Being Bayesian with LLMs

Riley Sinema

BYU CS 677

April 2025

Abstract

This paper explores the application of Bayesian methods to Large Language Models (LLMs) by implementing a deep ensemble approach with Llama-3.2-1B, following the methods in Duffield et al. 2024. While full Bayesian treatment of LLMs remains computationally prohibitive, I follow the parameter-efficient strategy in Duffield et al. 2024 that freezes most model weights and applies Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) only to an ensemble of final decoder layers. This approach is evaluated on two tasks: out-of-distribution detection using translated text and binary question answering. These results demonstrate that the Bayesian ensemble effectively captures epistemic uncertainty, achieving an AUROC of 0.98 in distinguishing between in-distribution and out-of-distribution languages. Performance on question answering showed modest improvements over the base model, though with less distinct uncertainty separation between correct and incorrect predictions. These findings suggest that even simplified Bayesian approaches can enhance uncertainty estimation in LLMs for specific applications, though significant challenges remain in scaling such methods to larger state-of-the-art models.

1 Introduction

Large Language Models (LLMs) have become so popular in recent years that nearly every person knows of a commercial generative LLM platform regardless of their domain knowledge about deep learning. These LLMs have begun to be integrated into almost every field because of their impressive generalization capabilities. Despite their extensive base knowledge, LLMs often generate nonfactual statements—commonly referred to as "hallucinations"—while exhibiting overconfidence in such outputs (Huang et al. 2025). This disconnect between confidence and accuracy is particularly concerning as LLM-generated information increasingly influences human decision-making and belief formation, potentially leading to harmful outcomes when based on incorrect information. This has also slowed the integration of LLMs into critical systems that require high accuracy, such as healthcare (Yang et al. 2023).

Uncertainty quantification (UQ) in LLMs has emerged as a critical research direction that addresses these concerns. Current approaches to LLM UQ can be broadly categorized into two paradigms: white-box approaches and black-box approaches. White-box methods require access to model weights, training procedures, or output logits (Lin, Hilton, and Evans 2022), while black-box methods treat the model as closed and evaluate UQ without access to internal components (Da et al. 2024, Gao et al. 2024).

In this paper, I follow the methods in Duffield et al. 2024 to train a deep ensemble LLM using Bayesian inference on the TQA dataset (D. Kim, S. Kim, and Kwak 2019). The model's ability to quantify uncertainty is evaluated following the methods in Duffield et al. 2024. Additional experiments assess whether this UQ method improves model performance on data from the BoolQ dataset (Clark et al. 2019).

2 Related Work

2.1 Black-box UQ with LLMs

As proprietary commercial generative LLMs have increased in popularity, research into UQ for these models has also expanded. Evaluating the uncertainty of a black-box model requires creative approaches because only the input prompt and model output are accessible. Many methods rely on sampling techniques, such as the work by Da et al. 2024, who sample multiple responses from the same input and construct a directional

graph with edges representing the semantic relationships between outputs. Uncertainty is then derived from properties of this constructed graph. Similarly, Gao et al. 2024 introduced a sampling method that systematically perturbs the input before collecting each sample, then aggregates the responses to calculate model uncertainty for the original input statement. These approaches are particularly valuable for commercial LLMs where internal model details remain inaccessible.

2.2 White-box UQ with LLMs

White-box UQ methods require deeper access to the model architecture or training process to evaluate uncertainty. These approaches typically leverage model weights, training procedures, or output logits. Lin, Hilton, and Evans 2022 employed a white-box approach by fine-tuning models to explicitly output confidence levels alongside answers to questions, directly encoding uncertainty estimation into the model's behavior. In a different approach, Kuhn, Gal, and Farquhar 2023 calculated semantic entropy from the output distribution of the model, providing a measure of uncertainty based on the variability in potential responses. My work also requires access to the model weights and output logits, specifically focusing on Bayesian methods for uncertainty quantification in open-source LLMs.

3 Methods

3.1 Model

The base model used in this paper is the open source Llama-3.2-1B (Grattafiori et al. 2024). This model was chosen to be compatible with the example from Duffield et al. 2024 and due to memory and compute restraints on available machines.

