Amortized Bayesian Meta-Learning for Low-Rank Adaptation of Large Language Models

Liyi Zhang
Princeton University
Department of Computer Science
zhang.liyi@princeton.edu

Jake Snell
Princeton University
Department of Computer Science
jsnell@princeton.edu

Thomas L. Griffiths
Princeton University
Department of Psychology and Computer Science
tomg@princeton.edu

Abstract

Fine-tuning large language models (LLMs) with low-rank adaptaion (LoRA) is a cost-effective way to incorporate information from a specific dataset. However, it is often unclear how well the fine-tuned LLM will generalize, i.e., how well it will perform on unseen datasets. Methods have been proposed to improve generalization by optimizing with in-context prompts, or by using meta-learning to fine-tune LLMs. However, these methods are expensive in memory and computation, requiring either long-context prompts or saving copies of parameters and using second-order gradient updates. To address these challenges, we propose Amortized Bayesian Meta-Learning for LoRA (ABMLL). This method builds on amortized Bayesian meta-learning for smaller models, adapting this approach to LLMs while maintaining its computational efficiency. We reframe task-specific and global parameters in the context of LoRA and use a set of new hyperparameters to balance reconstruction accuracy and the fidelity of task-specific parameters to the global ones. ABMLL provides effective generalization and scales to large models such as LLAMA3-8B. Furthermore, as a result of using a Bayesian framework, ABMLL provides improved uncertainty quantification. We test ABMLL on Unified-QA and Crossfit datasets and find that it outperforms existing methods on these benchmarks in terms of both accuracy and expected calibration error.

1 Introduction

Large language models (LLMs) handle a variety of tasks reasonably well (Radford et al., 2019). However, to tailor LLMs to specific domains, fine-tuning on specific datasets is often necessary. While methods such as low-rank adaptation (LoRA; Hu et al. (2021)) fine-tune a pretrained LLM cost-effectively, a fine-tuned LLM is limited to the domain it is trained on. Its performance may not improve in other domains and sometimes worsens as it suffers from catastrophic forgetting. Such catastrophic forgetting may result in overfitting and erasing existing capabilities of the pretrained LLM (Lazaridou et al., 2021; Luo et al., 2023).

Meta-learning is a strategy for solving this problem, training models on a variety of tasks in a way that supports generalization across tasks (Finn et al., 2017). However, meta-learning typically requires a large amount of computation and memory, making it challenging to apply to LLMs. One form of meta-learning that has been applied to LLMs involves fine-tuning models on in-context prompt-response examples (Min et al., 2022; Chen et al., 2022). Another more traditional approach, MAML-en-LLM (Sinha et al., 2024), adapts the Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) framework to LLMs. However, both methods are limited in the size of the language models that can be used: the former requires long-context prompts, whereas the latter uses second-order gradient updates and saves a model for each task.

Recent work on Amortized Bayesian Meta-Learning (ABML; Ravi and Beatson (2019)) addresses some of the computation and memory requirements of meta-learning. This approach posits a generative model over parameters where task-specific parameters are generated from global parameters, and inference over task-specific parameters is amortized. In other words, the conditional distribution over task-specific parameters is shared across tasks, implying that computation and memory costs stay constant with respect to the number of tasks. This approach thus offers a path towards efficient meta-learning for LLMs. However, several challenges exist. First, we need to specify the generative model over weight space in the context of LLMs. Second, the prior term used in ABML no longer adapts to the setting of fine-tuning a pretrained model because the spread of its weights mismatches that of an arbitrary prior used to train a model from scratch. Third, the enormous size

Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025), pages 194–199

of LLMs makes training difficult, as the scale of probabilities assigned to the model variables can overwhelm the influence of the data likelihood.

In this paper, we present a solution to these problems, taking a Bayesian approach to fine-tuning LLMs using ABML. To define the underlying generative model and efficiently characterize the distributions involved, we use LoRA to express both the model weights and their uncertainty. We introduce a new prior over global variables that accounts for the spread of the parameters learned in the pretrained model. We also introduce two adjustable hyperparameters that balance reconstruction accuracy and the fidelity of task-specific parameters to the global ones.

