UNCERTAINTY QUANTIFICATION IN FINE-TUNED LLMS USING LORA ENSEMBLES

Oleksandr Balabanov
Stockholm University
Department of Physics
Stockholm, Sweden
oleksandr.balabanov@fysik.su.se

Hampus Linander
Department of Mathematical Sciences
Chalmers university of technology & University of Gothenburg
Gothenburg, Sweden &
VERSES Research Lab
Los Angeles, California, USA
hampus.linander@chalmers.se

ABSTRACT

Fine-tuning large language models can improve task specific performance, although a general understanding of what the fine-tuned model has learned, forgotten and how to trust its predictions is still missing. We derive principled uncertainty quantification for fine-tuned LLMs with posterior approximations using computationally efficient low-rank adaptation ensembles. We analyze three common multiple-choice datasets using low-rank adaptation ensembles based on Mistral-7b, and draw quantitative and qualitative conclusions on their perceived complexity and balance between retained prior knowledge and domain specific adaptation during and after fine-tuning. We identify unexpected retention of acquired knowledge during fine-tuning in the overfitting regime.
github.com/oleksandr-balabanov/equivariant-posteriors

1 INTRODUCTION

Large language models (LLMs) learns conditional distributions of vocabularies useful for generative tasks, text classification and code completion. Trained on a corpus of sequences such as natural language text and program source code, these models have shown remarkable capabilities (Bubeck et al., 2023; Touvron et al., 2023a;b; Jiang et al., 2023). As the applications of LLMs continue to explode, informed use of these models hinge on a proper understanding of the uncertainty of their output (Huang et al., 2023; Kuhn et al., 2023; Malinin & Gales, 2021; Ren et al., 2023).

To align an LLM towards particular needs, such as answering factual questions and instructions, a common approach is to fine-tune the model using a specialised dataset targetting human interaction (Ouyang et al., 2022). Fine-tuning involves additional training of a pre-trained LLM on a smaller dataset generated with humans (Zhong et al., 2021) or by other LLMs (Peng et al., 2023; Zhang et al., 2023a). By fine-tuning, the model adapts its parameters to better capture the nuances, vocabulary, and style of the target domain(Peng et al., 2023).

Even full-model fine-tuning often requires orders of magnitude less training time, but still incur the same computational complexity and memory usage. Prior studies have shown that pre-trained language models can effectively adapt to specific tasks within smaller parameter subspaces Li et al. (2018); Aghajanyan et al. (2020), indicating an inherently low rank during the fine-tuning process. To further reduce the training time and computational complexity of fine-tuning, a common method used is Low-Rank Adaptation (LoRA) (Hu et al., 2021). As one member of the family of methods typically referred to as parameter efficient fine-tuning (PEFT), LoRA effectively reduces the number of parameters requiring training by keeping the pre-trained weights unchanged and integrating low-rank trainable matrices into each layer of the LLM transformer architecture.

Key questions arise after fine-tuning: What areas of knowledge remain outside the model's expertise? What knowledge is retained from the pre-trained model? What knowledge is gained during the fine-tuning process on the target dataset? These questions are fundamental in guiding us towards a more reliable, interpretable and trustworthy application of LLMs.

We contribute to addressing these questions by analyzing the uncertainty in LLMs that have been fine-tuned using LoRA for answering multiple-choice question answers (QAs). These tasks involve definite single-token target labels, which significantly simplifies the analysis while still requiring the model to have a thorough understanding of context. We use predictive entropy and mutual information to quantify inherent- and model uncertainty. These entropic uncertainty measures are calculated for a Bayesian posterior derived from an ensemble of LLMs fine-tuned with LoRA.

Using the pre-trained Mistral-7b Jiang et al. (2023) model as a prior, we fine-tune on the CommonsenseQA (CQA) Talmor et al. (2019) and evaluate its performance on both CQA and the Social Sciences and STEM components of the MMLU Hendrycks et al. (2021) multiple-choice QA datasets. By quantifying the evolution of knowledge using predictive entropy and mutual information, we demonstrate how these measures can be used to reason about the balance between retained prior knowledge and acquired domain specific knowledge during and after fine-tuning.

2 Contributions

- We show how uncertainty quantification with entropic measures can be used to reason about the balance between retained knowledge and domain specific adaptation.
- We derive posterior approximations for LLMs fine-tuned on target datasets using ensembles of LoRA members. On the way, we provide a Bayesian interpretation of fine-tuning, early-stopping and conditional generative tasks.
- We analyze three common multi-choice QA datasets, CQA, MMLU STEM and MMLU Social Sciences, using LoRA ensemble posteriors derived from a pre-trained Mistral-7b model, and identify unexpected knowledge retention in the overfitting regime.

3 Related Work

3.1 Uncertainty quantification in LLMs

There has been significant interest in uncertainty quantification across various tasks and domains in neural networks Gal & Ghahramani (2015); Gal & Ghahramani (2016); Malinin & Gales (2018); Ovadia et al. (2019b); Malinin et al. (2021); Lin et al. (2022); Kuhn et al. (2023); Lin et al. (2023).

This interest extends to the realm of Large Language Models (LLMs), where accurately quantifying prediction uncertainty is a key focus Xiao et al. (2022a); Lin et al. (2022); Mielke et al. (2022); Chen & Mueller (2023); Duan et al. (2023); Huang et al. (2023); Shorinwa et al. (2024). The application of LLMs in generative tasks introduces unique challenges, notably in measuring uncertainty of the generative outputs. Liu et al. (2019); Malinin & Gales (2021); Kuhn et al. (2023); Lin et al. (2023). The disentanglement of uncertainty into aleatoric and epistemic was recently discussed in the context of LLMs Hou et al. (2023); Ling et al. (2024). However, it was done via ensembling of the model inputs rather than the model instances, and not in the context of fine-tuning tasks.

3.2 Fine-tuning in LLMs

Fine-tuning has recently become an integral part of the LLM ecosystem where it is used to target specific tasks such as general instruction-following model, through methods like Reinforcement Learning from Human Feedback Houlsby et al. (2019); Hu et al. (2021); Liu et al. (2019); Ding et al. (2022; 2023).

