ICML 2026 ACCEPTED PAPER

Singular Bayesian Neural Networks

Mame Diarra Touré · David A. Stephens

$W = AB^T \rightarrow$ singular posterior geometry

theory-backed scalable uncertainty

A low-rank variational BNN whose induced posterior has singular geometric support.

arradiat.github.io/projects/singular-bnn · github.com/arradiat/SBNN · arXiv:2602.00387

CONTRIBUTION MAP

This paper shows that low-rank variational BNNs induce singular posterior geometry.

Method

End-to-end variational learning over low-rank factors $W = AB^T$.

Geometry

The induced qW is a pushforward posterior supported on the rank-r manifold.

Singularity

qW is singular w.r.t. Lebesgue measure when $r < \min(m,n)$.

Correlation

Mean-field factors still induce structured weight correlations in W.

Theory

Approximation, PAC-Bayes, and Gaussian-complexity guarantees.

Evidence

MLPs, LSTMs, Transformers; MIMIC, Beijing PM2.5, SST-2.

Core thesis: low-rank is not merely parameter economy; it changes posterior support, covariance, and capacity.

PROBLEM

Bayesian neural networks provide uncertainty, but full weight-space inference is expensive.

O(mn)

O(mn) Parameters

Standard mean-field MFVI requires 2 variational parameters per weight, doubling the count per layer. For a single matrix this scales as O(mn).

2x

Independence Assumption

Factorized posteriors ignore structured weight correlations, limiting expressiveness and the model's ability to represent epistemic uncertainty coherently.

5x

Ensemble Overhead

Deep Ensembles, the practical gold standard, require 5 full model copies, making them prohibitively expensive at modern scale.

Key Gap: No prior work trains low-rank BNNs end-to-end across diverse architectures (MLPs, LSTMs, Transformers) with rigorous theoretical guarantees.

Not LoRA - a different posterior object.

Approach	Backbone	Posterior object	Support of posterior	Gap
Post-hoc noise	Pretrained	Noise on fixed W	Full ambient $\mathbb{R}^{mn}$	Not end-to-end
Low-rank covariance	Full W means	Covariance approx.	Full ambient $\mathbb{R}^{mn}$	Still O(mn) means
Bayesian LoRA	Pretrained adapter	Adapter uncertainty	Rank-r adapter only	Not from-scratch BNN
SBNN (ours)	**From scratch**	**Singular q_W**	**Rank-r manifold M_r**	**—**

The differentiator is the support: SBNN puts probability on the rank-r manifold itself, not on ambient $\mathbb{R}^{mn}$.

Learn uncertainty in factor space, then map it into weight space.

Variational posterior

q(A,B)=qA(A)qB(B), with mean-field Gaussians over factor entries.

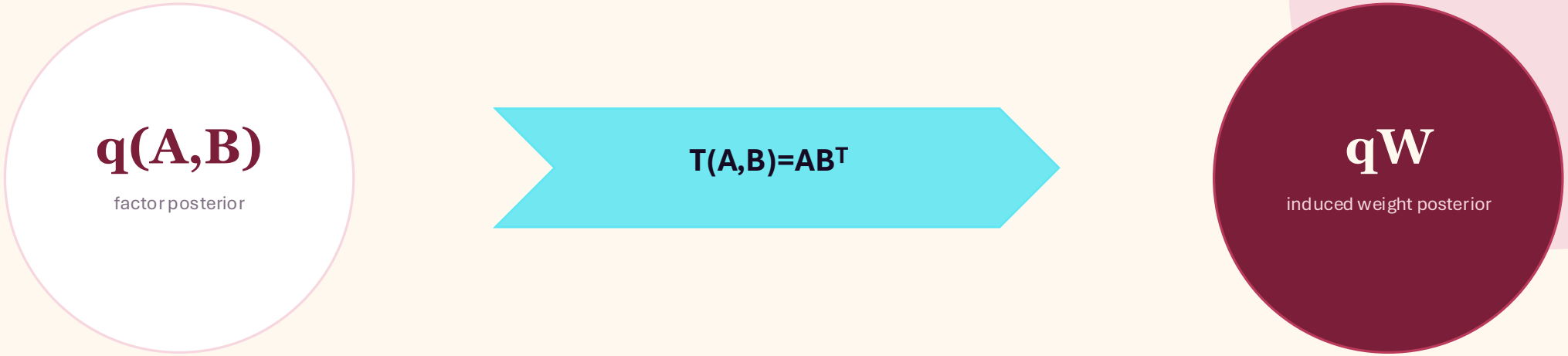
Parameter count

O(mn) becomes O(r(m+n)); rank controls expressiveness.

Drop-in layers

Implemented for dense layers, LSTM gates, and Transformer components.

The posterior over W is induced, not independently assigned.



This is the formal turn: the Bayesian object is a pushforward measure on weight space.

qW is singular with respect to Lebesgue measure.

