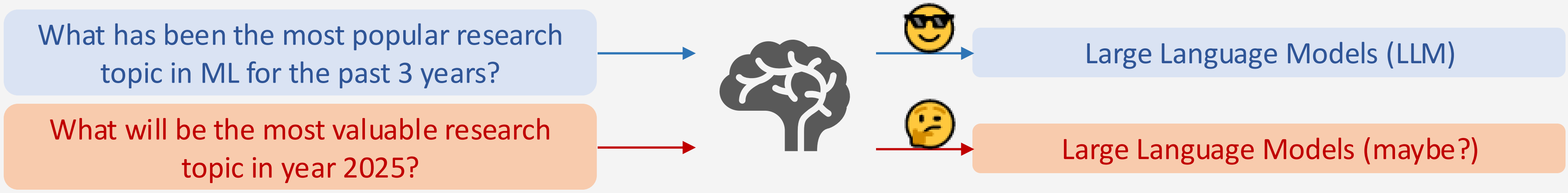
Training-Free Bayesianization for Low-Rank Adapters of Large Language Models

Haizhou Shi[*1], Yibin Wang[*2], Ligong Han[3], Huan Zhang[2], Hao Wang[1]

[*]Equal Contribution [1]Rutgers University [2]UIUC [3]MIT-IBM Waston AI Lab



Accurate Estimation of Uncertainty is Crucial to Trustworthy LLMs!!!

Content

- Background: Uncertainty Estimation of Large Language Models
  - Verbalized Uncertainty for Generation
  - Uncertainty Estimation for Adaptation (Classification)

- Training-Free Bayesianization for Low-Rank Adapters of LLMs

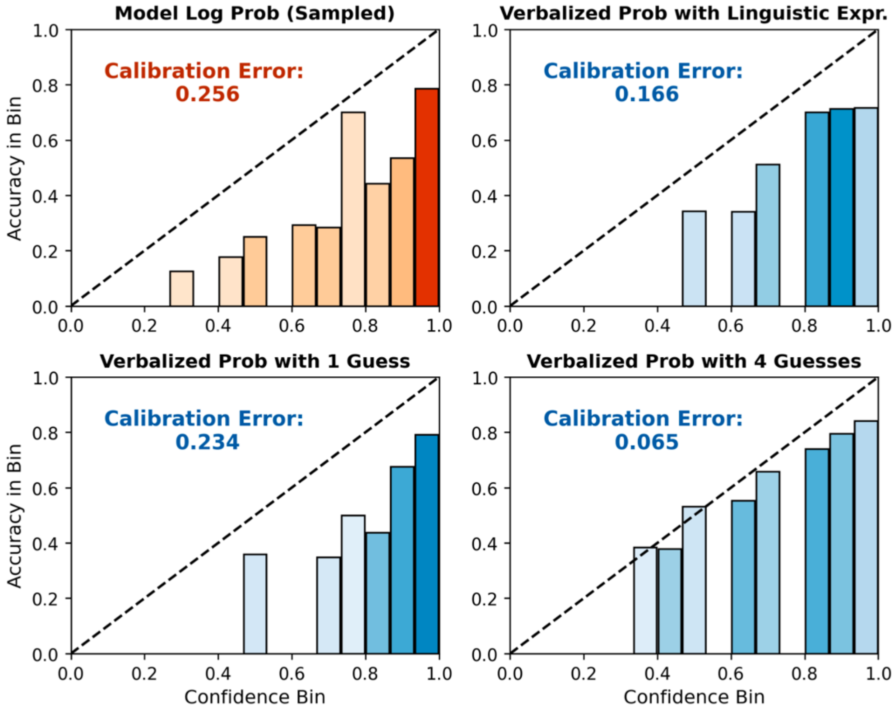
Background: Uncertainty Estimation of LLMs

RUTGERS UNIVERSITY

- Verbalized Uncertainty for Generation
  - Directly ask for uncertainty/confidence in the prompt.
  - It has been controversial.
  - It has been lacking theoretical supports.

Dataset	Acc	Method	ECE	AUROC	AUPRC-P	AUPRC-N
StrategyQA	59.90	Verbalized	39.04	50.34	60.06	40.27
		seq-prob	**7.14**	55.50	62.99	45.22
		len-norm-prob	37.65	55.50	62.99	45.22
		token-prob	32.43	**60.61**	**69.90**	**47.10**

"Comparisons with white-box methods indicate that while ***white-box methods perform better***, the gap is narrow."[1]



"Verbalized confidence scores (blue) are ***better-calibrated*** than log probabilities (orange) for gpt-3.5-turbo."[2]

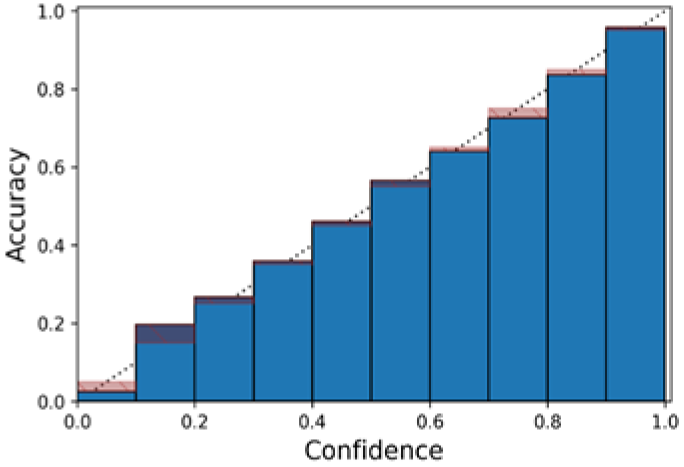
[1] Xiong et al. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.
[2] Tian et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *EMNLP*, 2023.