The large computational demands of training and fine-tuning LLMs have resulted in development of more efficient techniques, known as parameter-efficient fine-tuning (PEFT)Liu et al. (2022); Ding et al. (2022; 2023); Shi & Lipani (2024). These methods typically involve training a small number of additional parameters on top of a fixed, pre-trained LLM, with a prominent approach being the use of low-rank adapters (LoRA) for each weight matrix Hu et al. (2021).

3.3 LLMs WITHIN BAYESIAN METHODS

Bayesian inference techniques in neural networks have a long-standing history Denker et al. (1987); Tishby et al. (1989); Buntine & Weigend (1991); MacKay (1991). These methods establish a systematic approach to derive reliable and interpretable uncertainty estimates. Previous research on Bayesian language models has been primarily concentrated on the pretraining phase of language models, rather than on their fine-tuning Tran et al. (2019); Xue et al. (2021); Cinquin et al. (2021); Zhang et al. (2018); Chen & Li (2024).

Recent studies have begun exploring the fine-tuning of language models using a Bayesian approach. For instance, in Fan et al. (2020) and Zhang et al. (2021), the attention modules are sampled either from simplex-constrained attention distributions or using Bayesian Belief Networks. Neither of these studies address entropic uncertainty measures, nor do they utilize ensembling or PEFT methods for posterior approximation.

Yang et al. (2024) employs a post-hoc Laplace approximation Mackay (1992) to model LoRA parameters for fine-tuning. While this study does use LoRA for fine-tuning, it does not explore entropic uncertainty measures and focuses on the posterior over LoRA parameters rather than the model as a whole, which could limit straightforward Bayesian interpretability. There are also strong indications that deep ensembles provide more accurate posteriors compared to single-model stochastic methods like Laplace and Monte Carlo dropout Gustafsson et al. (2019); Ovadia et al. (2019a); Fort et al. (2019); Wilson & Izmailov (2020); Dwaracherla et al. (2022); Balabanov et al. (2023).

3.4 ENSEMBLING LLMs

Two recent studies have explored the use of ensembling in fine-tuning LLMs, specifically focusing on full model fine-tuning method where all weights are optimized Gleave & Irving (2022); Sun et al. (2022). This approach has large memory overhead by construction. While uncertainty quantification using variance across the ensembles is considered, their methods lack a Bayesian formalism that could provide a more grounded interpretation and understanding.

An alternative method, BatchEnsemble Wen et al. (2020), employs a base model with modifications through multiplicative, component-specific rank-1 matrices. This method has been applied to LLMs, but in the context of pre-training rather than fine-tuning Tran et al. (2022).

There are also recent endeavors employing LoRA ensembles for fine-tuning LLMs Wang et al. (2023); Zhai et al. (2023). Although these studies address uncertainty quantification, they do so without adopting a Bayesian framework and do not clearly distinguish between epistemic and aleatoric components, which is critical for data interpretability. In another work Arteaga et al. (2024), a separate predictor is trained on the LoRA ensembles uncertainty metrics to detect hallucinations. Niu et al. (2024) introduces a mixture-of-expert framework (UQ4CT) to explicitly capture and calibrate model uncertainty during fine-tuning. All these works are highly relevant yet differ from ours, which provides a principled Bayesian treatment for LoRA ensembles during fine-tuning and examines the evolution of uncertainty metrics to draw conclusions about the balance between retained prior knowledge and newly acquired domain adaptation.

4 BAYESIAN DEEP LEARNING FOR LLMs

When formulated as a conditional generative model, an LLM can be viewed as a function taking a sequence of tokens s^* as input, and predicting a distribution over the set of possible tokens t^* from the vocabulary as output. Such a model is typically trained to predict the next token in a sequence.

The prediction should be consistent with our observed data. This is encapsulated in the predictive probability distribution $p(t^*|s^*, \mathcal{D})$, with training data $\mathcal{D} = \{s_i, t_i\}_{i=1}^N$, giving a conditional distribution over tokens t^*. To obtain a specific prediction, we can sample from this distribution, for example, using maximum posterior estimation, $t^* = \arg\max_t p(t|s^*, \mathcal{D})$.

Given a model with parameters θ, the predictive distribution $p\big(t^* \mid s^*, \mathcal{D}\big) = \int p\big(t^* \mid s^*, \theta\big)\, p\big(\theta \mid \mathcal{D}\big)\, d\theta$ is obtained by marginalizing over the posterior distribution of the parameters. The posterior $p(\theta \mid \mathcal{D})$ represents the uncertainty associated with model parameters θ, quantifying how likely each choice of θ is, given the observed data. For a fixed θ, the term $p(t^* \mid s^*, \theta)$ captures data uncertainty,

reflecting the variability in the model's output t^* for the input s^*. By integrating over all possible parameter values θ, the final predictive distribution combines both uncertainty in the parameters with the inherent variability in the data Gawlikowski et al. (2023).

4.1 FINE-TUNING

The objective of fine-tuning is to use a specialized dataset to further improve the knowledge of a generally pre-trained model for a given task. This process can be conceptualized as a conditional generative task on a target domain, denoted as $\mathcal{D}_{\text{fine-tuning}}$.

Fine-tuning assumes prior knowledge about the models parameters, expressed as $p(\theta)$. This prior is typically centered around the pre-trained models parameters, $\omega_{\text{pre-trained}}$, to retain the existing knowledge and apply it to the target domain. We assume a normal prior distribution centered on the pre-trained parameter values: $p(\theta) = N(\theta; \omega_{\text{pretrained}}, \lambda^{-1} I_{\text{dim}[\theta]})$ with diagonal covariance $I_{\text{dim}[\theta]}$. By adjusting the prior variance λ^{-1}, we can control the extent to which the model is allowed to deviate from the pre-trained parameter values. In practice, this involves finding a balance between two extremes: a small λ^{-1} means that the fine-tuning is largely disregarded, whereas a large λ^{-1} implies that the pre-trained models knowledge is not utilized. As with any posterior considerations in a Bayesian formalism, fine-tuning requires careful adjustment of this parameter to target optimal performance for the target tasks.