If $r < \min(m,n)$, and $W = AB^T$ with $A \in R^{mxr}$, $B \in R^{nxr}$:

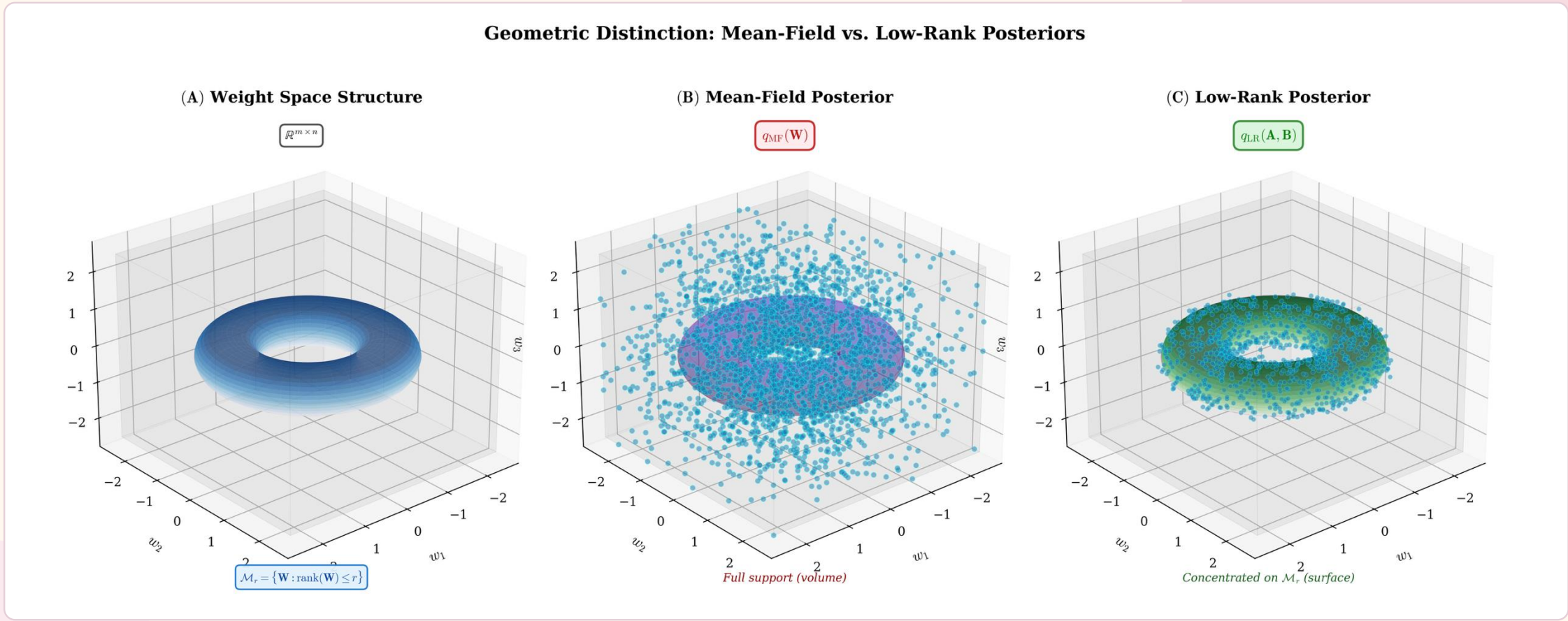
$$qW(Rr) = 1 \quad \text{and} \quad \lambda(Rr) = 0$$

therefore

$$qW \perp \lambda$$

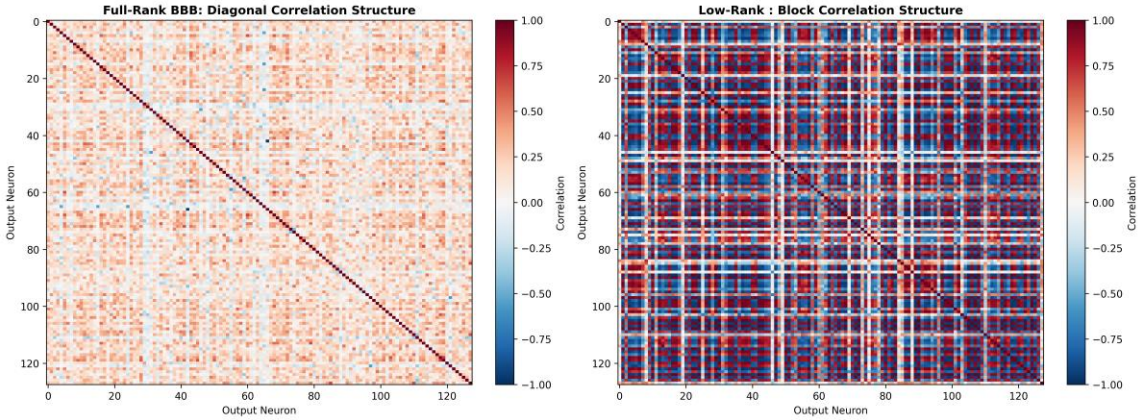
The posterior cannot be represented as a full-dimensional density over R^{mxn}. It lives on a zero-volume rank-r support.