Background: Uncertainty Estimation of LLMs

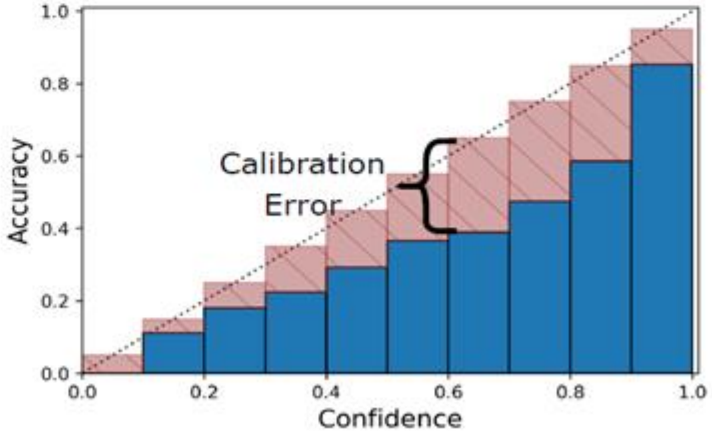
RUTGERS UNIVERSITY

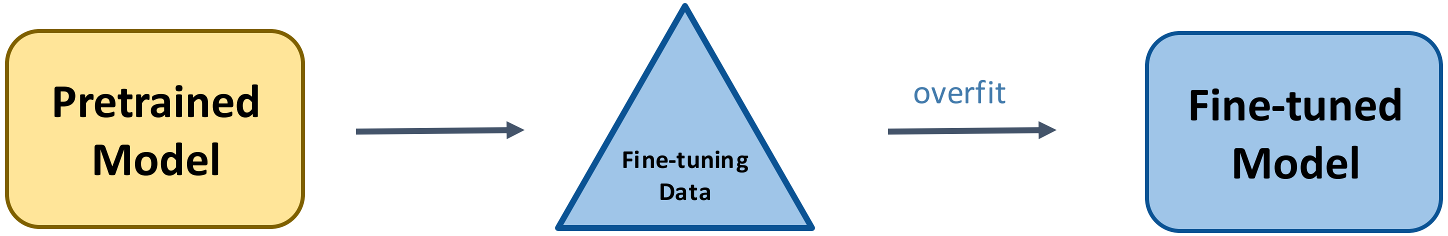
- Uncertainty Estimation for Downstream Adaptation
  - Data is usually ***scarce***, hence ***overconfidence*** is more likely.

well-calibrated



overconfident





Background: Uncertainty Estimation of LLMs

RUTGERS UNIVERSITY

- Uncertainty Estimation for Downstream Adaptation
  - Data is usually ***scarce***, hence ***overconfidence*** is more likely.
  - Closer to the traditional uncertainty estimation setting.
    - Bayesian Neural Networks are built for it!

$$\underbrace{P(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})}_{\text{predictive distribution}} = \int P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{W})\underbrace{P(\boldsymbol{W}|\mathcal{D})}_{\text{posterior distribution}}d\boldsymbol{W} \xleftarrow{\text{approximate}} \underbrace{q(\boldsymbol{W}|\boldsymbol{\theta})}_{\text{variational distribution}}$$

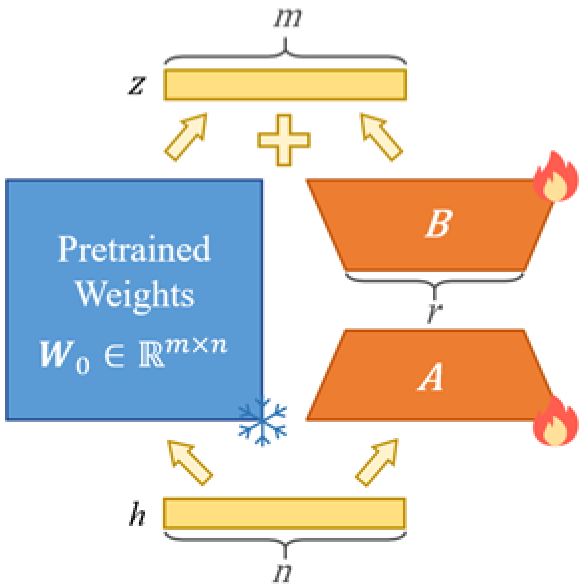
    - The true posterior is usually intractable!

$$\min_\theta \mathrm{KL}[q(\boldsymbol{W}|\boldsymbol{\theta})\|P(\boldsymbol{W}|\mathcal{D})] \quad \Leftrightarrow \quad \min_\theta \underbrace{-\mathbb{E}_{q(\boldsymbol{W}|\boldsymbol{\theta})}[\log P(\mathcal{D}|\boldsymbol{W})] + \mathrm{KL}[q(\boldsymbol{W}|\boldsymbol{\theta}) \parallel P(\boldsymbol{W})]}_{\text{Variational Free Energy}}$$

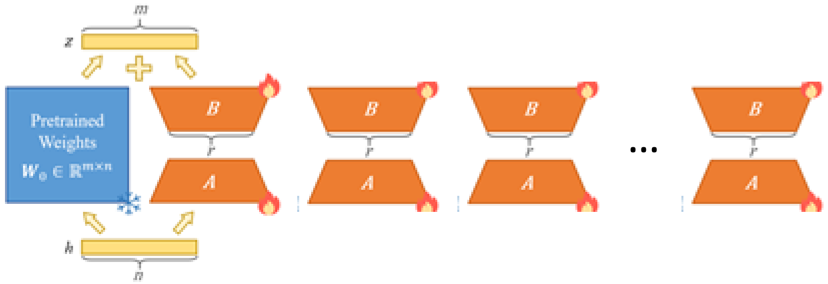
    - What about the extra cost of “Bayesianziation”?

Background: Uncertainty Estimation of LLMs

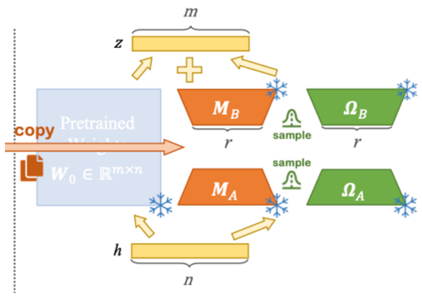
- Uncertainty Estimation for Downstream Adaptation
  - Data is usually ***scarce***, hence ***overconfidence*** is more likely.
  - Bayesian Neural Nets (BNNs) for uncertainty estimation.
  - Parameter-Efficient Fine-Tuning (PEFT) for parameter efficiency.