Note that in this formulation, there is no emphasis on the origin of the pre-trained parameters. Typically, for large language models, the training procedure consists of maximum likelihood optimization combined with supervised reinforcement learning Touvron et al. (2023b) and the specific details of the training procedure are often not publicly disclosed. In our formulation, we do not need to delve into these origins and instead formally take it as prior knowledge upon which we base our fine-tuning.

4.2 POSTERIOR APPROXIMATION METHODS

Variational inference Barber & Bishop (1998); Graves (2011); Blundell et al. (2015) offers a computationally efficient approximation method for the Bayesian posterior. It involves exploring a family of distributions $q_\omega(\theta)$, parameterized by ω, to approximate the posterior $p(\theta|\mathcal{D})$. A prevalent approach to identify the optimal $q_\omega(\theta)$ in this set involves minimizing the Kullback-Leibler (KL) divergence $\text{KL}(q_\omega(\theta) \,||\, p(\theta|\mathcal{D}))$ Gawlikowski et al. (2023). This divergence, which measures information loss when $q_\omega(\theta)$ approximates $p(\theta|\mathcal{D})$, can be expressed as:

$$\begin{aligned} \text{KL}(q_\omega(\theta) \,||\, p(\theta|\mathcal{D})) &= \int d\theta \, q_\omega(\theta) \, \log \frac{q_\omega(\theta)}{p(\theta|\mathcal{D})} \\ &= \text{KL}(q_\omega(\theta) \,||\, p(\theta)) - \mathbb{E}_{q_\omega(\theta)}[\log \, p(t|\theta, s)] + C, \end{aligned} \tag{1}$$

where $\mathcal{D} = \{(t, s)\}$ represents the training data, and C is a constant term independent of θ. Here we applied Bayes' theorem $p(\theta|t, s) = p(t|\theta, s)\, p(\theta)/p(t|s)$. Minimizing the loss function given in Eq. (1) results in an approximation of the posterior $p(\theta|\mathcal{D})$. This loss function includes two components: (1) the KL divergence against the prior $p(\theta)$ and (2) the expected negative log likelihood (ENLL), a common loss metric for classification tasks.

The optimization problem described in Eq. (1) does not specify the origin of the in-domain (observed) data $\mathcal{D}$, and it remains equally valid for a fine-tuning task when substituting $\mathcal{D} = \mathcal{D}_{\text{fine-tune}}$. However, it's important to stress that for a fine-tuning task, the prior is assumed to be centered around the pre-trained solution $\omega_{\text{pretrained}}$, whereas for a typical training procedure, the prior is usually considered to be centered around the origin in parameter space.

4.3 EARLY STOPPING

Early stopping can be used to enhance generalization to unseen in-domain data, thereby reducing the effect of overfitting. In the context of Bayesian deep learning, the objective is to approximate the posterior distribution $p(\theta|\mathcal{D})$, where $\mathcal{D}$ represents the entire task domain. Access to the full domain $\mathcal{D}$ is typically not available, and we work with a subset $\mathcal{D}^{\text{train}}$. Generally, $p(\theta|\mathcal{D}^{\text{train}})$ does not closely mirror $p(\theta|\mathcal{D})$, leading to overfitting, where $p(\theta|\mathcal{D}^{\text{train}})$ only acknowledges the training data and struggles to generalize.

Early stopping improves the approximation of $p(\theta|\mathcal{D})$. Within the variational inference picture, $q_\omega(\theta)$ is optimized by minimizing $\mathrm{KL}(q_\omega(\theta)\,||\,p(\theta|\mathcal{D}^{\text{train}}))$, but evaluated on the entire task domain $\mathcal{D}$ using $\mathrm{KL}(q_\omega(\theta)\,||\,p(\theta|\mathcal{D}))$. The prior contribution to the KL loss $\mathrm{KL}(q_\omega(\theta)\,||\,p(\theta|\mathcal{D}))$ can be dropped assuming the entire task domain $\mathcal{D}$ to have a large number of samples. In practice, the ENLL contribution to the KL loss is estimated using the validation dataset $\mathcal{D}^{\text{val}}$, assuming it represents the entire task domain effectively and was not part of the training.

5 Deep Ensembles with Low-Rank Adaptation

5.1 Low-Rank Adaptation

Low-Rank Adaptation (LoRA) provides a computationally efficient alternative to full-model fine-tuning by keeping the pre-trained weight matrices static and adding trainable low-rank matrices into each transformer layer, thereby decreasing the total number of parameters requiring training Hu et al. (2021). Given a pre-trained weight matrix $W_{\text{pretrained}}$ in the space $\mathbb{R}^{d\times k}$, the adaptations are represented as $W_{\text{pretrained}} + BA$, where B resides in $\mathbb{R}^{d\times r}$, A in $\mathbb{R}^{r\times k}$, and the rank r is substantially smaller than both d and k. Throughout this training approach, the variables A and B are tuned, while $W_{\text{pretrained}}$ is kept unchanged.

5.2 LoRA Deep Ensembles

We use LoRA ensembles for approximating the posterior $p(\theta|\mathcal{D}_{\text{fine-tune}})$. Ensembles of LoRA members have been introduced recently Wang et al. (2023), here we provide a Bayesian treatment.

Deep ensembles can be explicitly interpreted within the Bayesian variational framework by assuming an ansatz in the form of a sum of sharply peaked distributions around parameter realizations ω_k, with k indexing ensemble members Hoffmann & Elster (2021); Balabanov et al. (2023). In this context, the variational KL loss simplifies to a sum of conventional single-member log likelihood loss terms, augmented with L2 regularization.

We adopt the deep ensemble variational inference formulation to include LoRA reparametrization. The trainable low-rank LoRA matrices A_k and B_k are not treated as a subset of the variational parameters ω_k of the deep ensemble ansatz. Rather, they are auxiliary degrees of freedom that parameterize these variational parameters. That is, we multiply matrices A_k and B_k to obtain $W_k = (W_{\text{pretrained}} + A_k B_k) \subset \omega_k$, and W_k is used in the deep ensemble ansatz for approximating the posterior distribution. The LoRA parametrisation helps us to efficiently and inexpensively find good values for $W_k \subset \omega_k$.