Using amortized Bayesian meta-learning for LLM fine-tuning, we achieve both higher accuracy and better uncertainty estimation over unseen tasks compared with regular fine-tuning and other scalable methods in the meta-learning literature. Figure 1 illustrates an example where incorporating uncertainty estimation in fine-tuning leads to a more calibrated model response. Our method is scalable and avoids the computation and memory overhead of other meta-learning approaches, making it adaptable to larger models such as LLAMA 3 8B. We show that amortized Bayesian meta-learning provides fine-tuned LLMs that are accurate on domain-specific tasks, more generalizable to new tasks, and provide better uncertainty estimation.

2 Related Work

Meta-learning methods in LLMs. Extensive work has explored meta-learning for generalization, typically adopted for models in the pre-LLM era (Finn et al., 2017; Snell et al., 2017; Ravi and Beatson, 2019; Nichol et al., 2018). Sinha et al. (2024) adapted Model-Agnostic Meta-Learning (MAML), developed in Finn et al. (2017), to LLMs. However, this adaptation is more expensive in computation and memory than our method, requiring second-order gradient updates and saving a model for each task. More recently, Kim and Hospedales (2025) proposes a heirarchical Bayesian approach to LoRA meta-learning, but its parameters also increase linearly with number of tasks. As a result, we evaluate on larger models than those tried in these two papers.

As a different approach, Min et al. (2022) and Chen et al. (2022) explored meta-learning for LLMs using in-context learning. These works show that it is possible to fine-tune LLMs on in-context examples and achieve generalization. However, our approach does not require curation of such examples, does not place constraints on the size of the context window of a model, and is more scalable.

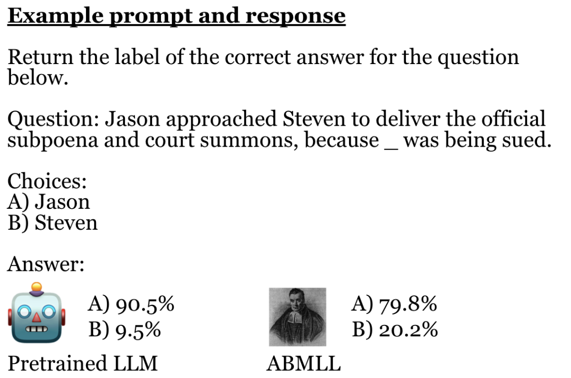


Figure 1: An example where better uncertainty calibration leads to a more reasonable response. This is a prompt and response from an unseen dataset, coming from a pretrained LLM (left) and an LLM fine-tuned with ABMLL (right), with both being updated with 10 gradient steps on other examples of this dataset as in the meta-learning literature. The label is B), so both LLMs are incorrect, but the question is ambiguous: it could interpreted as either Jason "asked" Steven to deliver, or Jason "came to" Steven to deliver, resulting in different answers. ABMLL results in a more calibrated response.

Uncertainty representation for LLMs. Approaches to capturing uncertainty for LLMs can rely on the intrinsic representation of uncertainty in the model or focus on capturing extrinsic uncertainty about model parameters. Intrinsic approaches produce better uncertainty calibration via prompt engineering and sampling (Gruver et al., 2023) or learning an external model (Shen et al., 2024). Extrinsic approaches include using fine-tuning methods to incorporate uncertainty, such as training LoRA with ensembles (Balabanov and Linander, 2024), Laplace approximation (Yang et al., 2023), and variational inference (Wang et al., 2024). Our work takes the extrinsic approach but differs from existing approaches by using the meta-learning setting to achieve strong uncertainty calibration through generalization across datasets.