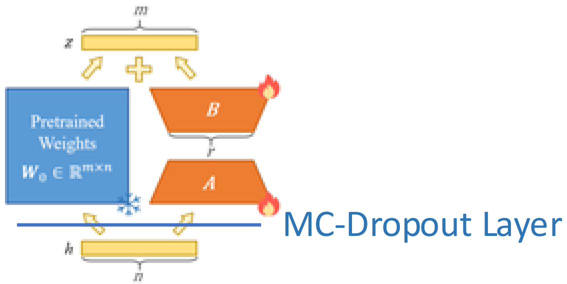
Low-Rank Adaptation (LoRA)[3]



LoRA Ensemble[4]



Laplace LoRA[6]



LoRA MC Dropout[5]

[3] Hu et al. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 2021
[4] Wang et al. LoRA Ensembles for Language Model Fine-tuning. *arXiv preprint arXiv:2310.00035*, 2023
[5] Gal et al. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *ICML*, 2016
[6] Yang et al. Bayesian Low-rank Adaptation for Large Language Models. *ICLR*, 2024

Background: Uncertainty Estimation of LLMs

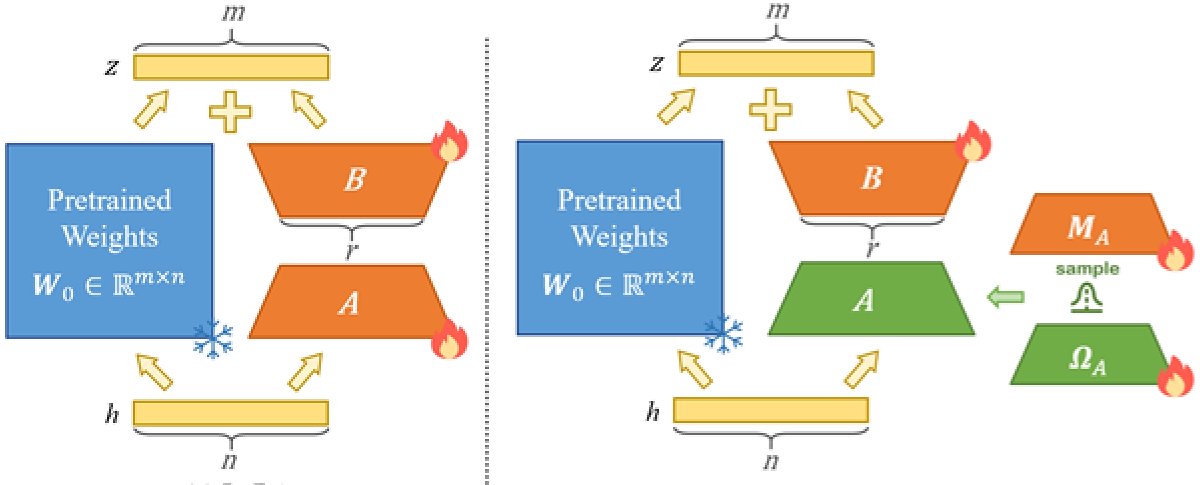
RUTGERS UNIVERSITY

- BLoB: Bayesian Low-Rank Adaptation by Backpropagation for Large Language Models[7]
  - Asymmetric Bayesianization (AB):
    - Elements of $\boldsymbol{A}$ are independent Gaussian:

$$q(\boldsymbol{A}|\boldsymbol{\theta} = \{\boldsymbol{M}, \boldsymbol{\Omega}\}) = \prod_{ij} \mathcal{N}(A_{ij}|M_{ij}, \Omega_{ij}^2)$$

    - Elements of $\boldsymbol{B}$ are deterministic:

$$W_{ij} = W_{0,ij} + \sum_{k=1}^{r} B_{ik} A_{kj},$$



(a) LoRA (b) BLoB (Ours)

Advantage:
- Reduces sampling noise -> Improves convergence!
- Reduces additional memory cost by 50%!

[7] Wang et al. BLoB: Bayesian Low-Rank Adaptation by Backpropagation for Large Language Models. *NeurIPS*, 2024

Training-Free Bayesianization (TFB): Motivation

RUTGERS UNIVERSITY

- Main Problems with Verbalized Uncertainty of LLMs:
  - No good theoretical guarantees
  - Empirically unstable

- Main Challenges of BLoB:
  - Training configuration might heavily depend on different data distributions
  - Hard to find the right training configuration yielding good uncertainty estimation ability

Can we "Bayesianize" a low-rank adapter in a ***theoretically sound*** and ***empirically simple*** way?

Training-Free Bayesianization (TFB)

- TFB Modeling
  - Variational posterior: low-rank isotropic Gaussians

$$q(\mathrm{vec}(\boldsymbol{W})|\boldsymbol{B},\boldsymbol{\theta}) = \mathcal{N}(\mathrm{vec}(\boldsymbol{W})|\boldsymbol{\mu}_q, \mathrm{proj}(\underline{\sigma_q^2 \boldsymbol{I}}))$$

controlled by single variable

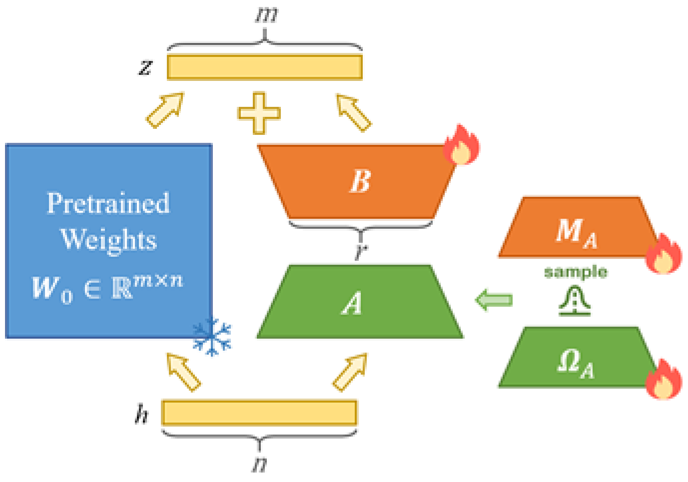


Figure: BLoB Bayesianization

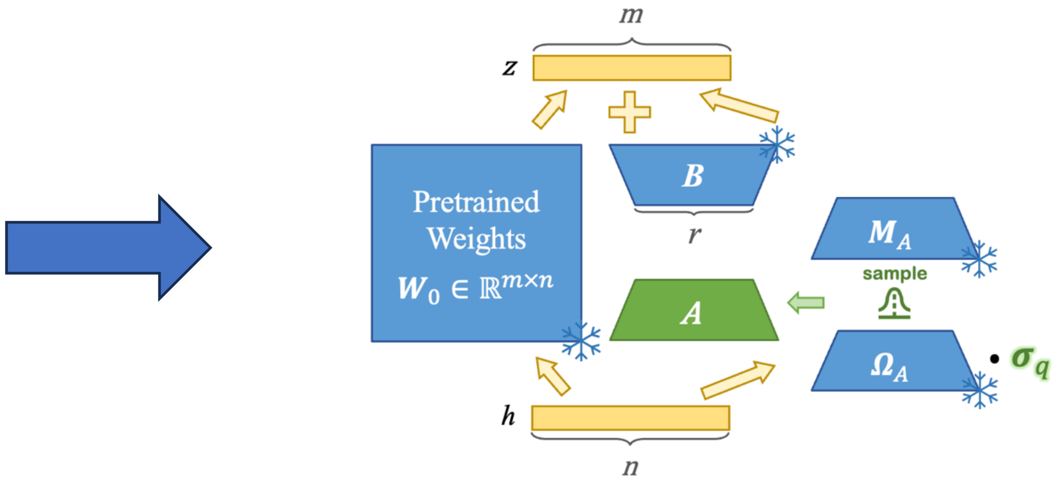


Figure: TFB Bayesianization

[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB)