For the Bayesian ensemble approach, I adopt a parameter-efficient fine-tuning strategy. While a full Bayesian treatment of all parameters would be ideal, the computational complexity of sampling the entire parameter space of a generative LLM—even this 1B model which is one of the smallest models available—is prohibitive. Following Duffield et al. 2024, the majority of the network is treated as a deterministic feature extractor by freezing all weights except those in the final decoder layer. This approach is motivated by practical considerations (for reducing computational cost) and also theoretical insights suggesting that uncertainty in the final layer can effectively approximate predictive uncertainty (Kristiadi, Hein, and Hennig 2020).

The deep ensemble consists of 10 instances of the final decoder layer with identical architecture but different parameter values. Each instance maintains the same input and output dimensionality as the original layer, ensuring compatibility with the frozen network components. During inference, each final layer processes the same features from the final frozen layer and produces its own distribution over the next token. The final prediction is determined by a majority voting scheme, where the token receiving the most votes across all 10 layers becomes the output.

3.2 Datasets

Following the methods in Duffield et al. 2024, the TQA dataset D. Kim, S. Kim, and Kwak 2019 was used to train the primary deep ensemble. This dataset contains question and answer data with context from middle school science curriculum.

A second experiment was conducted using the BoolQ dataset (Clark et al. 2019), a collection of yes/no questions paired with passages from Wikipedia articles. For this experiment, a separate deep ensemble with the same architecture was trained on examples from BoolQ.

For both datasets, input sequences were processed with a stride of 300 tokens and an overlap of 100 tokens during training. All model evaluations were performed exclusively on the validation sets to ensure unbiased assessment of model performance.

3.3 Training

The deep ensemble model was trained using Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) provided by the posteriors module introduced by Duffield et al. 2024. The 10 ensemble layers were reinitialized using Kaiming uniform distribution initialization (He et al. 2015) to ensure diversity in the posterior modes that are explored. For SGHMC training, hyperparameters were set to a learning rate of 1e-3, alpha of 1e-2,

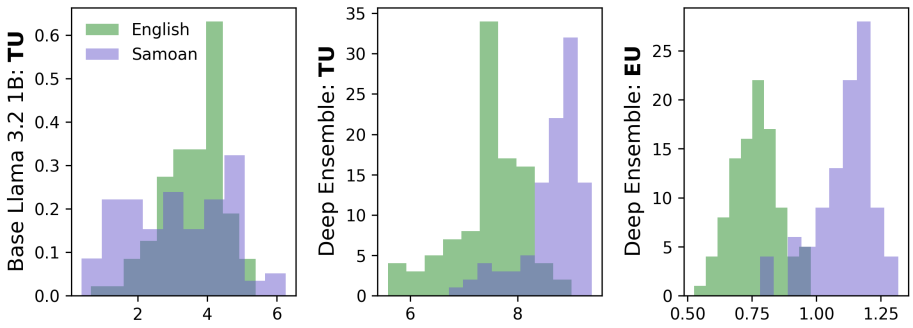

Figure 1: Distributions of uncertainty scores for the TQA statements in English and Samoan. The base model can only calculate the total uncertainty, while the ensemble model can separate the epistemic uncertainty from the total uncertainty. Note the clear separations of the English and Samoan uncertainty distributions for the epistemic uncertainty of the ensemble model.

beta of 0, and momenta of 0. The ensemble was trained for 20 epochs with a batch size of 10 on 5000 training examples.

3.4 Evaluation

Uncertainty evaluation in this paper follows the methodology outlined in Duffield et al. 2024, which decomposes predictive uncertainty into epistemic and aleatoric components. Given M heads in the deep ensemble model, each head outputs a vector of logits ℓ_m which is transformed via softmax to create the probability distribution $p_m = \text{softmax}(\ell_m)$. The mean of these distributions is denoted as $\bar{p}$.

The total uncertainty of the model is quantified as the entropy of the mean distribution:

$$U_{total} = -\sum_{i=1}^{C} \bar{p}_i \log \bar{p}_i \tag{1}$$

where C is the vocabulary size.

