Quantifying Uncertainty in Foundation Models via Ensembles

Meiqi Sun[1] **Wilson Yan**[1] **Pieter Abbeel**[12] **Igor Mordatch**[13]

1 Introduction

As large pre-trained foundation models begin to proliferate and have increasing impact in real-world applications, it is of utmost importance to guarantee that these models are trustworthy and reliable. An important capability towards this is for models to *know what they don't know* - models that are capable of quantifying uncertainty about their own outputs (potentially on inputs very different from what they were trained on). Uncertainty quantification can give guidance on when to trust the model. It can also lead to new capabilities, such as active learning (Settles, 2009) or exploration behavior in decision-making agents (Baranes & Oudeyer, 2013) driven by the goal of reducing uncertainty.

While past research investigated uncertainty quantification single number outputs (Gleave & Irving, 2022) or calibration (Jiang et al., 2021; Kadavath et al., 2022), we still lack a general investigation of uncertainty quantification for multi-token, open-ended outputs under inputs that can be very different from the training distribution.

We aim to fill this gap with our work by conducting a systematic study of uncertainty quantification spanning multiple tasks - a synthetic string task, a natural language arithmetic and a question-answering task. For each of the tasks, we construct a progression of inputs increasingly different from the model's training distribution. We expect model's uncertainty about its outputs under such inputs to increase, indicating that the outputs should not be trusted in such settings. Note that this evaluation is different from the question of model calibration: calibration asks whether the model is less certain about wrong answers over correct answers, but the questions always come from the training distribution.

We propose disagreement of model ensembles as an effective and compute-efficient method to quantify uncertainty. The idea is to train multiple models and ask them to make predictions - if predictions disagree, the system is not certain in the prediction. The method has been extensively applied in many areas of machine learning (Lakshminarayanan et al., 2017; Christiano et al., 2017; Vyas et al., 2018; Lowrey et al., 2018; Beluch et al., 2018), but so far has been relatively uncommon in foundation model research.

We investigate the behavior of such ensemble disagreement in foundation (specifically language) models and find that it is a great fit for this setting. One challenge with applying this method directly to foundation models is that pre-training a large ensemble of them is very expensive (training even a single foundation model can be a multi-million dollar investment). However, we primarily use foundation models for fine-tuning to specific tasks of interest. Fine-tuning is much cheaper as datasets are much smaller, training is shorter, or most model parameters can be frozen (Lu et al., 2021). Thus, it is relatively cheap to build an ensemble of fine-tuned models that will quantify uncertainty. Importantly, the uncertainty will be about fine-tuning task, not pre-training task - and because the former has less data it is usually what presents reliability concerns.

Our study looks at behavior of fine-tuned ensembles stemming from a pre-trained GPT2 language model (Radford et al., 2019). We find that for single token generation, ensemble disagreement is a very good predictor of model uncertainty. For generations that are multi-token sentences, we find

[1]UC Berkeley [2]Covariant [3]Google Brain Correspondence to: meiqisun@berkeley.edu

Workshop on Robustness in Sequence Modeling, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

a more subtle relationship: for inputs very different from training distribution, total likelihood of a sequence becomes very low and is thus acts as a good predictor of uncertainty. However, for inputs that are not very far from the training distribution, total likelihood is similar to test set likelihoods (and thus is not a reliable predictor of uncertainty), yet we observe high ensemble disagreement in this case. We also find that an ensemble of models is crucial and that estimating uncertainty based on disagreement of multiple samples generated by one model does not lead to accurate predictions.

To summarize, our contributions are twofold. Firstly, we propose a benchmark to evaluate uncertainty quantification spanning multiple tasks and progression of increasingly out of distribution inputs. Secondly, we propose an argument supported by evaluation results for disagreement of ensembles of fine-tuned models as both a compute-efficient and effective way to measure uncertainty in pre-trained foundation models. We hope that our investigation and results encourage more research in the area of uncertainty quantification and exploration in foundation models and the use of fine-tuned ensembles as a tool in this research.

2 Method

Learning Ensembles of Foundation Models For each downstream language task, we construct an ensemble of $M = 10$ models by independently fine-tuning M pretrained GPT-2 models. We use different random seeds for each training run so that fine-tuned models do not converge to the same solution, an important factor when using our ensemble to quantify uncertainty.