RUTGERS UNIVERSITY

- TFB Modeling
  - Variational posterior: low-rank isotropic Gaussians

- TFB Modeling
  - Variational posterior: low-rank isotropic Gaussians
- In practice:
  - Given the LoRA adapter $\{\boldsymbol{B}, \boldsymbol{A}\}$;
  - Compact Singular Value Decomposition (SVD) of $\boldsymbol{B}$:
  - Transform the original LoRA into a new pair:
  - Calculate the variance matrix $\boldsymbol{\Omega}$:

controlled by single variable

- In practice:
  - Given the LoRA adapter $\{\boldsymbol{B}, \boldsymbol{A}\}$;
  - Compact Singular Value Decomposition (SVD) of $\boldsymbol{B}$:

$$\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top,$$

  - Transform the original LoRA into a new pair:

$$\{\boldsymbol{B}' = \boldsymbol{U}\boldsymbol{D}, \boldsymbol{A}' = \boldsymbol{V}^\top\boldsymbol{A}\},$$

  - Calculate the variance matrix $\boldsymbol{\Omega}$:

$$\Omega_{ij} = {}^{\sigma_q}/_{d_i}, \quad \forall i \in [r], \forall j \in [n].$$

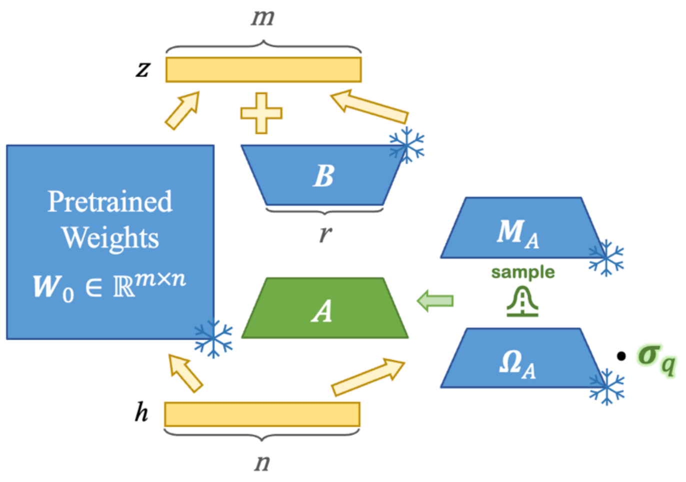


Figure: TFB Bayesianization

[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB)

RUTGERS UNIVERSITY

- TFB Modeling
  - Instead of performing Variational Inference (VI), we perform maximal variance search:

$$\begin{aligned} \max \quad & \sigma_q \\ s.t. \quad & |l(\mathcal{D}|\boldsymbol{B}', \boldsymbol{M}, \boldsymbol{\Omega}(\sigma_q)) - l(\mathcal{D}|\boldsymbol{B}, \boldsymbol{A})| \le \epsilon, \end{aligned}$$

where

- $l(\mathcal{D}|\boldsymbol{B}, \boldsymbol{A})$ is the original performance
- $l(\mathcal{D}|\boldsymbol{B}', \boldsymbol{M}, \boldsymbol{\Omega}(\sigma_q))$ is the post-Bayesianization performance
- ϵ is the max tolerance of the performance change

[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB)

- TFB Modeling
  - Instead of performing Var … variance search:

where

- $l(\mathcal{D}|\boldsymbol{B}, \boldsymbol{A})$ is the origin…
- $l(\mathcal{D}|\boldsymbol{B}', \boldsymbol{M}, \boldsymbol{\Omega}(\sigma_q))$ is …
- ϵ is the max tolerance of t…

Algorithm 1 Training-Free Bayesianization (`TFB`)

input $\mathcal{D}$: Anchor Dataset;
input $\{\boldsymbol{B}, \boldsymbol{A}\}$: Low-Rank Component;
input l: Model Evaluation Metric;
input ϵ: Performance Change Tolerance;
input $[\sigma_{q_{\min}}, \sigma_{q_{\max}}]$: search range of the posterior STD.

Evaluate the original performance: $p_0 \leftarrow l(\mathcal{D}|\boldsymbol{B}, \boldsymbol{A})$.
Singular Value Decomposition on $\boldsymbol{B}$:
  $\boldsymbol{U}, \boldsymbol{D}, \boldsymbol{V} \leftarrow \mathrm{SVD}(\boldsymbol{B})$. ▷ Eqn. 4.
Get an equivalent pair of the low-rank component:
  $\boldsymbol{B}' \leftarrow \boldsymbol{U}\boldsymbol{D}$; $\boldsymbol{A}' \leftarrow \boldsymbol{V}^\top \boldsymbol{A}$. ▷ Eqn. 5.
while σ_q not converged **do**
  $\sigma_q \leftarrow (\sigma_{q_{\max}} + \sigma_{q_{\min}})/2$.
  Calculate the STD matrix $\boldsymbol{\Omega}$ for $\boldsymbol{A}'$:
    $\Omega_{ij} = \sigma_q / D_{ii}$. ▷ Eqn. 6.
  Evaluate the performance:
    $p \leftarrow l(\mathcal{D}|\boldsymbol{B}', \boldsymbol{A}', \boldsymbol{\Omega})$.
  if $|p - p_0| < \epsilon$ **then**
    $\sigma_{q_{\min}} \leftarrow \sigma_q$.
  else
    $\sigma_{q_{\max}} \leftarrow \sigma_q$.
  end if
end while
output $\{\boldsymbol{B}', \boldsymbol{A}', \boldsymbol{\Omega}\}$: Bayesianized Low-Rank Adapter.