Mean-field fills volume. Low-rank mass lives on a manifold.



This is why "singular BNN" is the right phrase: the posterior support itself has changed.

Mean-field factors do not imply mean-field weights.



Covariance lemma

Even when A and B have independent entries, W_{ij} and $W_{i'j'}$ can be correlated through shared latent dimensions.

Inductive bias

The model trades independence bias for correlation bias: row/column-level uncertainty propagates coherently.

Rank r controls how expressive these correlations can be.

The paper gives three theory certificates, not one.

Approximation

Eckart–Young–Mirsky controls rank-r approximation by the tail singular values of W*.

Learned factors

Error decomposes into learning error ||W−Wr*||F plus unavoidable rank bias σ>r.

Generalization

PAC-Bayes complexity scales as √r(m+n), and Gaussian complexity transfers to Bayesian predictive means.

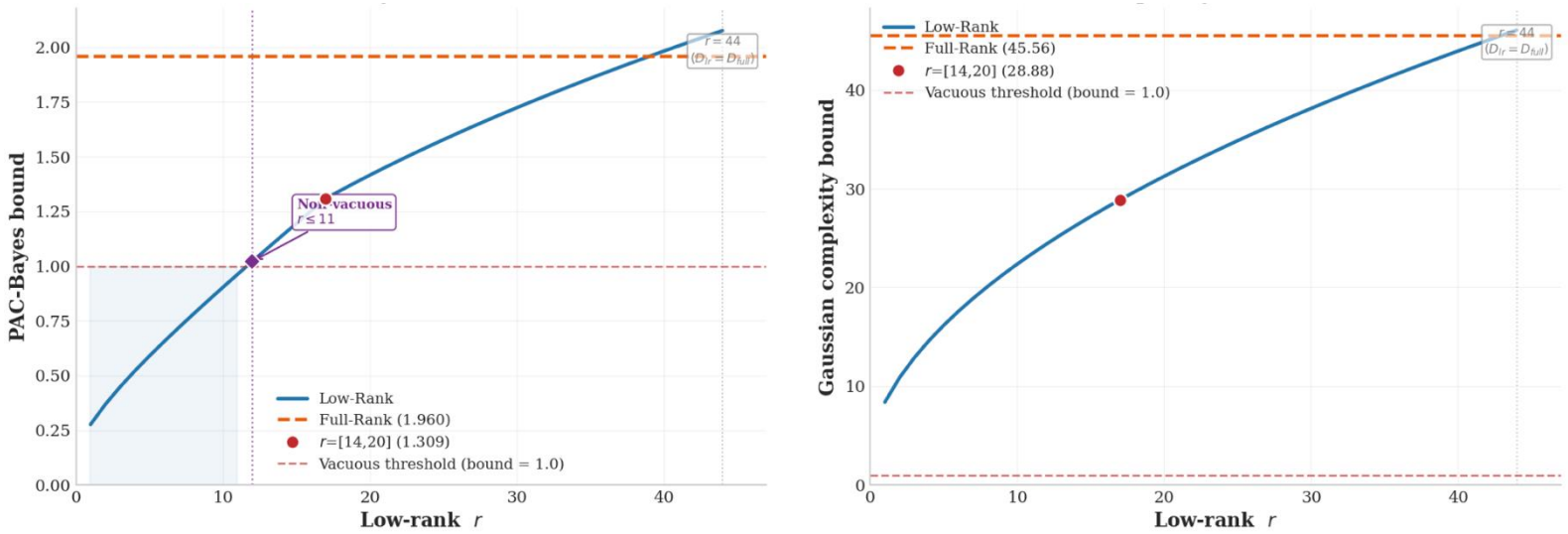


Figure result: PAC-Bayes shows a critical rank transition; Gaussian complexity decreases with rank reduction.

Three Certificates — Formal Statements

① Loss Approximation (Eckart-Young-Mirsky)

Loss gap between W and best rank-r approximation:*

$$|\mathbb{E}[\ell(W^*x,y)] - \mathbb{E}[\ell(W^*_r x,y)]| \le L \cdot R \cdot \sqrt{\Sigma_{i=r+1}\, \sigma_i^2(W^*)}$$

Rapid singular-value decay ⇒ small rank-induced bias.

② PAC-Bayes Generalization Bounds

Complexity ratio, low-rank vs. full-rank posterior:

$$\text{Complexity(Q\_LR)} \,/\, \text{Complexity(Q\_full)} \approx \sqrt{(r(1/m + 1/n))} \ll 1$$

O(r(m+n)) vs O(mn) — tighter bounds when r ≪ min(m,n).

③ Gaussian Complexity Transfer

BNN mean lies in closed convex hull of support class; Gaussian complexity is closure and convex-invariant:

$$\mathcal{G}(\text{F\^{}BNN}) \le \mathcal{G}(\text{F\^{}Pinto(C,r)})$$

Deterministic bounds transfer to Bayesian means.