The aleatoric uncertainty (capturing data-inherent randomness) is computed as the average entropy of each ensemble head:

$$U_{aleatoric} = \frac{1}{M} \sum_{i=1}^{M} \left(-\sum_{j=1}^{C} p_{i,j} \log p_{i,j} \right) \tag{2}$$

The epistemic uncertainty (representing model uncertainty) is then derived as:

$$U_{epistemic} = U_{total} - U_{aleatoric} \tag{3}$$

To evaluate the model's uncertainty calibration, I follow Duffield et al. 2024 to utilize the Area Under the Receiver Operating Characteristic (AUROC) metric. This metric assesses how well the uncertainty estimates correlate with prediction errors, with higher values indicating better discrimination between correct and incorrect predictions.

4 Experiments

4.1 Out of Distribution Detection

The models were evaluated on 100 simple scientific statements with the final word removed from the TQA validation set (D. Kim, S. Kim, and Kwak 2019). While the statements in the dataset are originally in English, for this experiment they were evaluated both in their original form and after being translated to Samoan to create an out-of-distribution scenario.

Both models were prompted to complete each statement, and uncertainty metrics were computed on their outputs. For the base Llama-3.2-1B model, only total uncertainty could be evaluated due to its non-ensemble architecture. For the deep ensemble model, both total uncertainty and epistemic uncertainty were calculated following the methodology described in Section 3.4.

As shown in Figure 1, the uncertainty distributions of the base model exhibit substantial overlap between English and Samoan statements, making it difficult to distinguish between in-distribution and out-of-distribution samples based on uncertainty alone. In contrast, the deep ensemble model demonstrates a clear separation between uncertainty distributions for English versus Samoan statements, particularly in its epistemic uncertainty measure.

Table 1: AUROC scores of base model and deep ensemble on TQA statements.

	Base Llama 3.2 1B	Deep Ensemble
	TU	**TU EU**
AUROC ↑	0.42	0.92 **0.98**
Loss on English statements ↓	**3.98**	4.64
Loss on Samoan statements ↓	12.67	**10.06**

The AUROC scores presented in Table 1 quantify this separation, with the ensemble's epistemic uncertainty achieving an AUROC of 0.98 for distinguishing between in-distribution (English) and out-of-distribution (Samoan) examples. This high score indicates that the epistemic uncertainty strongly correlates with the model's prediction confidence.

4.2 Binary Question Answering

The models were evaluated on 1000 samples randomly selected from the BoolQ validation set (Clark et al. 2019). Each model was provided with the context passage and corresponding question, then prompted to answer with either 'True' or 'False'. For the deep ensemble, each of the 10 heads voted on an output token, and the token receiving the most votes was selected as the final answer. As a baseline comparison, we included the performance of a naive classifier that always outputs 'True', which is a common benchmark for this dataset. The resulting accuracies are presented in Table 2.

	Base Llama 3.2 1B	Deep Ensemble	Baseline
Accuracy ↑	0.48	0.57	**0.59**

Table 2: The accuracies of the Base Llama-3.2-1B, Deep Ensemble, and baseline on 1000 samples from the BoolQ dataset. The baseline is outputting 'True' for every question.

The deep ensemble model substantially outperformed the base Llama-3.2-1B model, achieving an accuracy improvement of 0.09 percentage points. However, it fell short of the naive baseline by 0.589 percentage points, suggesting room for further optimization. Figure 2 illustrates the uncertainty distributions for both correct and incorrect predictions from the ensemble model.

Both uncertainty distributions exhibit notable right-skew, indicating that the model was highly confident in most of its predictions. However, the mean uncertainty for incorrect predictions (TU: 2.68, EU: 0.30) was higher than for correct predictions (TU: 2.4, EU: 0.27), confirming that the model's uncertainty estimates somewhat correlate with its prediction accuracy.

5 Conclusion

This paper investigated the application of Bayesian methods to LLMs following the approach of Duffield et al. 2024. I constructed a deep ensemble model using Llama-3.2-1B (Grattafiori et al. 2024) as the base architecture and evaluated its performance against the standard model on both out-of-distribution detection and question answering tasks using the TQA (D. Kim, S. Kim, and Kwak 2019) and BoolQ (Clark et al. 2019) datasets.

