and Xu, Z. 2020. Improving contextual language models for response retrieval in multi-turn conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1805–1808.
- Minka, T. P. 2001. *A family of algorithms for approximate Bayesian inference*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.
- Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Pei, J.; Wang, C.; and Szarvas, G. 2022. Transformer uncertainty estimation with hierarchical stochastic attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11147–11155.
- Penha, G.; Balan, A.; and Hauff, C. 2019. Introducing MAN-tIS: a novel multi-domain information seeking dialogues dataset. *arXiv preprint arXiv:1912.04639*.
- Penha, G.; and Hauff, C. 2021. On the calibration and uncertainty of neural learning to rank models for conversational search. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 160–170.



Polson, N. G.; Scott, J. G.; and Windle, J. 2013. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association*, 108(504): 1339–1349.

Qu, C.; Yang, L.; Croft, W. B.; Trippas, J. R.; Zhang, Y.; and Qiu, M. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st international acm sigir conference on research & development in information retrieval*, 989–992.

Rasmussen, C. E. 2003. Gaussian processes in machine learning. In *Summer school on machine learning*, 63–71. Springer.

Rifkin, R. M.; and Klautau, A. 2004. In Defense of One-Vs-All Classification. *J. Mach. Learn. Res.*, 5: 101–141.

Roy, N.; Pineau, J.; and Thrun, S. 2000. Spoken dialogue management using probabilistic reasoning. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, 93–100.

Snell, J.; and Zemel, R. S. 2021. Bayesian Few-Shot Classification with One-vs-Each Pólya-Gamma Augmented Gaussian Processes. In *9th International Conference on Learning Representations, ICLR*.

Tegho, C.; Budzianowski, P.; and Gašić, M. 2017. Uncertainty estimates for efficient neural network-based dialogue policy optimisation. *arXiv preprint arXiv:1711.11486*.

van Niekerc, C.; Heck, M.; Geishauser, C.; Lin, H.; Lubis, N.; Moresi, M.; and Gasic, M. 2020. Knowing What You Know: Calibrating Dialogue Belief State Distributions via Ensembles. In *Findings of the Association for Computational Linguistics*, 3096–3102.

Wenzel, F.; Galy-Fajou, T.; Donner, C.; Kloft, M.; and Oppel, M. 2019. Efficient Gaussian process classification using Pólya-Gamma data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5417–5424.

Whang, T.; Lee, D.; Oh, D.; Lee, C.; Han, K.; Lee, D.-h.; and Lee, S. 2021. Do response selection models really know what’s next? utterance manipulation strategies for multi-turn response selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14041–14049.

Williams, C. K. I.; and Barber, D. 1998. Bayesian Classification With Gaussian Processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12): 1342–1351.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. 2016. Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144.

Zhu, J.; Wang, J.; Cox, I. J.; and Taylor, M. J. 2009. Risky business: modeling and exploiting uncertainty in information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 99–106.