PAC-Bayes: Complexity Reduction in Detail

Full-Rank MFVI

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{C_{\max} \cdot mn + log(2\sqrt{N}/\delta)}{2N}}$$

Complexity scales as O(mn)

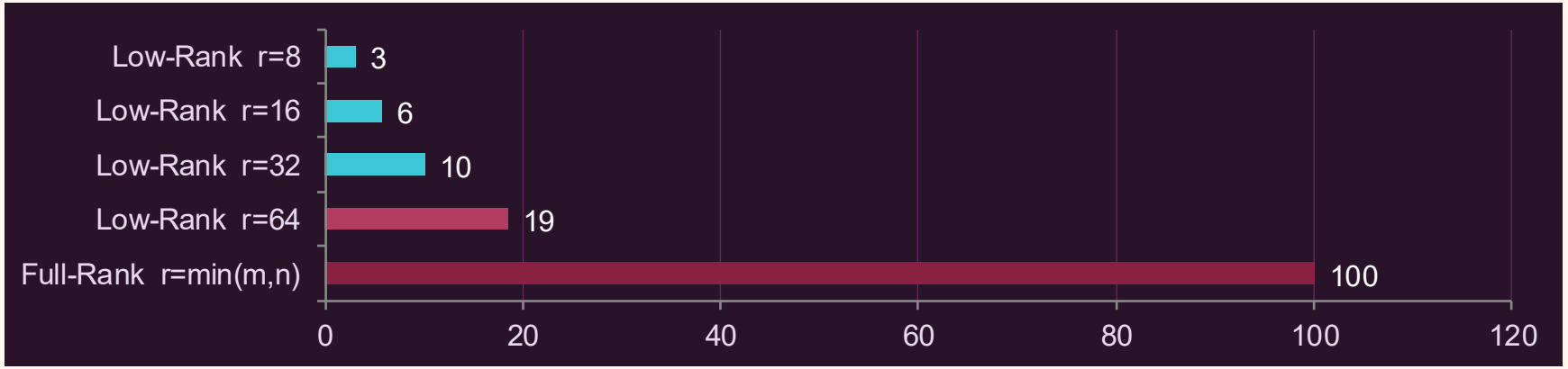
Low-Rank (Ours)

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{C_{\max} \cdot r(m+n) + log(2\sqrt{N}/\delta)}{2N}}$$

Complexity scales as O(r(m+n))

Complexity Ratio:

$$\frac{\text{Complexity}(Q_{LR})}{\text{Complexity}(Q_{full})} = \sqrt{r\left(\frac{1}{m} + \frac{1}{n}\right)} \ll 1 \quad \text{when} \quad r \ll \min(m, n)$$



Relative complexity (%, lower is better)

Practical Numbers (m=512, n=512)

r=64:	25× fewer
r=16:	64× fewer
r=8:	128× fewer

The framework is not architecture-specific.

MLP

standard dense layers
$W_\ell = A_\ell B_\ell^T$

LSTM

input-to-hidden + hidden-to-hidden
sample factors once per batch

Transformer

position-wise factorization
rank r=16 in SST-2 experiments

Drop-in variational layers make the geometry portable across model families.