5.3 Prior Selection

The choice of an optimal prior is crucial for obtaining a posterior distribution that exhibits optimal performance. Our approach is to assume a normally distributed prior around the pre-trained model, $p(\theta) = N(\theta; \omega_{\text{pretrained}}, \lambda^{-1} I_{\dim[\theta]})$, defined with a certain variance λ^{-1}. This variance is a hyperparameter that must be carefully chosen based on specific objectives.

An inappropriate selection of the prior can lead to a posterior that performs poorly on the task at hand. Consequently, we adjust the variance of the prior with the aim of optimizing the posterior quality, as measured by the log likelihood on the validation data. Different choice for the prior variance results in adjusting L2 regularization loss Hoffmann & Elster (2021); Balabanov et al. (2023). In the case of LoRA ensembles, this loss is given by

$$\mathrm{L2} = \frac{\lambda}{2}\sum_i ||W^{(i)} - W^{(i)}_{\text{pretrained}}||^2 = \frac{\lambda}{2}\sum_i ||B^{(i)}A^{(i)}||^2, \tag{2}$$

where $A^{(i)}$ and $B^{(i)}$ are LoRA matrices associated with the weights $W^{(i)}$.

6 UNCERTAINTY QUANTIFICATION

6.1 EPISTEMIC MODEL AND ALEATORIC DATA UNCERTAINTY

Uncertainty quantification is an important topic that assesses the reliability of model predictions (Gal & Ghahramani, 2016). In general, the total uncertainty in a model's prediction can be attributed to multiple sources. One widely used categorization is the decomposition into epistemic (or model) uncertainty—arising from insufficient knowledge or data to fully learn the underlying distribution—and aleatoric (or data) uncertainty—stemming from the inherent noise and randomness in the data generation process.

An example of aleatoric data uncertainty is given by the input sequence "There is a single item in the box. The color of that item is ". One would expect a model trained on a large corpus of text to produce a predictive distribution with support on the different colors. This posterior distribution tells us that the model is fairly certain that the next token is a color, but that there is an uncertainty about which color. The uncertainty about the color cannot be decreased by improving the training data; it is inherent in the input sequence.

On the other hand, the input sequence "Large language models are " presents a simple conditional completion task, but we would not trust the output predictions from a model trained on a text corpus compiled before the concept was introduced. This epistemic model uncertainty is not captured in the output distribution, but rather in the uncertainty of the model parameters themselves. This model uncertainty can be systematically calculated within a Bayesian formalism, where it corresponds to the shape of the posterior distribution of model parameters conditioned on the training data.

6.2 ENTROPIC UNCERTAINTY METRICS

Depending on the approach used to estimate uncertainty, existing methods can be broadly divided into intrinsic and extrinsic methods. Extrinsic methods involve post-hoc analysis with auxiliary models to estimate uncertainty Lahlou et al. (2023); Kristiadi et al. (2021). In contrast, intrinsic methods incorporate uncertainty estimation directly within the model. Here, we employ a common intrinsic methodology based on entropic metrics Malinin & Gales (2021) to evaluate uncertainty within a Bayesian formulation, applying it to our LoRA ensemble approximation of the Bayesian posterior for fine-tuned models.

The total predictive uncertainty, encompassing both aleatoric data and epistemic model components, can be quantified using the total predictive entropy:

$$H(t^*|s^*, \mathcal{D}) = -\sum_c p(t^* = c|s^*, \mathcal{D}) \log p(t^* = c|s^*, \mathcal{D}), \tag{3}$$

where s^* is the input string, $\mathcal{D}$ is the training dataset, and c denotes all possible output tokens. The output probabilities $p(c \mid s^*, \mathcal{D})$ are computed by averaging softmax outputs across model parameters θ sampled from the Bayesian posterior $p(\theta \mid \mathcal{D})$.

Epistemic model uncertainty is defined as the variability in predictions that arises from sampling model parameters θ from the posterior distribution $p(\theta \mid \mathcal{D})$. This variability captures how the distribution of predicted outputs shifts with each parameter set, weighted by its posterior probability. A common metric for this purpose is the average mutual information between the model parameters and output, conditioned on the training data Mackay (1992); Malinin & Gales (2021):

$$\text{MI}(\theta, t^*|s^*, \mathcal{D}) = H(t^*|s^*, \mathcal{D}) - \mathbb{E}_{\theta \sim p(\theta|\mathcal{D})} \left[H(t^*|s^*, \theta) \right]. \tag{4}$$

For ensemble posteriors, zero mutual information implies that all ensemble members yield the same predictive distribution, indicating no variability in the outputs and thus zero epistemic uncertainty. On the contrary, if the mutual information is high, there is significant variability in the predictive distribution across the ensemble, reflecting high epistemic uncertainty.

Using these metrics, we can identify three regimes that describe how the model perceives data:

- *Known-certain*: Low total and epistemic uncertainty; the posterior ensemble strongly agrees on a narrow, confident prediction distribution.

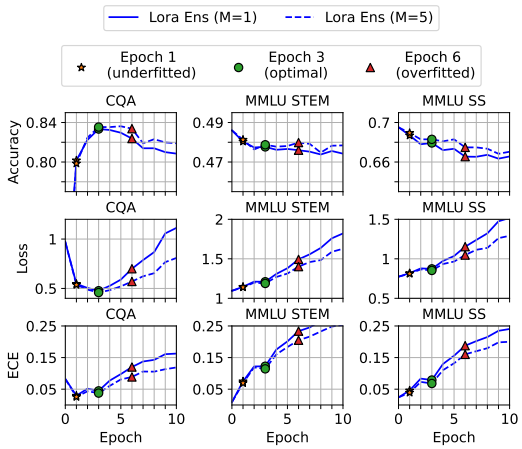

Figure 1: Performance of the LoRA ensembles trained on CQA and evaluated on either CQA, MMLU STEM, or MMLU SS datasets. The metrics (evaluated for the ensemble mean distribution) include accuracy, log likelihood loss, and expected calibration error (ECE). The number of ensemble members is either $M = 5$ or $M = 1$. For a LoRA ensemble with $M = 1$ the average results over 5 distinct realizations are shown. We also mark the metric values corresponding to epoch 1 (underfitting), epoch 3 (optimal, minimal loss), and epoch 6 (overfitting).