[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB): Theory

- Theoretical Analysis
  - **(Thm.1)** TFB produces low-rank isotropic Gaussian posteriors.

$$
\begin{aligned}
q(\mathrm{vec}(\boldsymbol{W})|\sigma_q) &= \mathcal{N}(\mathrm{vec}(\boldsymbol{W})|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q),\\
\textit{where} \quad \boldsymbol{\mu}_q &= \mathrm{vec}(\boldsymbol{W}_0 + \boldsymbol{B}'\boldsymbol{M}),\\
\boldsymbol{\Sigma}_q &= \sigma_q^2 \cdot \boldsymbol{I}_n \otimes \begin{bmatrix} \boldsymbol{I}_r & \\ & \boldsymbol{O}_{m-r} \end{bmatrix}.
\end{aligned} \tag{10}
$$

  - **(Thm.2)** TFB is equivalent to Variational Inference.

$$
\underbrace{\begin{aligned} \max \quad & \sigma_q \\ \textit{s.t.} \quad & l_{\mathcal{D}}(\sigma_q) \le \epsilon, \end{aligned}}_{\text{Variational Inference}} \Leftrightarrow \underbrace{\min_{\sigma_q} \quad l_{\mathcal{D}}(\sigma_q) + \lambda \, \mathrm{KL}[q(\boldsymbol{W}|\sigma_q) \parallel P(\boldsymbol{W})],}_{\text{Variational Inference}}
$$

    - If $l_{\mathcal{D}}$ is the NLL loss and locally convex $[0, \epsilon_0)$;
    - And the prior standard deviation $\sigma_p > \epsilon_0$.

[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB): Experiments

RUTGERS UNIVERSITY

- Main Conclusions
  - TFB ***improves accuracy & uncertainty estimation*** across trained LoRA checkpoints (MLE, MAP, BLoB).

Metric	Method	TF?	In-Distribution Datasets						Out-of-Distribution Datasets (OBQA→X)			
									Small Shift		*Large Shift*	
			WG-S	ARC-C	ARC-E	WG-M	OBQA	BoolQ	ARC-C	ARC-E	Chem	Phy
ACC (↑)	MCD	✗	78.03±0.61	81.64±1.79	91.37±0.38	83.18±0.84	87.20±1.02	89.93±0.16	81.42±1.38	87.27±0.84	47.92±2.25	**46.53±0.49**
	ENS	✗	**78.82±0.52**	82.55±0.42	**91.84±0.36**	**83.99±0.74**	87.37±0.67	**90.50±0.14**	79.62±0.57	86.56±0.60	49.65±3.22	44.44±1.96
	LAP	BP	76.05±0.92	79.95±0.42	90.73±0.08	82.83±0.85	87.90±0.20	89.36±0.52	81.08±1.20	87.21±1.20	48.26±3.93	46.18±1.30
	MonteCLoRA	✗	69.20±0.18	78.38±0.89	90.79±0.62	74.79±0.23	84.13±0.31	89.17±0.30	79.63±0.87	86.58±0.49	**50.00±1.04**	42.01±2.41
	BLoB	✗	76.45±0.37	82.32±1.15	91.14±0.54	82.01±0.56	87.57±0.21	89.65±0.15	79.75±0.43	87.13±0.00	42.71±3.71	44.79±6.64
	MLE	-	77.87±0.54	81.08±0.48	91.67±0.36	82.30±0.53	87.90±0.87	89.58±0.26	81.48±2.41	86.83±0.87	45.83±0.85	42.36±1.77
	+ TFB (Ours)	✓	77.44±0.30	82.53±1.00	91.33±0.37	82.53±0.56	**88.53±0.57**	89.75±0.25	79.76±1.24	85.52±0.56	44.33±4.03	37.00±2.16
	MAP	-	76.90±0.97	81.08±2.48	91.61±0.44	82.59±0.28	85.73±0.19	90.09±0.28	79.98±0.87	86.58±0.79	43.40±4.98	38.54±3.40
	+ TFB (Ours)	✓	76.43±0.72	82.80±1.42	91.39±0.37	82.64±0.58	86.00±0.16	89.96±0.18	80.61±1.24	86.30±0.89	45.33±2.87	35.67±4.11
	BLoB-Mean	✗	77.72±0.12	82.60±0.60	91.64±0.55	83.92±0.48	88.00±0.80	89.86±0.05	82.06±1.15	**88.54±0.31**	39.93±5.20	39.93±4.02
	+ TFB (Ours)	✓	77.81±0.36	**83.33±0.19**	91.76±0.48	83.81±0.39	87.80±0.16	90.11±0.28	**82.93±1.54**	87.64±0.51	39.67±7.32	37.33±6.65
ECE (↓)	MCD	✗	16.13±0.54	13.69±1.11	6.73±0.71	13.05±0.99	9.76±0.71	7.95±0.17	13.63±1.18	9.27±0.60	30.91±3.57	33.08±1.40
	ENS	✗	14.72±0.17	13.45±1.19	6.59±0.45	11.17±0.92	8.17±0.86	7.35±0.55	11.37±1.82	7.21±1.13	18.92±6.03	26.80±3.23
	LAP	BP	**4.18±0.11**	9.26±3.08	5.27±0.51	**3.50±0.78**	8.93±0.34	**1.93±0.22**	7.83±1.49	7.80±1.99	**14.49±0.57**	13.17±2.14
	MonteCLoRA	✗	18.29±0.27	12.22±0.75	7.23±0.71	15.97±0.45	9.79±0.07	7.09±0.52	10.65±0.53	8.18±0.26	23.21±0.17	30.39±4.76
	BLoB	✗	9.93±0.22	**5.41±1.17**	2.70±0.87	4.28±0.64	2.91±0.92	2.58±0.25	**5.61±0.40**	**2.48±0.43**	16.67±0.87	**12.78±4.18**
	MLE	-	17.02±0.46	16.35±0.68	7.00±0.53	13.83±0.65	9.77±0.81	8.69±0.21	14.45±2.19	10.78±0.50	32.46±2.60	38.41±4.44
	+ TFB (Ours)	✓	12.98±0.37	11.63±0.68	5.14±0.14	10.01±0.70	7.20±0.47	7.39±0.26	6.54±0.53	5.69±1.64	14.63±1.46	19.68±3.27
	MAP	-	18.71±0.74	15.77±1.60	6.62±0.64	14.26±0.92	12.19±0.55	8.40±0.25	16.46±0.44	11.36±0.58	34.79±3.76	38.50±2.18
	+ TFB (Ours)	✓	14.95±0.65	11.27±2.53	5.76±0.63	10.97±1.19	9.70±0.69	6.86±0.31	13.25±0.95	9.22±0.91	27.21±2.62	35.91±4.12
	BLoB-Mean	✗	15.43±0.15	12.41±1.52	4.91±0.28	9.37±1.33	6.44±0.15	6.26±0.29	11.22±0.38	6.34±0.71	26.65±3.06	25.40±5.40
	+ TFB (Ours)	✓	8.16±0.48	6.48±0.36	**2.44±0.50**	3.83±0.43	**2.67±0.18**	3.10±0.59	6.69±1.63	3.61±0.87	18.45±6.75	20.53±6.27
NLL (↓)	MCD	✗	0.83±0.01	0.99±0.10	0.45±0.06	0.64±0.03	0.62±0.08	0.49±0.01	1.03±0.02	0.61±0.03	1.91±0.18	2.02±0.15
	ENS	✗	0.75±0.02	0.80±0.11	0.38±0.03	0.55±0.02	0.45±0.05	0.42±0.05	0.72±0.07	0.44±0.03	1.40±0.18	1.50±0.13
	LAP	BP	0.56±0.00	1.18±0.02	1.04±0.01	0.51±0.00	0.94±0.00	0.43±0.00	1.17±0.01	1.11±0.00	**1.27±0.01**	**1.28±0.00**
	MonteCLoRA	✗	0.82±0.02	0.71±0.03	0.51±0.04	0.74±0.02	0.55±0.02	0.36±0.02	0.68±0.03	0.49±0.01	1.43±0.00	1.44±0.06
	BLoB	✗	0.58±0.00	**0.51±0.03**	**0.23±0.01**	0.43±0.01	0.34±0.01	**0.26±0.01**	0.56±0.02	**0.35±0.02**	1.34±0.04	1.35±0.10
	MLE	-	0.88±0.04	1.20±0.11	0.46±0.04	0.68±0.01	0.61±0.06	0.52±0.01	1.07±0.06	0.72±0.06	1.91±0.16	2.25±0.21
	+ TFB (Ours)	✓	0.68±0.03	0.85±0.02	0.33±0.03	0.53±0.01	0.46±0.04	0.42±0.00	0.66±0.02	0.44±0.01	1.39±0.11	1.49±0.05
	MAP	-	0.99±0.07	1.12±0.23	0.46±0.03	0.74±0.07	0.79±0.02	0.52±0.01	1.19±0.04	0.83±0.06	1.97±0.13	2.32±0.10
	+ TFB (Ours)	✓	0.77±0.05	0.80±0.15	0.38±0.03	0.57±0.05	0.61±0.03	0.40±0.01	0.96±0.08	0.66±0.06	1.69±0.16	2.12±0.08
	BLoB-Mean	✗	0.74±0.02	0.73±0.04	0.29±0.03	0.47±0.03	0.37±0.02	0.32±0.02	0.67±0.07	0.39±0.03	1.53±0.13	1.54±0.15
	+ TFB (Ours)	✓	**0.55±0.01**	0.53±0.04	**0.23±0.02**	**0.40±0.01**	**0.33±0.02**	0.27±0.01	**0.52±0.05**	**0.35±0.02**	1.36±0.13	1.46±0.11