EXPERIMENTAL DESIGN

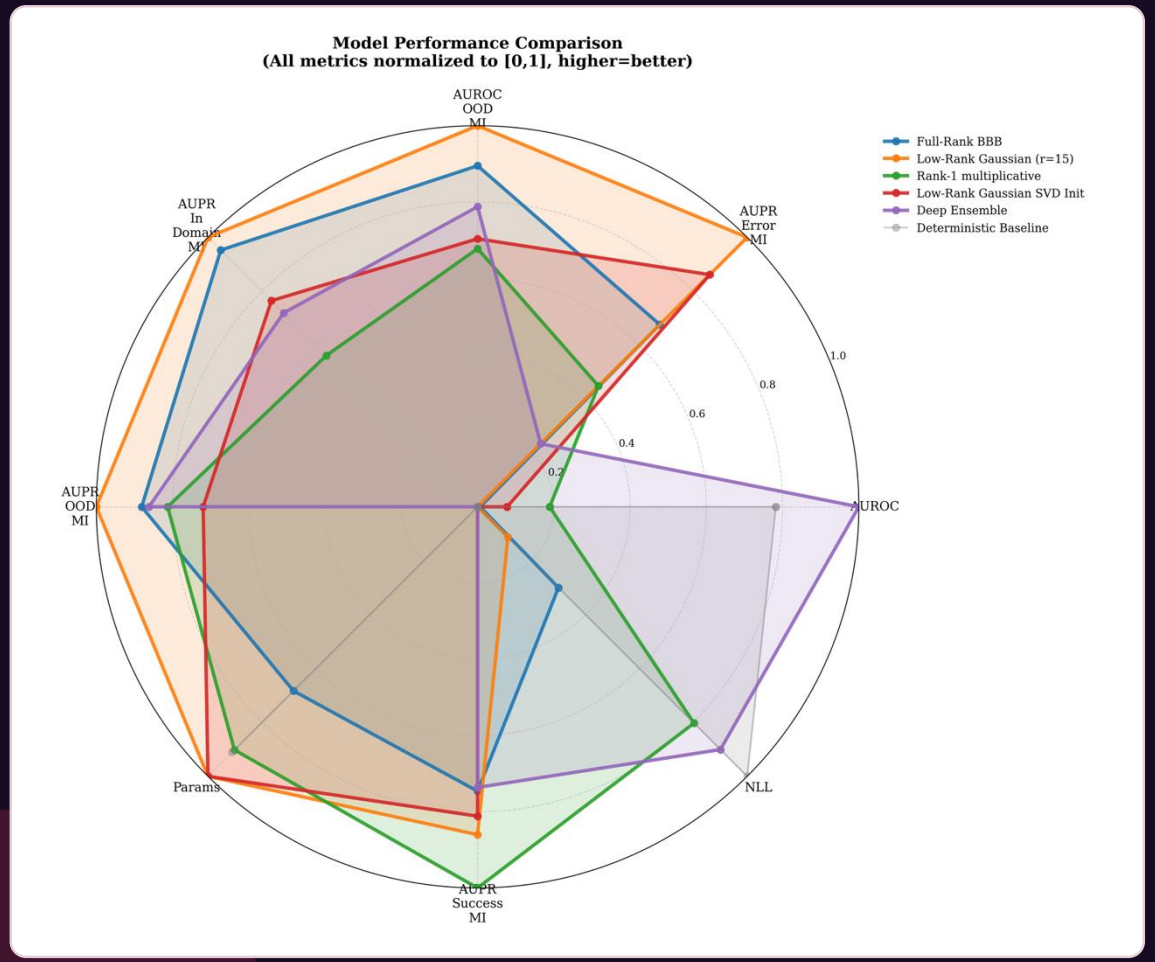
Three architectures. Three data regimes. Six uncertainty baselines.

Dataset	Architecture	Shift	Rank
MIMIC-III ICU mortality	2-layer MLP	adult ICU → newborn ICU	r=15
Beijing PM2.5 forecasting	2-layer LSTM	Beijing → Guangzhou	r=14/20
SST-2 sentiment	4-layer Transformer	movie reviews → AGNews	r=16

Baselines: deterministic, Deep Ensemble, Full-Rank BBB, Low-Rank random init, Low-Rank SVD init, Rank-1 multiplicative; SWAG added as supplementary comparator.

RESULT: MIMIC-III

On clinical shift, low-rank gives the strongest OOD uncertainty.



0.802
AUC-OOD, best overall

0.788
AUPR-OOD, best overall

0.824
AUPR-In, best overall

Nuance

Deep Ensemble keeps stronger in-domain AUROC and NLL. Low-rank prioritizes epistemic separation under shift.

Parameter reduction: 70% fewer than Full-Rank BBB; 88% fewer than Deep Ensemble.

MIMIC-III: ICU Mortality Prediction (MLP)

Model	AUROC↑	AUPR-Err↑	AUC-OOD↑	AUPR-OOD↑	AUPR-In↑	NLL↓	Params↓
Deterministic	.922	.145	.500	.544	.456	.284	22.4K
Deep Ensemble	.929	.237	.738	.754	.721	.300	112K
Full-Rank BBB	.895	.412	.770	.759	.807	.401	44.8K
Low-Rank (ours)	.895	.540	**.802**	**.788**	**.824**	.433	13.6K

★ **Low-Rank achieves BEST OOD detection with 88% fewer parameters than Deep Ensemble**