- *Known-uncertain*: High total uncertainty, low epistemic uncertainty; uncertainty arises from data noise or model expressiveness limitations, with the ensemble members agreeing on a broad distribution.
- *Unknown*: High total and epistemic uncertainty; the posterior members disagree on the prediction distribution, highlighting a lack of knowledge.

In this work, we use uncertainty quantification from entropic metrics of the LoRA ensemble posterior approximation to better understand how fine-tuning reshapes the knowledge of the model, revealing shifts in data characteristics and how the model adapts to new information. In particular, we study the dynamics of the uncertainty metrics during fine-tuning.

7 Numerical Results

7.1 Datasets

We evaluate our methods on three multiple-choice QA datasets: CommonsenseQA (CQA) Talmor et al. (2019), and the Social Sciences (SS) as well as STEM components from MMLU Hendrycks et al. (2021). The ensemble is trained on the CQA training dataset but tested on three different validation datasets: the (in-domain) CQA validation dataset and the (out-of-domain) MMLU STEM and MMLU SS validation datasets. Detailed information about the training and validation (test) splits is available in Appendix A.

7.2 LoRA Ensembles

We train LoRA ensembles using the pre-trained 7 billion parameter Mistral-7b model Jiang et al. (2023). The LoRA parametrization was implemented using PEFT Mangrulkar et al. (2022). Our ensembles contain 5 members ($M = 5$), but we also investigate larger ensembles in our ablation studies. Adhering to the method of Hu et al. Hu et al. (2021), we applied the adapter $\Delta W = \alpha BA$ exclusively to the query and value matrices within the self-attention modules of Mistral-7b, with α set to 32. The rank was set to $r = 8$, which gives $3,407,872$ trainable parameters. The adapter matrices B are initialized to zero, while the entries of A are initialized using a Kaiming-uniform distribution He et al. (2015).

The ensemble members are trained using the prescribed regularisation from Section 5.3, corresponding to using the pretrained Mistral-7b model as a prior. We use the Adam optimizer, training for 10 epochs with learning step sizes of $5 \cdot 10^{-6}$ and batch sizes of 8. This learning rate is smaller compared to those in Wang et al. (2023) to better refine the learning curve before the models become overfitted. The token size limit is fixed at 128 for CQA and 512 for MMLU. We performed an ablation study on L2 regularization $\lambda/2 = 0.01, 0.1, 1, 10$, finding optimal performance at $\lambda/2 = 1$. The training focuses solely on the single token output representing the answer, with all other token outputs being masked.

To calculate performance metrics and uncertainty estimates, we reduced the output dimension from 32,000 to 6. Five of these dimensions represent QA choice tokens (a, b, c, d, e), and the sixth aggregates the softmax prediction scores associated with all other tokens. We found that, even before fine-tuning, the pretrained Mistral-7b checkpoint assigns a near-zero probability to the sixths class with very few exceptions. This suggests that the model reliably comprehends the format of multiple-choice questions and answers, consistently producing appropriate tokens.

7.3 Performance metrics

In Fig. 1, we present the performance metrics for LoRA ensembles with both $M = 1$ and $M = 5$ members. The ensembles are trained on multiple-choice questions from the CQA training dataset and evaluated on three validation datasets: CQA (first column), MMLU STEM (second column), and MMLU Social Studies (third column).

The ensemble members start over-fitting after a couple of epochs as can be seen by the validation loss in the second row of panels, but interestingly without considerable drop in accuracy in the first row of panels. As will be seen in more detail in Section 7.4, this signals that the ensemble members become more overconfident on predictions that are wrong, corroborated by the increasing expected calibration error with epoch seen in the bottom row of Fig. 1. This is consistent with the recent reports that fine-tuned LLMs often exhibit overconfidence Jiang et al. (2021); Lin et al. (2022); Xiao et al. (2022b); He et al. (2023); Tian et al. (2023); OpenAI et al. (2023).

For fine-tuning on CQA seen in the first column, the overfitting is significantly reduced for ensembles compared to individual members, as can be seen by the gap between the loss curves for $M = 1$ and $M = 5$ for later epochs. This means that ensemble members output more confident, but different predictions. This is a common feature of Bayesian posteriors Blundell et al. (2015); Zhang et al. (2020); Kristiadi et al. (2020); Ober & Aitchison (2021); Fortuin et al. (2022); Aitchison et al. (2021); Yang et al. (2024). This signals high epistemic uncertainty in this regime, indicating that the validation set is perceived as "unknown" to the model.

7.4 Dynamics of uncertainty metrics

We analyze data-resolved dynamics of uncertainty measures using two-dimensional histogram plots of mutual information and entropy Linander et al. (2023). This approach allows us to effectively disentangle different sources of uncertainty, providing a detailed view of how the model perceives data across the "known-certain," "known-uncertain," and "unknown" regimes during fine-tuning. By visualizing the relationship between mutual information and entropy, we can track how the model's knowledge evolves, identify patterns in data complexity, and assess the model's ability to adapt to fed information. This method not only highlights the dynamics of learning, but also helps identify limitations of the model, offering actionable insights for further improvement.