3 Background

3.1 Low-Rank Adaptation (LoRA)

LoRA (Hu et al., 2021) fine-tunes LLM weights on a low-rank space to improve efficiency compared with regular fine-tuning. Let $\mathbf{W}_0$ of size d_{out}-by-d_{in} denote a weight matrix from a pretrained LLM. Let

$\mathbf{x}$ denote the input to $\mathbf{W}_0$, and $\mathbf{z}$ denote the output of $\mathbf{W}_0$, LoRA fine-tunes pretrained weight $\mathbf{W}_0$ by adding perturbation on the low-ranked space,

$$\mathbf{z} = (\mathbf{W}_0 + \Delta\mathbf{W}_0)\mathbf{x} = (\mathbf{W}_0 + \mathbf{BA})\mathbf{x}.$$

The trainable matrices $\mathbf{B}$ and $\mathbf{A}$ are known as *LoRA adapters*. The sizes of $\mathbf{B}$ and $\mathbf{A}$ are d_{out}-by-d_{rank} and d_{rank}-by-d_{in}, respectively, with d_{rank} being significantly smaller than the original dimensions. Therefore, the number of parameters to be updated are $(d_{\text{out}} + d_{\text{in}})d_{\text{rank}}$, significantly fewer than the original $d_{\text{out}}d_{\text{in}}$.

3.2 Amortized Bayesian Meta-Learning

Amortized Bayesian Meta-Learning (ABML) Ravi and Beatson (2019) improves upon MAML-based meta-learning frameworks by representing uncertainty with a Bayesian approach. It also amortizes inference over the parameters so that memory no longer increases linearly with the number of tasks.

Let θ denote global parameters such that a few steps of gradient descent will produce local parameters ϕ_i on task i with dataset D_i. ABML treats θ as random variables, and minimizes a negative evidence lower bound using variational inference,

$$\text{argmin}_\theta \Big[\sum_{i=1}^{M} -E_{q_\theta(\phi_i|D_i)}[\log p(D_i|\phi_i)] + \text{KL}\big(q_\theta(\phi_i|D_i)\big|\big|p(\phi_i|\theta)\big)\Big] + \text{KL}(q(\theta)||p(\theta)). \quad (1)$$

The variational distribution $q_\theta(\phi_i|D_i)$ is represented by the Gaussian distribution $N(\mu_\phi, \sigma_\phi^2)$ with μ_ϕ, σ_ϕ as trainable parameters.

4 Method

Our method extends Amortized Bayesian Meta-Learning, making it possible to apply to LLMs. This approach combines the advantages of meta-learning for adapting to new tasks with Bayesian inference for uncertainty representation.

We use the the objective of Eq. 1 from ABML. In our setting, θ and ϕ_i are the global and task-specific model parameters produced as the output of LoRA adapters. On a high level, the generative process is

$$\theta \sim p(\theta),$$
$$\phi_i \sim p(\phi_i|\theta),$$
$$D_i \sim \text{LLM}(\phi_i),$$

Algorithm 1 One epoch in the ABMLL algorithm. The "test section" does not need to be performed every epoch.

Input: Likelihood model $p(D_i|\phi_i)$, prior $p(\theta)$ and $p(\phi|\theta)$, variational posterior $q_\theta(\phi_i|D_i)$, with trainable parameters $\mathbf{B}, \mathbf{A}$; constant c, β; number of tasks M and inner-loop size K.

Training section
for task $i \in \{1, 2, ..., M\}$ **do**
  Draw batch D_i from task i dataset.
  Inner-loop:
  for iter $k \in \{1, 2, ..., K\}$ **do**
    Run a step gradient descent to minimize w.r.t. ϕ_i: $-E_{q_\theta(\phi_i|D_i)}[\log p(D_i|\phi_i)] + \beta\text{KL}\big(q_\theta(\phi_i|D_i)\big|\big|p(\phi_i|\theta)\big)$.
  end for
  Outer-loop: Run a step gradient descent to minimize w.r.t. θ: $-E_{q_\theta(\phi_i|D_i)}[\log p(D_i|\phi_i)] + \beta\text{KL}\big(q_\theta(\phi_i|D_i)\big|\big|p(\phi_i|\theta)\big) + \beta\text{KL}(q(\theta)||p(\theta))$.
end for
Test section
Take unseen task i. Create a copy of the above weights, and on the new weights:
for iter $k \in \{1, 2, ..., K\}$ **do**
  Draw batch D_i from task i dataset.
  Run a step gradient descent to minimize w.r.t ϕ_i: $-E_{q_\theta(\phi_i|D_i)}[\log p(D_i|\phi_i)] + \beta\text{KL}\big(q_\theta(\phi_i|D_i)\big|\big|p(\phi_i|\theta)\big)$.
end for
Evaluate on rest of data in task i.
Delete the weights copy and reload the weights at the end of training section.
Output: $\mathbf{B}$, $\mathbf{A}$.