70%

fewer params than
Full-Rank BBB

88%

fewer params than
Deep Ensemble

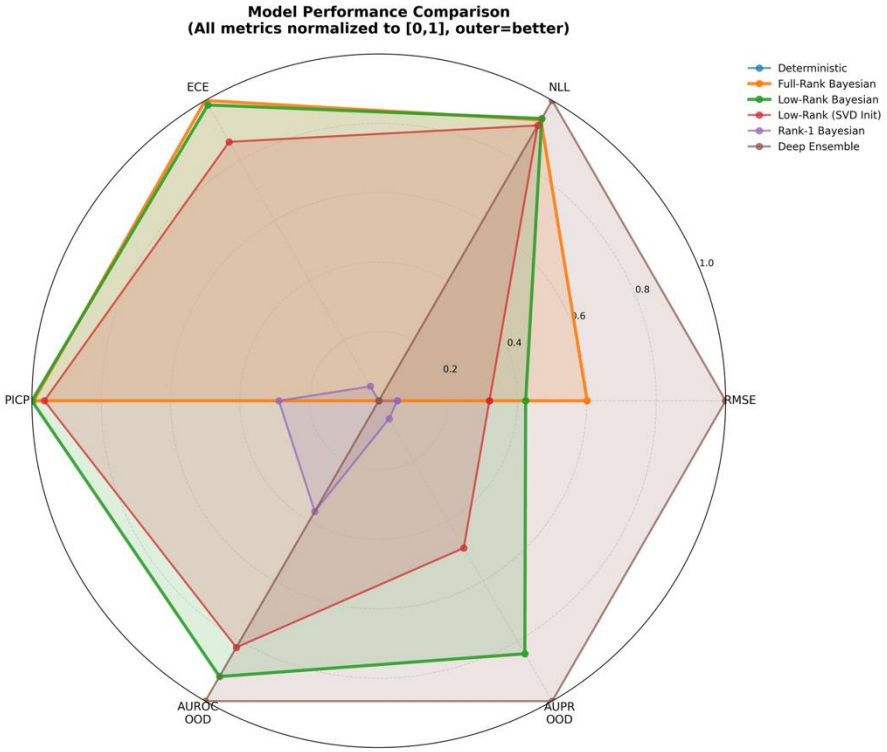
0.540

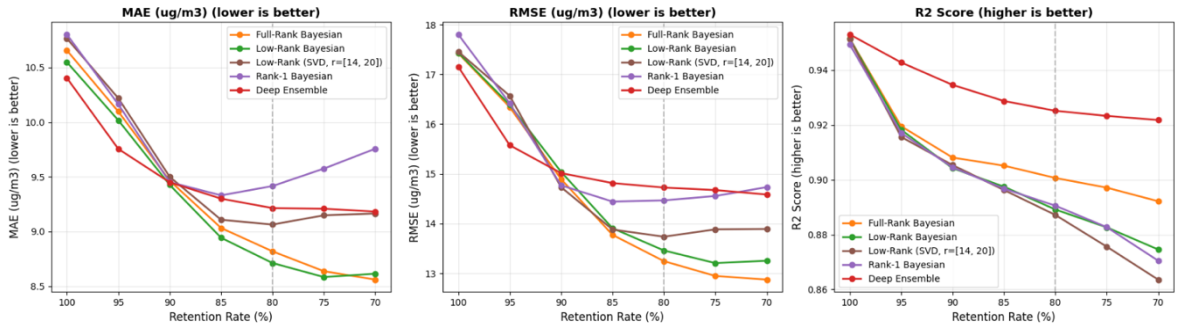
AUPR-Error
(best in class)

arradiat.github.io/projects/singular-bnn · github.com/arradiat/SBNN · arXiv:2602.00387

RESULT: BEIJING PM2.5

For time-series forecasting, uncertainty quality shows up in coverage and abstention.





Why Low-Rank wins selective prediction

Structured rank-r correlations yield **better-calibrated abstention** than mean-field or Rank-1 or Deep-Ensembles when the model abstains on the 20% lowest-confidence inputs, residual MAE drops 17.4%, beating Deep Ensembles at 6.6× fewer params.

Beijing $PM_{2.5}$: Time-Series Forecasting (LSTM)

Model	MAE↓	ECE↓	PICP↑	AUROC-OOD↑	AUPR-OOD↑	Params↓
Deterministic	10.79	—	—	0.500	0.500	33K
Full-Rank BBB	10.55	0.111	0.788	0.492	0.743	132K
Low-Rank (ours)	**10.63**	**0.114**	**0.790**	**0.710**	**0.861**	**47K**
Rank-1 Mult.	10.80	0.307	0.449	0.580	0.751	66K
Deep Ensemble	10.45	0.317	0.310	0.730	0.883	330K

0.790

PICP Coverage
Best prediction
interval calibration

17.4%

MAE Reduction
At 80% retention
(selective prediction)