In Fig. 2, we show the epoch evolution of uncertainty measures for a LoRA ensemble that was fine-tuned and evaluated on the CQA dataset. The mutual information (MI) on the vertical axis, and the predictive entropy (Entropy) on the horisontal axis, are binned over every sample of the validation set of CQA and the number of samples in each bin is indicated by the colorbar. We distinguish between incorrect and correct predictions to enhance interpretability. The ensemble attains optimal validation loss at epoch 3, as shown in Fig. 1. Epochs 1 and 6 exhibit underfitting and overfitting, respectively. The true posterior $p(\theta|D)$ should yield nearly zero mutual information when evaluated on the validation dataset, which should belong to "known" data by design. However, in each of the underfitted, optimal, and overfitted regimes, there is a fraction of validation samples that exhibit a moderate spread in mutual information. Samples with high mutual information indicate that parts of

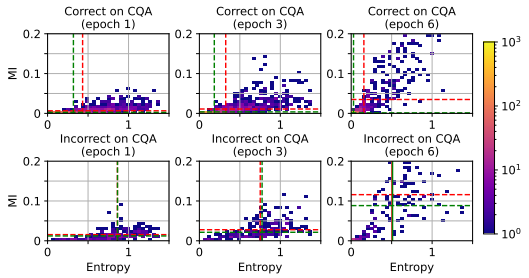

Figure 2: In-distribution evolution of uncertainty during fine-tuning. Histograms of predictive entropy and mutual information for a LoRA ensemble trained and evaluated on the CQA dataset. The ensemble consists of $M = 5$ members, with a LoRA rank of 8, $\alpha = 32$, and an L2 LoRA loss of 1. Evolution of uncertainty measures for in-domain data across training epochs (columns), differentiated by correct (top row) and incorrect (bottom row) predictions, with mean (red) and median (green). Color represents the number count in each bin.

the target domain are perceived as "unknown" by the ensemble, highlighting regions where expanding or refining the training data could most improve performance.

By examining the mean and median of the entropic metrics in Fig. 2, we can observe overall patterns in the validation data perception by the ensemble. Interestingly, the ensemble becomes progressively more confident on correct QAs, reaching its highest confidence once the ensemble overfits, where most correct QAs are predicted with almost zero uncertainty. Although overall performance (as measured by loss shown in Fig. 1) degrades at epoch 6 compared to epoch 3, the model has actually learned the bulk of the validation dataset more effectively. The loss increase arises because the ensemble has also become more confident about incorrect QAs, increasing the cross-entropy. However, almost all of these incorrect predictions are marked by high mutual information, placing them in the "unknown" regime. Since these points can be identified by their high mutual information, they can be filtered out at inference time or used to refine the training procedure (e.g., by including more instances of these data). This result is noteworthy because it shows that the ensemble continues to improve even after reaching its minimum loss at epoch 3. By epoch 6, it performs better on most of the dataset, although its increased confidence in the remaining poorly performing data inflates the overall loss. Nonetheless, these problematic data points remain mostly distinguished by high epistemic uncertainty.

Fig. 3 shows the uncertainty histograms for MMLU STEM and MMLU SS, calculated using an ensemble trained on CQA. Starting with MMLU STEM in the first and second row of panels, we draw two key insights. First, among the wrongly classified samples for the optimal epoch 3 (second row, second column), a large fraction falls into the "known-uncertain" regime, characterized by high predictive entropy but low mutual information. This is in contrast to MMLU SS, which exhibits a more moderate spread of predictive entropy in the corresponding panel (last row, second column). This observation highlights the relatively higher complexity of STEM compared to SS datasets. The ensemble perceives the STEM samples as equally "known" as SS samples (similar mutual information) but assigns them broader predictive distributions, indicating greater "uncertain" characteristics due to a lack of precise knowledge.

Second, continuing with MMLU SS in the third and fourth rows of Fig.3, we observe a significant decrease in predictive entropy as the number of epochs increases, indicating that the ensemble learns features that enable it to classify samples with greater confidence. This behavior is similar to CQA in Fig.2, but the key difference is that, in this case, the data is from an out-of-domain dataset. This demonstrates that the ensemble continues to improve on the majority of MMLU SS samples, despite not being specifically trained on them, suggesting a similarity between the CQA and MMLU SS domains.

Interestingly, the evolution of entropic quantities reveals a key difference between the datasets. For MMLU STEM, there is a notable density of "known-uncertain" samples (high entropy, low mutual information) (see Fig. 3, second row), whereas for MMLU SS, high-entropy samples are more evenly distributed in terms of mutual information. The density of "known-uncertain" MMLU

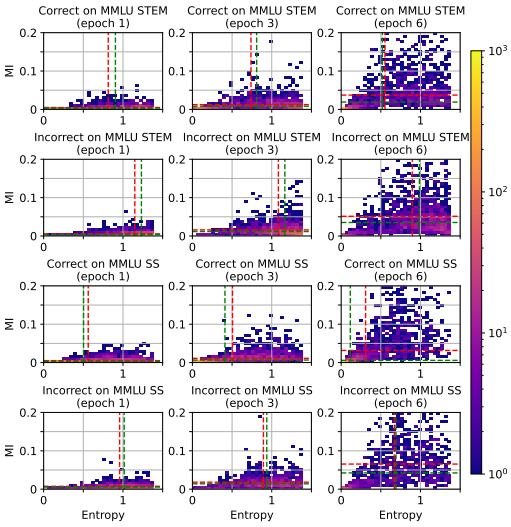

Figure 3: Out-of-distribution evolution of uncertainty during fine-tuning. Histograms of predictive entropy (Entropy) and mutual information (MI) for a LoRA ensemble trained on CQA and evaluated on MMLU STEM (top rows) and MMLU SS (bottom rows) datasets. The ensemble consists of $M = 5$ members, with a LoRA rank of 8, $\alpha = 32$, and an L2 LoRA loss of 1. Evolution of uncertainty measures for out-of-training-domain data across training epochs (columns left to right), differentiated by correct (first and third row) and incorrect (second and forth row) predictions, with mean (red) and median (green).

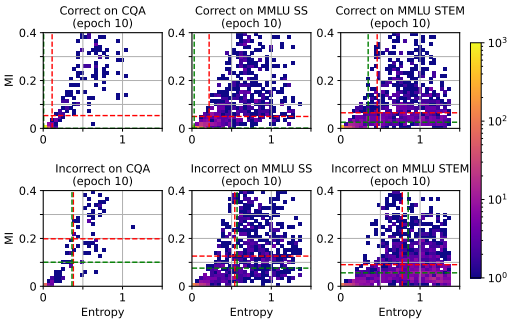

Figure 4: Late overfitting regime (epoch 10). Histograms of predictive entropy (Entropy) and mutual information (MI) for a LoRA ensemble trained on CQA (left column) and evaluated on MMLU SS (middle column) and MMLU STEM (right column) datasets. The ensemble consists of $M = 5$ members, with a LoRA rank of 8, $\alpha = 32$, and an L2 LoRA loss of 1. There is still a significant fraction of samples (lower right panel) in MMLU STEM with low mutual information and high total uncertainty.