[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB): Experiments

RUTGERS UNIVERSITY

- Main Conclusions
  - TFB ***improves accuracy & uncertainty estimation*** across trained LoRA checkpoints (MLE, MAP, BLoB).
  - TFB works perfectly with ***small amount of data*** (for search).

Table 7: Dataset Statistics. The size of the Anchor Set $\mathcal{D}$ is used in Table 1, 3 and 14.

	WG-S	ARC-C	ARC-E	WG-M	OBQA	BoolQ	Combined
Size of Label Space	2	5	5	2	4	2	7
Size of Training Set	640	1,119	2,251	2,258	4,957	9,427	20,652
Size of Anchor Set $\mathcal{D}$	500 (78%)	500 (45%)	500 (22%)	500 (22%)	500 (10%)	500 (5%)	500 (2%)
Size of Test Set	1,267	299	570	1,267	500	3,270	7,173

[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB): Experiments

RUTGERS UNIVERSITY

- Main Conclusions
  - TFB ***improves accuracy & uncertainty estimation*** across trained LoRA checkpoints (MLE, MAP, BLoB).
  - TFB works perfectly with ***small amount of data*** (for search).
  - TFB works the best among ***other posterior families***.

Metric	**Method**	**In-Distribution Datasets**							**Out-of-Distribution Datasets** (OBQA→X)			
									Small Shift		*Large Shift*	
		WG-S	ARC-C	ARC-E	WG-M	OBQA	BoolQ	Rk. (↓)	ARC-C	ARC-E	Chem	Phy
ACC (↑)	BLoB-Mean	77.72±0.12	82.60±0.60	91.64±0.55	**83.92±0.48**	88.00±0.80	89.86±0.05	2.50	82.06±1.15	**88.54±0.31**	39.93±5.20	39.93±4.02
	+ TFB (FR)	75.57±0.25	83.20±0.65	91.58±0.67	82.19±1.09	**88.73±0.41**	89.46±0.17	2.83	81.33±0.82	88.06±0.75	42.00±2.16	**41.33±5.44**
	+ TFB (C-STD)	76.35±0.08	83.20±0.33	91.33±0.70	81.79±0.51	88.20±0.57	89.65±0.08	3.00	81.73±0.68	88.18±0.65	**43.00±1.41**	39.33±3.86
	+ TFB (Final)	**77.81±0.36**	**83.33±0.19**	**91.76±0.48**	83.81±0.39	87.80±0.16	**90.11±0.28**	**1.67**	**82.93±1.54**	87.64±0.51	39.67±7.32	37.33±6.65
ECE (↓)	BLoB-Mean	15.43±0.15	12.41±1.52	4.91±0.28	9.37±1.33	6.44±0.15	6.26±0.29	4.00	11.22±0.38	6.34±0.71	26.65±3.06	25.40±5.40
	+ TFB (FR)	10.42±0.29	7.45±0.88	**2.01±1.03**	4.36±0.68	3.70±1.04	3.62±0.10	2.67	7.19±1.40	3.29±1.03	**17.78±1.01**	**19.14±4.01**
	+ TFB (C-STD)	9.23±0.20	**5.98±0.32**	2.94±0.67	3.86±0.45	3.17±0.21	**2.82±0.62**	1.83	6.89±0.89	**2.76±0.88**	18.27±2.52	19.45±3.46
	+ TFB (Final)	**8.16±0.48**	6.48±0.36	2.44±0.50	**3.83±0.43**	**2.67±0.18**	3.10±0.59	**1.50**	**6.69±1.63**	3.61±0.87	18.45±6.75	20.53±6.27
NLL (↓)	BLoB-Mean	0.74±0.02	0.73±0.04	0.29±0.03	0.47±0.03	0.37±0.02	0.32±0.02	3.67	0.67±0.07	0.39±0.03	1.53±0.13	1.54±0.15
	+ TFB (FR)	0.60±0.01	0.53±0.03	0.23±0.02	0.43±0.01	**0.33±0.02**	0.27±0.01	2.00	0.57±0.04	0.34±0.02	**1.34±0.07**	1.42±0.09
	+ TFB (C-STD)	0.57±0.01	**0.51±0.02**	**0.22±0.01**	0.43±0.01	**0.33±0.01**	**0.26±0.01**	**1.33**	0.56±0.04	**0.33±0.02**	**1.34±0.08**	**1.41±0.09**
	+ TFB (Final)	**0.55±0.01**	0.53±0.04	0.23±0.02	**0.40±0.01**	**0.33±0.02**	0.27±0.01	1.50	**0.52±0.05**	0.35±0.02	1.36±0.13	1.46±0.11