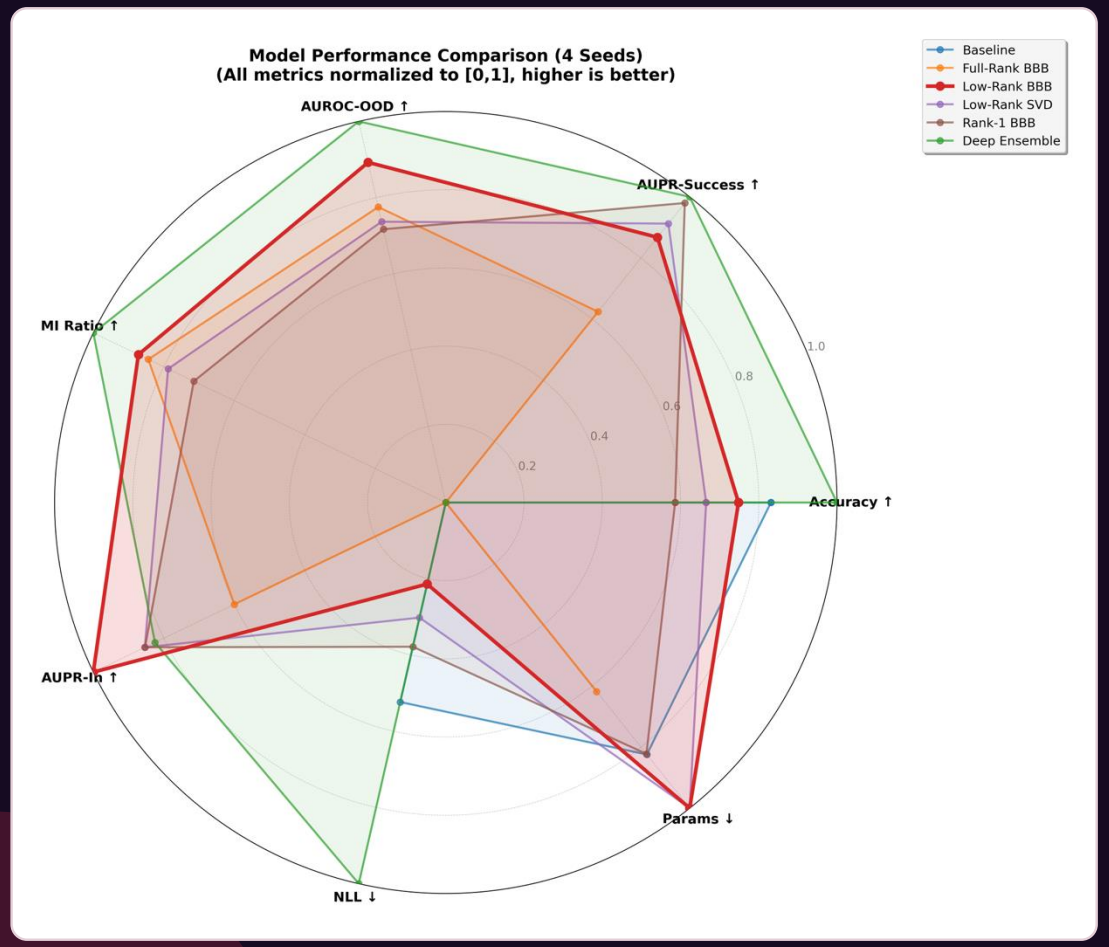
64%

Param Reduction
vs Full-Rank BBB,
with better OOD

Key finding: Bayesian methods outperform Deep Ensemble when filtering uncertain predictions. Low-Rank achieves the lowest MAE at 80% retention (8.71 vs 9.21 for Deep Ensemble), demonstrating that structured correlations improve selective prediction quality.

RESULT: SST-2 TRANSFORMER

At Transformer scale, the parameter story becomes decisive.



1.5M
Low-Rank BBB parameters

13×
fewer than Full-Rank BBB

33×
fewer than Deep Ensemble

Performance profile

Low-Rank BBB: 0.806 accuracy, best AUPR-In, second-best AUROC-OOD. Deep Ensemble remains strongest overall but costs 49.6M parameters.

Training time

Low-Rank trains in **8.2** min vs **23.1** for Full-Rank BBB and **64.7** for Deep Ensemble.

SST-2 Sentiment — Transformer Efficiency

Model	Acc↑	NLL↓	AUROC-OOD↑	MI Ratio↑	AUPR-In↑	Params↓	Time
Deterministic	.812	.490	.500	.00	.102	9.9M	7.7min
Deep Ensemble	.825	.434	.657	1.55	.267	49.6M	64.7min
Full-Rank BBB	.752	.552	.622	1.31	.222	19.8M	23.1min
Low-Rank (ours)	**.806**	**.527**	**.640**	**1.35**	**.302**	**1.5M**	**8.2min**