STEM samples largely persists despite overfitting (see Fig. 4), suggesting that the fine-tuned model ensemble perceives confusing samples from MMLU STEM as being more consistently "known-uncertain," whereas confusing samples from MMLU SS are forgotten more rapidly, transitioning into the "unknown" regime.

8 Conclusions

We derived a principled, and practically efficient method for posterior approximation of fine-tuned LLMs using ensembles of low-rank adaptation models. Using the posterior approximation, we derived uncertainty quantification using entropic metrics, and followed their evolution during fine-tuning. This enabled us to draw a number of conclusions regarding the balance of retained knowledge and domain specific adaptation. First, the low-rank ensemble shows retained accuracy while also decreasing both predictive uncertainty and mutual information in the overfitting regime. At the same time, incorrect predictions are increasingly correlated with higher mutual information, opening up for inference time filtering and active learning. Second, during fine-tuning towards CQA, there is a distinct difference between the change in uncertainty for MMLU STEM and MMLU social science. For the STEM dataset, there is a large fraction of samples with low mutual information and high predictive entropy, hinting at in-domain but uncertain predictions. This can be connected to the lower prior accuracy of MMLU STEM compared to MMLU SS and CQA, indicating low representative power of the model in this domain. These STEM samples exhibit increased robustness, through broad yet consistent predictive distributions across ensemble members, to a large extent surviving overfitting. As fine-tuning continues to be an important tool in shaping model output Zhang et al. (2023b), it is also increasingly important to understand how the knowledge of the pre-trained model changes as the model is adapted to a target domain. The methods presented here are generally applicable to different fine-tuning scenarios, and provide a systematic treatment of prediction uncertainty for fine-tuned LLMs.

9 IMPACT STATEMENT

This paper advances the field of Machine Learning by enhancing the reliability and understanding of large language models (LLMs) through uncertainty quantification using LoRA ensembles. Our work contributes to the understanding and trustworthiness of LLM predictions, a step forward in AI reliability and applicability.

REFERENCES

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning, 2020.

Laurence Aitchison, Adam Yang, and Sebastian W Ober. Deep kernel processes. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 130–140. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/aitchison21a.html.

Gabriel Y. Arteaga, Thomas B. Schön, and Nicolas Pielawski. Hallucination detection in llms: Fast and memory-efficient fine-tuned models, 2024. URL https://arxiv.org/abs/2409.02976.

Oleksandr Balabanov, Bernhard Mehlig, and Hampus Linander. Bayesian posterior approximation with stochastic ensembles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13701–13711, June 2023.

David Barber and Charles M. Bishop. Ensemble learning in bayesian neural networks. 1998.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1613–1622. JMLR.org, 2015.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Wray L. Buntine and Andreas S. Weigend. Bayesian back-propagation. *Complex Syst.*, 5, 1991.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness, 2023.

Wenlong Chen and Yingzhen Li. Calibrating transformers via sparse gaussian processes, 2024.

Tristan Cinquin, Alexander Immer, Max Horn, and Vincent Fortuin. Pathologies in priors and inference for bayesian transformers, 2021.

John S. Denker, Daniel B. Schwartz, Ben S. Wittner, Sara A. Solla, Richard E. Howard, Lawrence D. Jackel, and John J. Hopfield. Large automatic learning, rule extraction, and generalization. *Complex Syst.*, 1, 1987.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models, 2022.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5 (3):220–235, Mar 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00626-4. URL https://doi.org/10.1038/s42256-023-00626-4.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the uncertainty estimation of large language models, 2023.

Vikranth Dwaracherla, Zheng Wen, Ian Osband, Xiuyuan Lu, Seyed Mohammad Asghari, and Benjamin Van Roy. Ensembles for Uncertainty Estimation: Benefits of Prior Functions and Bootstrapping. *arXiv e-prints*, art. arXiv:2206.03633, June 2022.

Xinjie Fan, Shujian Zhang, Bo Chen, and Mingyuan Zhou. Bayesian attention modules. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16362–16376. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/bcff3f632fd16ff099a49c2f0932b47a-Paper.pdf.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep Ensembles: A Loss Landscape Perspective. *arXiv e-prints*, art. arXiv:1912.02757, December 2019.

Vincent Fortuin, Adrià Garriga-Alonso, Sebastian W. Ober, Florian Wenzel, Gunnar Rätsch, Richard E. Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited, 2022.

Yarin Gal and Zoubin Ghahramani. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. *arXiv e-prints*, art. arXiv:1506.02158, June 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(1):1513–1589, Oct 2023. ISSN 1573-7462. doi: 10.1007/s10462-023-10562-9. URL https://doi.org/10.1007/s10462-023-10562-9.

Adam Gleave and Geoffrey Irving. Uncertainty estimation for language reward models, 2022.

Alex Graves. Practical variational inference for neural networks. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf.

Fredrik K. Gustafsson, Martin Danelljan, and Thomas B. Schön. Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision. *arXiv e-prints*, art. arXiv:1906.01620, June 2019.

Guande He, Jianfei Chen, and Jun Zhu. Preserving pre-trained features helps calibrate fine-tuned language models, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Lara Hoffmann and Clemens Elster. Deep Ensembles from a Bayesian Perspective. *arXiv e-prints*, art. arXiv:2105.13283, May 2021.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling, 2023.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICL2022*, 2021.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models, October 2023. URL http://arxiv.org/abs/2307.10236. arXiv:2307.10236 [cs].

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks, 2020.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Learnable uncertainty under laplace approximations, 2021. URL https://arxiv.org/abs/2010.02720.