where i represents any task i, and $\text{LLM}(\phi_i)$ denotes the LLM considered as a probabilistic model that takes ϕ_i as its weights and outputs token sequences with joint probabilities defined by the LLM's autoregressive predictive distribution. We provide a pseudocode, Algorithm 1, to illustrate our approach. For any LLM layer with pretrained weights $\mathbf{W_0}$, the quantities for our extension to ABML are:

$$\mu_\theta = \mathbf{B}_{\mu_\theta}\mathbf{A}_{\mu_\theta},$$
$$\log\sigma_\theta^2 = \mathbf{B}_{\sigma_\theta}\mathbf{A}_{\sigma_\theta} + c\mathbf{I},$$
$$\mu_\phi = \mathbf{B}_{\mu_\phi}\mathbf{A}_{\mu_\phi},$$
$$\log\sigma_\phi^2 = \mathbf{B}_{\sigma_\phi}\mathbf{A}_{\sigma_\phi} + c\mathbf{I},$$

$$
\begin{aligned}
p(\phi_i|\theta) &= N(\phi_i; \mu_\theta + \mathbf{W_0}, \sigma_\theta^2), \\
q_\theta(\phi_i|D_i) &= N(\phi_i; \mu_\phi + \mathbf{W_0}, \sigma_\phi^2), \\
p(\theta) &= p(\mu_\theta, \sigma_\theta) \\
&= N(\mu_\theta; 0, \mathbf{I}) \cdot \mathrm{Gamma}(\frac{1}{\sigma_\theta^2}; a_0, b_0), \\
\mathrm{KL}(q(\theta)||p(\theta)) &= -\log p(\theta).
\end{aligned}
$$

Lastly, $p(D_i|\phi_i)$ is defined as the joint probability assigned to D_i where the LLM takes ϕ_i as its weights. The trainable parameters are the LoRA adapters $\mathbf{A}$ and $\mathbf{B}$. However, we introduce four pairs of these adapters to compute both the mean and variance of the LoRA outputs on local and global model weights. $\mathbf{I}$ is identity matrix, and c is a hyperparameter constant dependent on the spread of pretrained LLM weights. a_0 and b_0 are hyperparameters, and the simplification of the KL term as $-\log p(\theta)$ follows Ravi and Beatson (2019).

Balancing the reconstruction error. LLMs are often overparameterized. As a result, probabilistic quantities on the space of weights, $\mathrm{KL}\big(q_\theta(\phi_i|D_i)\big|\big|p(\phi_i|\theta)\big)$ and $\mathrm{KL}(q(\theta)||p(\theta))$, can overwhelm quantities on the data space, $\log p(D_i|\phi_i)$. $\beta-$VAE (Higgins et al., 2016) and Bayesian neural network approaches by Trinh et al. (2022) introduce hyperparameters to temper the likelihood versus regularization terms. Inspired by this idea, we introduce hyperparameters β, γ, resulting in the following objective,

$$
\begin{aligned}
\mathrm{argmin}_\theta \Big[\sum_{i=1}^{M} -E_{q_\theta(\phi_i|D_i)}[\log p(D_i|\phi_i)] + \\
\beta \mathrm{KL}\big(q_\theta(\phi_i|D_i)\big|\big|p(\phi_i|\theta)\big)\Big] + \gamma \mathrm{KL}(q(\theta)||p(\theta)).
\end{aligned} \tag{2}
$$

This provides a flexible way to control how close the global parameters θ are to the prior $p(\theta)$, and how close the task-specific parameters ϕ_i are to θ.