[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB): Experiments

RUTGERS UNIVERSITY

- Main Conclusions
  - TFB ***improves accuracy & uncertainty estimation*** across trained LoRA checkpoints (MLE, MAP, BLoB).
  - TFB works perfectly with ***small amount of data*** (for search).
  - TFB works the best among ***other posterior families***.
  - TFB works for ***various LLM backbones***.

Table 3. Performance of different **LLM backbones** on the combined dataset of six commonsense reasoning tasks.

Method	**ACC** (↑)	**ECE** (↓)	**NLL** (↓)
Llama2-7B	**81.41**$\pm$**0.64**	4.50$\pm$0.37	**0.43**$\pm$**0.00**
+ TFB (Ours)	81.32$\pm$0.51	**1.24**$\pm$**0.22**	**0.43**$\pm$**0.00**
Llama3-8B	**86.93**$\pm$**0.09**	4.28$\pm$0.54	**0.34**$\pm$**0.00**
+ TFB (Ours)	86.61$\pm$0.20	**1.64**$\pm$**0.64**	**0.34**$\pm$**0.00**
Llama3.1-8B	**86.70**$\pm$**0.08**	4.74$\pm$0.28	0.35$\pm$0.00
+ TFB (Ours)	86.45$\pm$0.33	**1.05**$\pm$**0.06**	**0.34**$\pm$**0.00**
Mistral-7B-v0.3	**86.88**$\pm$**0.51**	5.05$\pm$0.88	0.35$\pm$0.02
+ TFB (Ours)	86.64$\pm$0.28	**1.68**$\pm$**0.53**	**0.33**$\pm$**0.01**

[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB): Experiments

- Main Conclusions
  - TFB ***improves accuracy & uncertainty estimation*** across trained LoRA checkpoints (MLE, MAP, BLoB).
  - TFB works perfectly with ***small amount of data*** (for search).
  - TFB works the best among ***other posterior families***.
  - TFB works for ***various LLM backbones***.
  - TFB works ***beyond LoRA adapters***.

Table 4. Performance of different **LoRA-like PEFT methods** on the combined dataset of six commonsense reasoning tasks.

Method	ACC (↑)	ECE (↓)	NLL (↓)
LoRA	**86.70**$\pm$**0.08**	4.74$\pm$0.28	0.35$\pm$0.00
+ TFB (Ours)	86.45$\pm$0.33	**1.05**$\pm$**0.06**	**0.34**$\pm$**0.00**
VeRA	**84.93**$\pm$**0.50**	5.11$\pm$0.55	0.39$\pm$0.01
+ TFB (Ours)	84.28$\pm$0.48	**1.44**$\pm$**0.44**	**0.38**$\pm$**0.01**
PiSSA	**86.83**$\pm$**0.51**	4.26$\pm$0.14	0.35$\pm$0.00
+ TFB (Ours)	86.61$\pm$0.43	**1.17**$\pm$**0.22**	**0.33**$\pm$**0.00**

[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB): Experiments

RUTGERS UNIVERSITY

- Main Conclusions
  - TFB ***improves accuracy & uncertainty estimation*** across trained LoRA checkpoints (MLE, MAP, BLoB).
  - TFB works perfectly with ***small amount of data*** (for search).
  - TFB works the best among ***other posterior families***.
  - TFB works for ***various LLM backbones***.
  - TFB works ***beyond LoRA adapters***.
  - TFB achieves significant improvement in terms of ***efficiency.***

Method	**Batch Size**	WG-S		ARC-C		ARC-E		WG-M		OBQA		BoolQ	
		Time (s)	Mem. (MB)	Time (s)	Mem. (MB)	Time (s)	Mem. (MB)	Time (s)	Mem. (MB)	Time (s)	Mem. (MB)	Time (s)	Mem. (MB)
LoRA	4	338	12,894	632	19,762	1,238	18,640	1,339	13,164	2,692	17,208	6,489	29,450
BLoB	4	371 (1.10x)	13,194 (1.02x)	685 (1.08x)	21,736 (1.10x)	1,360 (1.10x)	20,700 (1.11x)	1,476 (1.10x)	13,194 (1.00x)	3,257 (1.21x)	18,046 (1.05x)	7,251 (1.12x)	30,578 (1.04x)
TFB (Ours)	4	1,203 (3.56x)	10,372 (0.80x)	1,257 (1.99x)	11,966 (0.61x)	1,246 (1.01x)	11,202 (0.60x)	1,237 (0.92x)	10,344 (0.79x)	1,238 (0.46x)	10,376 (0.60x)	1,452 (0.22x)	16,340 (0.55x)
TFB (Ours)	8	628 (1.86x)	10,666 (0.83x)	731 (1.16x)	15,286 (0.77x)	702 (0.57x)	12,598 (0.68x)	634 (0.47x)	10,662 (0.81x)	642 (0.24x)	12,116 (0.70x)	1,015 (0.16x)	22,146 (0.75x)
TFB (Ours)	12	446 (1.31x)	12,064 (0.93x)	599 (0.94x)	18,204 (0.92x)	540 (0.43x)	14,310 (0.76x)	441 (0.32x)	11,370 (0.86x)	487 (0.18x)	13,410 (0.77x)	908 (0.13x)	25,220 (0.85x)