The experiments in Section 4 demonstrated that the Bayesian deep ensemble approach offers promising capabilities in uncertainty estimation. As shown in Section 4.1, the model effectively identified out-of-distribution languages, with the epistemic uncertainty distributions in Figure 1 revealing clear separation between in-distribution and out-of-distribution inputs. The high AUROC score for epistemic uncertainty demonstrated a strong correlation between model accuracy and uncertainty estimates. For binary question answering in Section 4.2, while the deep ensemble achieved modestly higher accuracy than the base

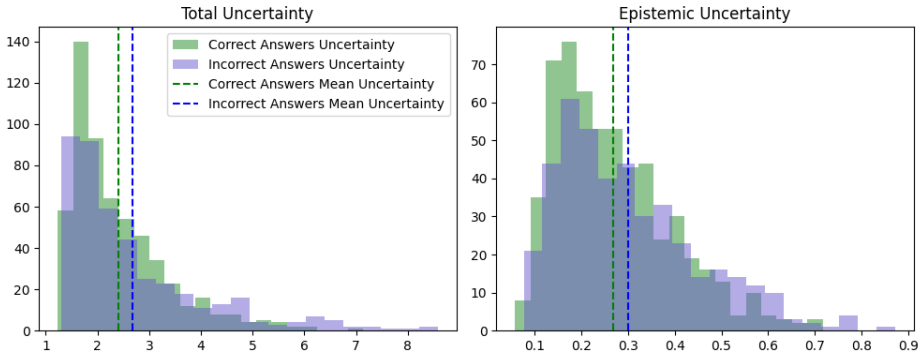

Figure 2: Distributions of uncertainty for correct and incorrect answers from the deep ensemble model. The mean uncertainties for the correct answers are 2.4 (TU) and 0.27 (EU). The mean uncertainties for the incorrect answers are 2.68 (TU) and 0.30 (EU).

model and showed slightly elevated uncertainty for incorrect answers, the uncertainty distributions exhibited considerable overlap. This suggests that Bayesian methods show promise for specific applications like out-of-distribution detection, but their general applicability across all NLP tasks requires further investigation.

Despite these encouraging results, this work highlights significant practical challenges in applying Bayesian methods to modern LLMs. The base model used in this paper, Llama-3.2-1B (Grattafiori et al. 2024), is already among the smallest and simplest to work with open-source models available—yet still required substantial computational resources, technical domain skills, and theoretical knowledge for Bayesian treatment. With state-of-the-art models now around 600 times larger, scaling Bayesian approaches accordingly presents formidable challenges. Duffield et al. 2024 address this in their introduction of the `posteriors` library, but much work is still needed in this area. Future research might explore more parameter-efficient Bayesian approximations or hybrid approaches that balance computational feasibility with uncertainty quantification quality. As LLMs continue to be deployed in critical applications, developing practical uncertainty estimation techniques remains an important research direction, even if full Bayesian treatments may be computationally prohibitive for the largest models.

References

He, Kaiming et al. (Dec. 2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Clark, Christopher et al. (2019). "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions". In: *NAACL*.

Kim, Daesik, Seonhoon Kim, and Nojun Kwak (2019). *Textbook Question Answering with Multi-modal Context Graph Understanding and Self-supervised Open-set Comprehension*. arXiv: 1811.00232 [cs.CL]. URL: https://arxiv.org/abs/1811.00232.

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

Lin, Stephanie, Jacob Hilton, and Owain Evans (2022). *Teaching Models to Express Their Uncertainty in Words*. arXiv: 2205.14334 [cs.CL]. URL: https://arxiv.org/abs/2205.14334.

Kuhn, Lorenz, Yarin Gal, and Sebastian Farquhar (2023). *Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation*. arXiv: 2302.09664 [cs.CL]. URL: https://arxiv.org/abs/2302.09664.

Yang, Rui et al. (2023). "Large language models in health care: Development, applications, and challenges". In: *Health Care Science* 2.4, pp. 255–263.

Da, Longchao et al. (2024). "Llm uncertainty quantification through directional entailment graph and claim level response augmentation". In: *arXiv preprint arXiv:2407.00994*.

Duffield, Samuel et al. (2024). "Scalable Bayesian Learning with posteriors". In: *arXiv preprint arXiv:2406.00104*.

Gao, Xiang et al. (2024). "Spuq: Perturbation-based uncertainty quantification for large language models". In: *arXiv preprint arXiv:2403.02509*.

Grattafiori, Aaron et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

Huang, Lei et al. (Jan. 2025). "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Trans. Inf. Syst.* 43.2. ISSN: 1046-8188. DOI: 10.1145/3703155. URL: https://doi.org/10.1145/3703155.