Uncertainty Quantification through Ensemble Disagreement We propose a simple yet effective method to quantify uncertainty through evaluating ensemble disagreement. Specifically, for a given text query, we generate a sample for each model in the ensemble and compute the mode answer as the most commonly sampled answer. In the case where there is more than one mode answer, we randomly choose one of the modes. We compute ensemble disagreement by calculating the standard deviation of the likelihoods of the sampled mode answer for all models in the ensemble. For proper evaluation, we compare our method against several alternatives using a *single* fine-tuned foundation model. All three investigated methods are summarized as follows

1) **Single Model**: We generate a single sample and computes its likelihood

2) **Single Model Multiple Samples**: We generate 10 samples and compute the mean and standard deviation of the likelihoods for each sample

3) **Model Ensemble (ours)**: We generate 1 sample for each model in the ensemble, and compute the mean and standard deviation of the likelihoods for the *mode* sample

3 Results

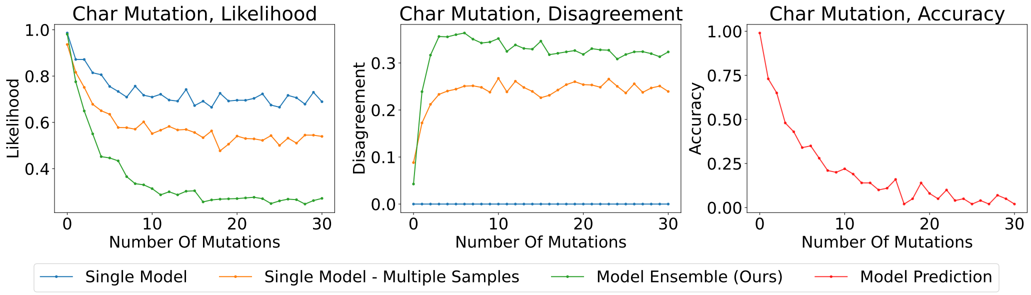

Figure 1: Results for Character Mutation

3.1 Character Mutation Task

Task Description We construct a simple synthetic language task through the process described as follows:

1) Initialize an alphabet sequence: `"abcdefghijklmnopqrstuvwxyz"`
2) Replace a random letter with an underscore
3) Define the expected answer as the replaced letter (the next token for GPT-2 to generate)

During evaluation, we construct out-of-distribution variants of our synthetic task by introducing an increasing number of mutations in the data. Specifically, we define the following modified process:

1) Initialize an alphabet sequence: `"abcdefghijklmnopqrstuvwxyz"`
2) Replace a random letter with an underscore
3) *Excluding* the originally replaced letter, replace N random letters each with another random letter

where N is the number of mutations, ranging from 0 to 30

Results Figure 1 shows that model prediction accuracy decreases significantly as the number of mutations increases. As the number of mutations increases, all three scenarios lead to a decrease in overall likelihood, suggesting that they, to a certain degree, quantify increasing levels of uncertainty. However, the likelihood generated by the ten model ensemble drops faster than the single model comparisons and more accurately reflects the true accuracy. Thus, ensembles are more sensitive to the change in its level of confidence as the challenges grow harder and harder.

Contrary to likelihood, a sharp increase in ensemble disagreement is observed as the number of mutations increases. Here, single model single answer always has a disagreement of 0. All models start off with a similarly low levels of disagreement, yet only the ensemble approach more accurately reflects the drop in model prediction accuracy through a sharp increase in its disagreement. This suggests that the disagreements amongst an ensemble can more smoothly capture different levels of uncertainty to out-of-distribution data.

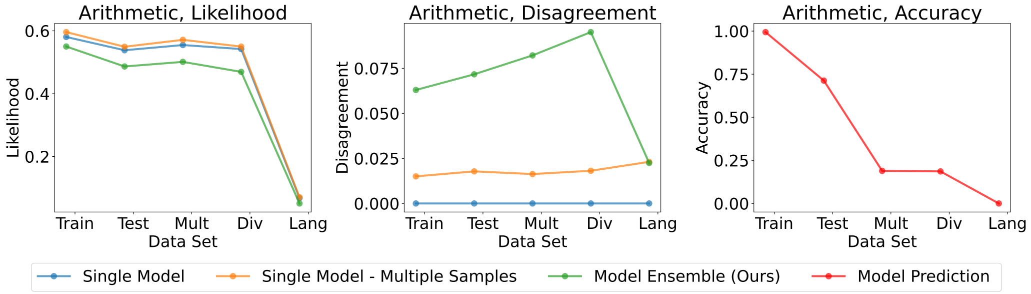

Figure 2: Results for Deepmind Arithmetic

3.2 Deepmind Arithmetic Task

Task Description We train on the addition and subtraction subset of the Deepmind Arithmetic Task (Saxton et al., 2019). Dataset examples consist of language-dependent addition or subtraction tasks such as ‘‘`What is` $3446185 + 0.005678116$?’’ and ‘‘`Subtract 3 from` -0.036713428’’. During evaluation, we consider five categories of examples to our models, representing varying degrees of out-of-distribution data:

1) **Training Set**: Selected addition and subtraction questions from the training set
2) **Test Set**: Selected addition and subtraction questions from the unseen test set

3) **Multiplication**: Modified examples from the training set with operators replaced to specify multiplication
4) **Division**: Modified examples from the training set, with operators replaced to specify division
5) **Language**: Completely different task of question-answer trivia

Results Figure 2 shows quantitative evaluation for the arithmetic task. We see that the single models have similar outputs regardless of the number of samples we output with both lines corresponding to the single models staying very close to one another. In the ensemble disagreement plot, despite facing tasks with different out-of-distribution levels (leading to very different levels of uncertainty), the single model approach always generates a similar level of ensemble disagreement. The ensemble approach, however, is more effective: showing a decrease in likelihood as the tasks become harder, in addition to a sharp increase in ensemble disagreement.

It is worth noting that when the overall likelihood is relatively high (i.e. in the first four groups of datasets), likelihood alone may not be enough for quantifying uncertainty, as overall likelihood is relatively unchanged across the first four categories (in-distribution and slightly out-of-distribution). However, with additional information from the ensemble disagreement, we attain a better metric for uncertainty quantification. The sharp increase in standard deviation as we move from one task to another informs increasing model uncertainty. In cases where the likelihood of the models are very low (i.e. when the task is very out of distribution), ensemble disagreement is also low, meaning that the models are confused about the question, and choose to express “I don’t know”.

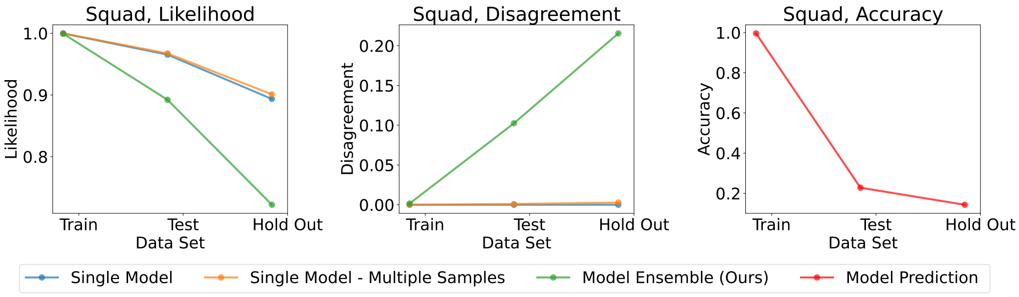

Figure 3: Results for SQuAD

3.3 SQuAD (The Stanford Question Answering Dataset) Task

Task Description Lastly, we consider the SQuAD dataset (Rajpurkar et al., 2016), a dataset of trivia questions with (Question-Context, Answer) pairs, in which the question answers can be found in the provided context.

During evaluation, we test ensemble performance on data of varying out-of-distributions levels, consisting of the **Train Set**, the **Test Set** (similar topics as the training set), and a **Hold Out Set** with unseen topics.

Results Figure 3 shows quantitative results on SQuAD evaluations. We see similar performances between the baseline one model one sample and one model ten samples. The drop in likelihood is not obvious, and the model disagreement is almost unchanged between data sets.

On the other hand, the ensemble is more informative. From the likelihood plot, there is a sharp drop as evaluation data becomes more out-of-distribution. On the model disagreement graph, there are sharp increases in ensemble disagreement less familiar questions and topics. The combined results from the ten distinctive models help us certify that the model is very uncertain about the results it generated for the hold out set (and their accuracy indeed turned out to be low), and the model is fairly certain about samples from training (and their prediction accuracy for training is very high). The higher information gain from the ensemble over a single model makes us believe that creating an ensemble is both necessary and worthwhile.

References

Baranes, A. and Oudeyer, P.-Y. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013.

Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9368–9377, 2018.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Gleave, A. and Irving, G. Uncertainty estimation for language reward models. *arXiv preprint arXiv:2203.07472*, 2022.

Jiang, Z., Araki, J., Ding, H., and Neubig, G. 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, 09 2021. doi: 10.1162/tacl_a_00407.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z. H., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lowrey, K., Rajeswaran, A., Kakade, S., Todorov, E., and Mordatch, I. Plan online, learn offline: Efficient learning and exploration via model-based control. *arXiv preprint arXiv:1811.01848*, 2018.

Lu, K., Grover, A., Abbeel, P., and Mordatch, I. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 2021.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Saxton, D., Grefenstette, E., Hill, F., and Kohli, P. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*, 2019.

Settles, B. Active learning literature survey. 2009.

Vyas, A., Jammalamadaka, N., Zhu, X., Das, D., Kaul, B., and Willke, T. L. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 550–564, 2018.