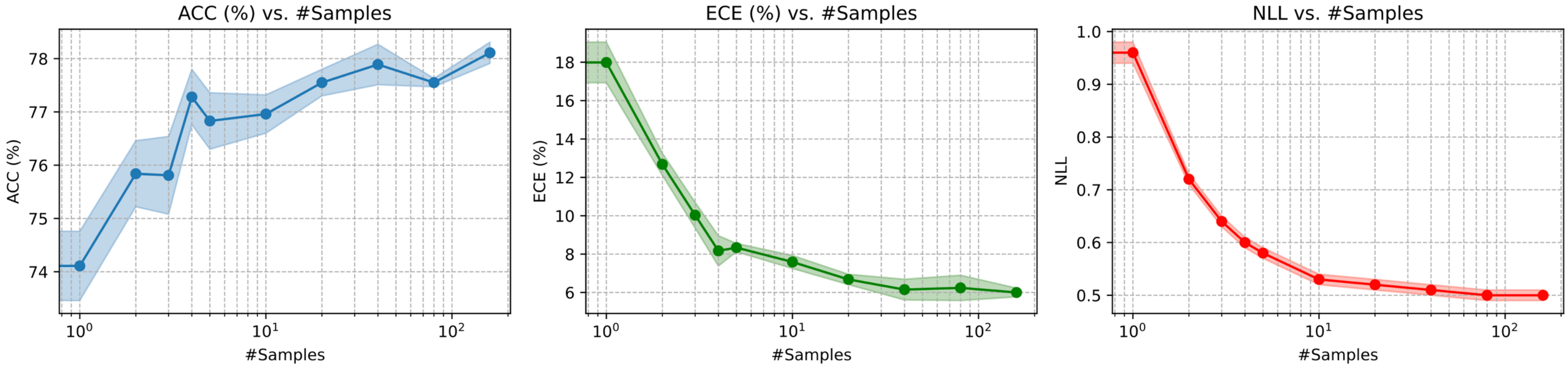
Datasets: WG-S, ARC-C, ARC-E, WG-M, OBQA, BoolQ

[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB): Experiments

RUTGERS UNIVERSITY

- Main Conclusions
  - TFB ***improves accuracy & uncertainty estimation*** across trained LoRA checkpoints (MLE, MAP, BLoB).
  - TFB works perfectly with ***small amount of data*** (for search).
  - TFB works the best among ***other posterior families***.
  - TFB works for ***various LLM backbones***.
  - TFB works ***beyond LoRA adapters***.
  - TFB achieves significant improvement in terms of ***efficiency.***
  - TFB improves w/ ***Scaling Test-Time Compute***

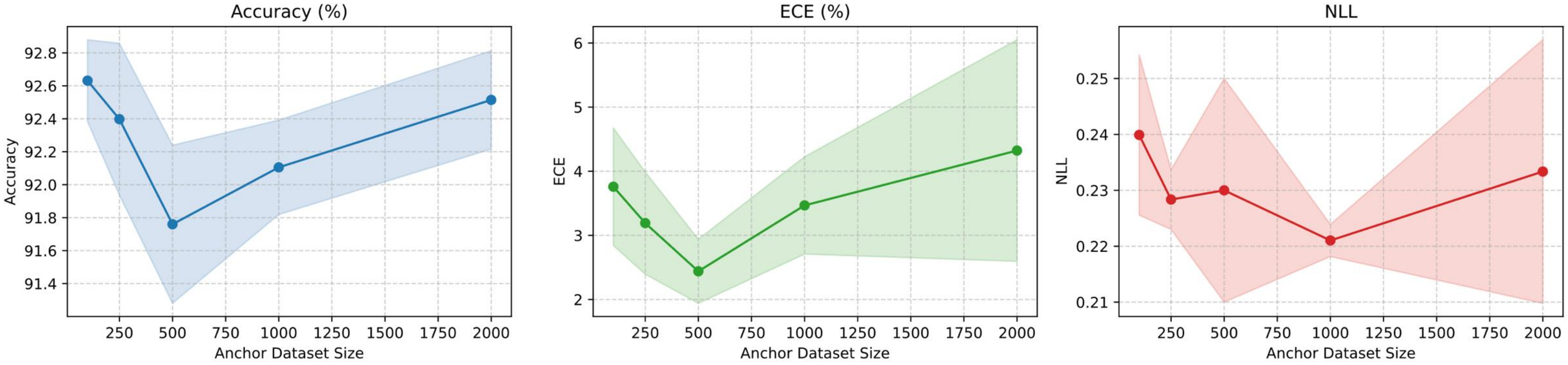


[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

Training-Free Bayesianization (TFB): Experiments

RUTGERS UNIVERSITY

- Main Conclusions
  - TFB ***improves accuracy & uncertainty estimation*** across trained LoRA checkpoints (MLE, MAP, BLoB).
  - TFB works perfectly with ***small amount of data*** (for search).
  - TFB works the best among ***other posterior families***.
  - TFB works for ***various LLM backbones***.
  - TFB works ***beyond LoRA adapters***.
  - TFB achieves significant improvement in terms of ***efficiency.***
  - TFB improves w/ ***Scaling Test-Time Compute.***
  - TFB is ***robust*** against ***anchor dataset size.***



[10] Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. *arXiv preprint arXiv:2412.05723*, 2024

References

[1] Xiong et al. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. arXiv preprint arXiv:2306.13063, 2023.

[2] Tian et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. EMNLP, 2023.

[3] Hu et al. LoRA: Low-Rank Adaptation of Large Language Models. ICLR, 2021

[4] Wang et al. LoRA Ensembles for Language Model Fine-tuning. arXiv preprint arXiv:2310.00035, 2023

[5] Gal et al. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. ICML, 2016

[6] Yang et al. Bayesian Low-rank Adaptation for Large Language Models. ICLR, 2024

[7] ***Wang et al. BLoB: Bayesian Low-Rank Adaptation by Backpropagation for Large Language Models. NeurIPS, 2024***

[8] Wen et al. Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches. ICLR, 2018

[9] Blundell et al. Weight Uncertainty in Neural Networks. ICML, 2015

[10] ***Shi et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. arXiv preprint arXiv:2412.05723, 2024***