5 Empirical Evaluations

Model and datasets. We fine-tune LLAMA3-8B on CrossFit (Ye et al., 2021) and UnifiedQA (Ye et al., 2021), textual datasets commonly used to train meta-learning models. Because a key aim of our paper is to study uncertainty quantification, we filter for multiple choice datasets, leading to a subset of CrossFit and UnifiedQA with 34 datasets with 68K training datapoints in total. They feature problems such as sentiment analysis, natural language inference, and identifying particular traits or topics in a given text. For evaluation on an unseen task, we use Winogrande (Sakaguchi et al., 2021), a multiple choice dataset evaluating common sense reasoning.

Metrics. We use accuracy to evaluate general performance and expected calibration error (ECE) to evaluate uncertainty estimation.

Implementation details. All methods use batchsize of 2 and inner-loops with 5 gradient steps. LoRA adapters follow standard practice with rank $= 8$, and learning rate is tuned in $[10^{-6}, 5 \cdot 10^{-5}]$. For ABMLL, $\beta = 5 \cdot 10^{-10}$, $\gamma = 10^{-6}$, $c = e^{-20}$. For the gamma prior, $a_0 = 1, b_0 = 0.01$, following Ravi and Beatson (2019). During validation on the unseen dataset, all models train 10 gradient steps on 10 batches from this dataset and evaluate on the rest.

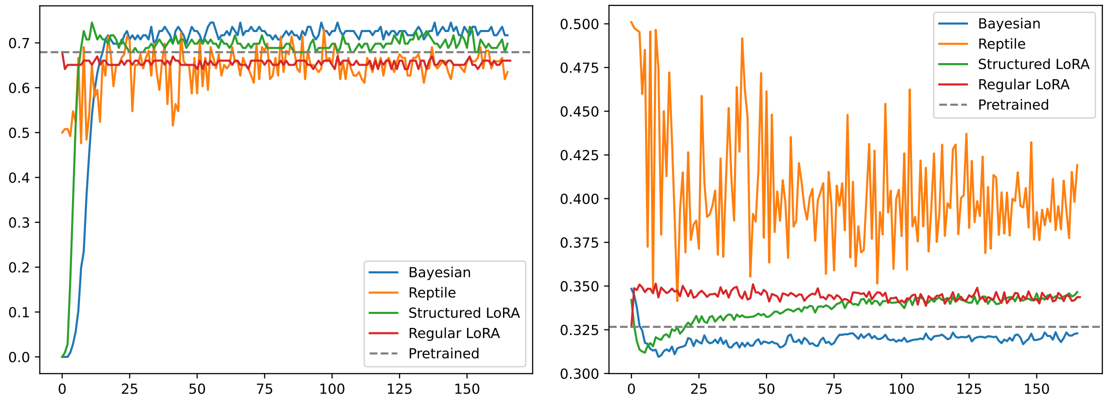
Baselines. We use four baseline methods that can viably scale to LLAMA3-8B. *Pretrained* is the off-the-shelf LLM. *Regular LoRA* is the default LoRA method trained on the whole randomly shuffled training dataset. *Structured LoRA* also uses the default LoRA, but the training dataset follows the same "structure" as our method: it is iteratively trained 5 gradient steps on one task at a time. Thus, it tests the effect of our generative model on performance. The *Reptile* (Nichol et al., 2018) algorithm primarily uses a weighted average between new weights and previous weights to achieve meta-learning.

Results. Figure 2 shows validation accuracy and ECE over epochs across methods. We observe that ABMLL consistently achieves higher accuracy. On ECE, ABMLL also consistently achieves the best performance, whereas structured LoRA worsens as training continues. For fairness, the result at the end of every sixth epoch is reported from regular LoRA, because both ABMLL and Reptile run six instead of one gradient step during each epoch's training.

Table 1 reports the best validation score for each model from three random seeds, showing a statistically significant advantage for ABMLL.