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

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. Deup: Direct epistemic uncertainty prediction, 2023. URL https://arxiv.org/abs/2102.08501.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes, 2018.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching Models to Express Their Uncertainty in Words. *arXiv e-prints*, art. arXiv:2205.14334, May 2022. doi: 10.48550/arXiv.2205.14334.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *arXiv e-prints*, art. arXiv:2305.19187, May 2023. doi: 10.48550/arXiv.2305.19187.

Hampus Linander, Oleksandr Balabanov, Henry Yang, and Bernhard Mehlig. Looking at the posterior: accuracy and uncertainty of neural-network predictions. *Machine Learning: Science and Technology*, 4(4):045032, 2023.

Chen Ling, Xujiang Zhao, Wei Cheng, Yanchi Liu, Yiyou Sun, Xuchao Zhang, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, et al. Uncertainty decomposition and quantification for in-context learning of large language models. *arXiv preprint arXiv:2402.10189*, 2024.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

D. J. C. Mackay. Information-based objective functions for active data selection. *Neural Computation*, 4(2):550–604, 1992.

David MacKay. Bayesian model comparison and backprop nets. In J. Moody, S. Hanson, and R.P. Lippmann (eds.), *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper/1991/file/c3c59e5f8b3e9753913f4d435b53c308-Paper.pdf.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks, 2018.

Andrey Malinin and Mark Gales. Uncertainty Estimation in Autoregressive Structured Prediction, February 2021. URL http://arxiv.org/abs/2002.07650. arXiv:2002.07650 [cs, stat].

Andrey Malinin, Liudmila Prokhorenkova, and Aleksei Ustimenko. Uncertainty in gradient boosting via ensembles, 2021.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/tacl_a_00494. URL https://aclanthology.org/2022.tacl-1.50.

Ruijia Niu, Dongxia Wu, Rose Yu, and Yi-An Ma. Functional-level uncertainty quantification for calibrated fine-tuning on llms, 2024. URL https://arxiv.org/abs/2410.06431.

Sebastian W Ober and Laurence Aitchison. Global inducing point variational posteriors for bayesian neural networks and deep gaussian processes. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8248–8259. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ober21a.html.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo

Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, 2019b.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-Distribution Detection and Selective Generation for Conditional Language Models, March 2023. URL http://arxiv.org/abs/2209.15558. arXiv:2209.15558 [cs].

Zhengxiang Shi and Aldo Lipani. Dept: Decomposed prompt tuning for parameter-efficient fine-tuning, 2024.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions, 2024. URL https://arxiv.org/abs/2412.05563.

Meiqi Sun, Wilson Yan, Pieter Abbeel, and Igor Mordatch. Quantifying uncertainty in foundation models via ensembles. In *NeurIPS 2022 Workshop on Robustness in Sequence Modeling*, 2022. URL https://openreview.net/forum?id=LpBlkATV24M.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, 2023.

Tishby, Levin, and Solla. Consistent inference of probabilities in layered networks: predictions and generalizations. In *International 1989 Joint Conference on Neural Networks*, pp. 403–409 vol.2, 1989. doi: 10.1109/IJCNN.1989.118274.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Dustin Tran, Mike Dusenberry, Mark van der Wilk, and Danijar Hafner. Bayesian layers: A module for neural network uncertainty. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/154ff8944e6eac05d0675c95b5b8889d-Paper.pdf.

Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards reliability using pretrained large model extensions, 2022.

Xi Wang, Laurence Aitchison, and Maja Rudolph. Lora ensembles for large language model fine-tuning, 2023.

Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: An alternative approach to efficient ensemble and lifelong learning, 2020.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 7273–7284, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.538. URL https://aclanthology.org/2022.findings-emnlp.538.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis, 2022b.

Boyang Xue, Jianwei Yu, Junhao Xu, Shansong Liu, Shoukang Hu, Zi Ye, Mengzhe Geng, Xunying Liu, and Helen Meng. Bayesian transformer language models for speech recognition. pp. 7378–7382, 06 2021. doi: 10.1109/ICASSP39728.2021.9414046.

Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models, 2024.

Yuanzhao Zhai, Han Zhang, Yu Lei, Yue Yu, Kele Xu, Dawei Feng, Bo Ding, and Huaimin Wang. Uncertainty-penalized reinforcement learning from human feedback with diverse reward lora ensembles, 2023.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023a.

Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning, 2020.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023b.

Shujian Zhang, Xinjie Fan, Bo Chen, and Mingyuan Zhou. Bayesian attention belief networks, 2021.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. *arXiv preprint arXiv:2104.04670*, 2021.

A DATASETS

A.1 SIZE

For MMLU STEM and Social Studies datasets, we use the original test split as the validation set, as in Wang et al. (2023). For CQA dataset we use the standard training and validation splits. The topics in the MMLU dataset are categorized into STEM, Social Studies, Humanities, and Other, following the classification defined in Hendrycks et al. (2021).

Dataset	**Train Size**	**Validation Size**
CQA	9741	1221
MMLU SS	397	3077
MMLU STEM	430	3153

Table 1: Summary of the CQA and MMLU dataset sizes.

A.2 EXAMPLES

The datasets have been preprocessed into a single, consistently formatted prompt question-answer. Below, we present examples of these questions for each of the considered dataset.

Q: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what?

Answer Choices:

(a) bank

(b) library

(c) department store

(d) mall

(e) new york

A: (a).

CommonsenseQA

Q: Which one of the following is the most appropriate definition of a 99% confidence interval?

Answer Choices:

(a) 99% of the time in repeated samples, the interval would contain the true value of the parameter

(b) 99% of the time in repeated samples, the interval would contain the estimated value of the parameter

(c) 99% of the time in repeated samples, the null hypothesis will be rejected

(d) 99% of the time in repeated samples, the null hypothesis will not be rejected when it was false

A: (a).

MMLU SS

Q: Find all c in Z_3 such that Z_3[x]/(x^2 + c) is a field.

Answer Choices:

(a) 0

(b) 1

(c) 2

(d) 3

A: (b).

MMLU STEM