Conclusion

Meta-learning is an effective method for supporting better generalization across datasets, but its demands on computation and memory can make it difficult to apply to large language models. We have shown how meta-learning can be used to adapt LLMs by combining Amortized Bayesian



(a) Accuracy.

(b) Expected calibration error (lower is better).

Figure 2: Validation accuracy and ECE on the vertical axis over epochs on the horizontal axis across our method (ABMLL) and four benchmarks. On accuracy, ABMLL consistently achieves higher. On ECE, ABMLL also consistently achieves the best performance, whereas structured LoRA, the second best performer on accuracy, worsens on uncertainty calibration as training continues.

Table 1: Validation accuracy and ECE across three random seeds, with standard error.

Method	Accuracy ↑	ECE ↓
Pretrained	$68.2\% \pm 0.3\%$	0.327 ± 0.000
Regular LoRA	$68.2\% \pm 0.3\%$	0.327 ± 0.000
Structured LoRA	$73.6\% \pm 0.6\%$	0.320 ± 0.001
Reptile	$73.5\% \pm 0.2\%$	0.370 ± 0.005
ABMLL	$\mathbf{74.8\% \pm 0.3\%}$	$\mathbf{0.317 \pm 0.001}$

Meta-Learning with Low-Rank Adaptation. This approach results not just in better accuracy across several benchmarks, but also in better calibration.

Limitations

One limitation of the paper is the scope of empirical evaluation regarding datasets and models. While the datasets feature natural text that can occur in the real world, it would be beneficial to evaluate on more test datasets to confirm the method's consistency. Additionally, the paper's method can be naturally extended to other models, so evaluating on more models would be a reasonable venue for future work.

As a meta-learning method, our approach must be trained on datasets that can be naturally divided into different tasks, a requirement that is not always available to practioners seeking significant model improvement on one particular domain.

While our empirical results suggest that our approach provides more accurate and calibrated responses, theoretical convergence is not guaranteed due to the need for approximate inference and various design choices, including limitations of the variational family $q_\theta(\phi_i|D_i)$.

Acknowledgments

We thank Qinyuan Ye for helpful discussion. We also thank the anonymous reviewers for helpful comments. This work was supported by grant N00014-23-1-2510 from the Office of Naval Research.

References

Oleksandr Balabanov and Hampus Linander. 2024. Uncertainty quantification in fine-tuned llms using lora ensembles. *ArXiv*, abs/2402.12264.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730. Association for Computational Linguistics.

Chelsea Finn, P. Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016.

beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Minyoung Kim and Timothy M Hospedales. 2025. Lift: Learning to fine-tune via bayesian parameter efficient meta fine-tuning. The Thirteenth International Conference on Learning Representations, 2025.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Giménez, Cyprien de Masson d'Autume, Tomás Kociský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Neural Information Processing Systems*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *ArXiv*, abs/2308.08747.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *ArXiv*, abs/1803.02999.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sachin Ravi and Alex Beatson. 2019. Amortized bayesian meta-learning. In *International Conference on Learning Representations*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory W. Wornell, and Soumya Ghosh. 2024. Thermometer: Towards universal calibration for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 44687–44711. PMLR.

Sanchit Sinha, Yuguang Yue, Victor Soto, Mayank Kulkarni, Jianhua Lu, and Aidong Zhang. 2024. Maml-en-llm: Model agnostic meta-training of llms for improved in-context learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 2711–2720. Association for Computing Machinery.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NeurIPS'17, page 4080–4090, Red Hook, NY, USA. Curran Associates Inc.

Trung Trinh, Markus Heinonen, Luigi Acerbi, and Samuel Kaski. 2022. Tackling covariate shift with node-based bayesian neural networks. In *International Conference on Machine Learning*, pages 21759–21774. PMLR.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris N. Metaxas, and Hao Wang. 2024. BLob: Bayesian low-rank adaptation by backpropagation for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. 2023. Bayesian low-rank adaptation for large language models. *ArXiv*, abs/2308.13111.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189. Association for Computational Linguistics.