33×

fewer params than
Deep Ensemble

13×

fewer params than
Full-Rank BBB

8×

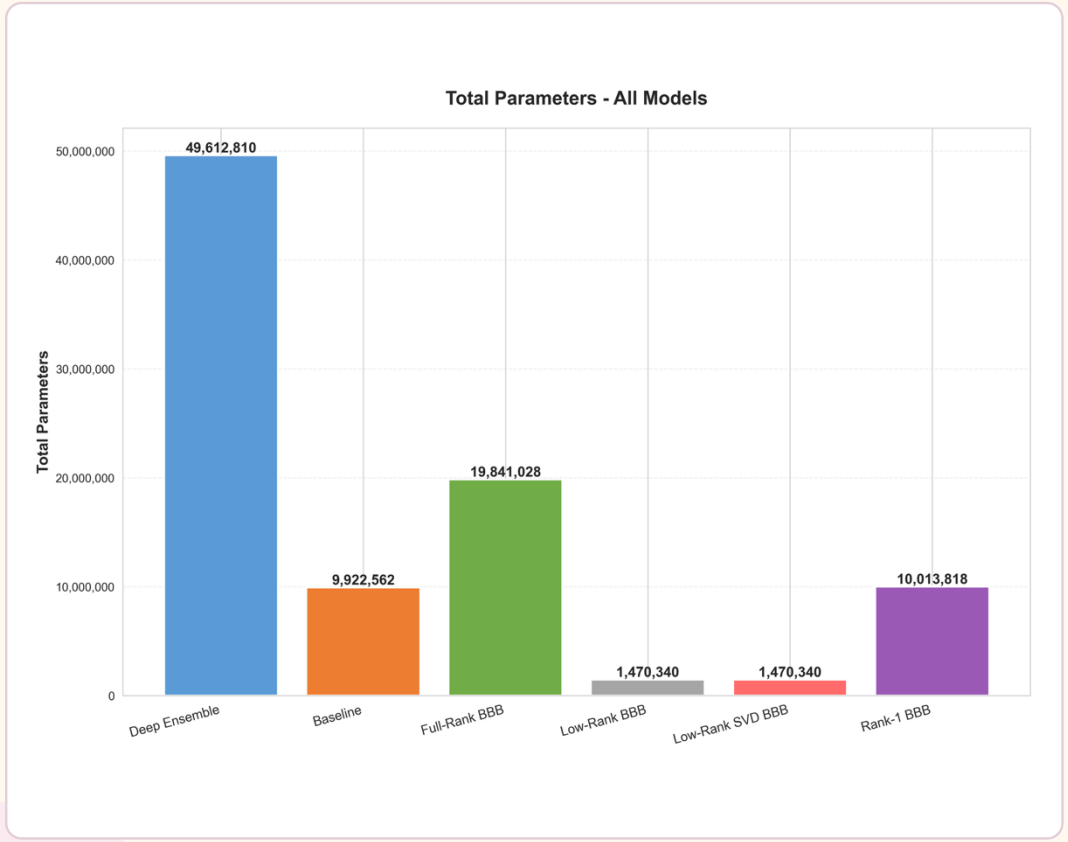
faster training
vs Deep Ensemble

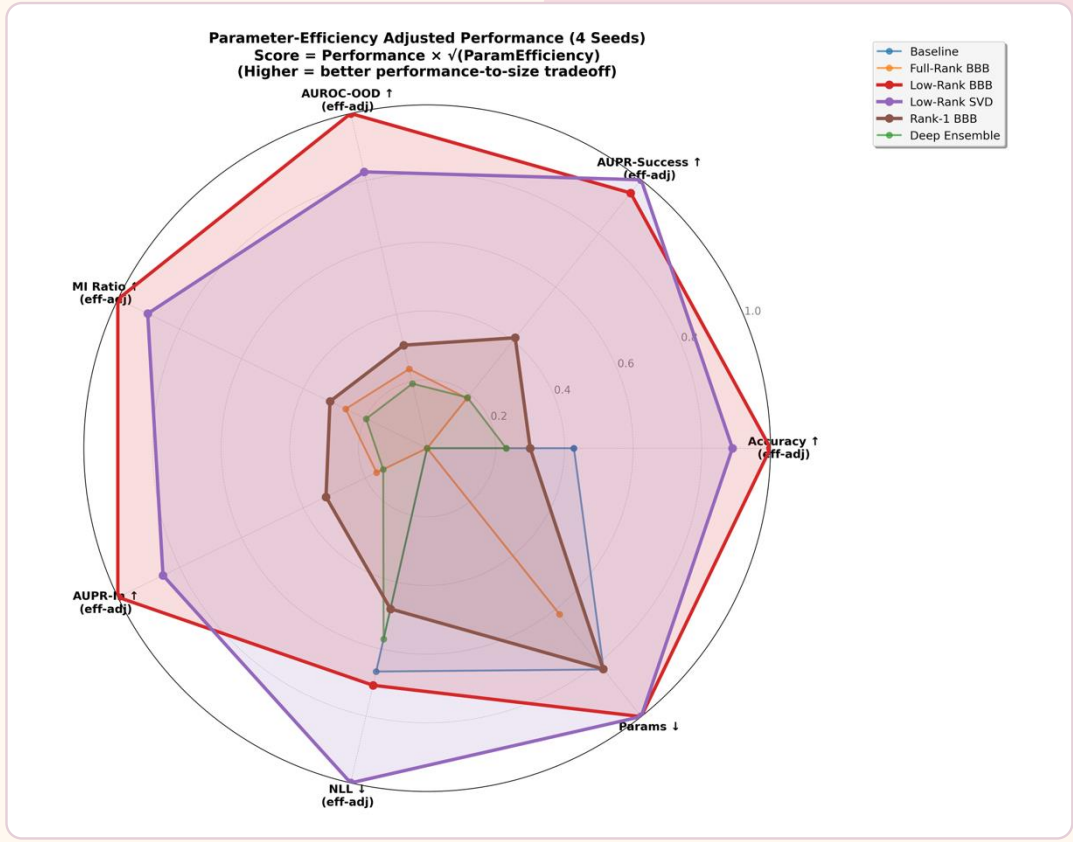
⚡ **Controlled GPU Profiling** — *same GPU, fixed steps*

Model	Params	Peak Memory	Epoch Time
★ **Low-Rank BBB (ours)**	**1.47M**	**357.5 MB**	**5.88s**
Full-Rank BBB	19.84M	721.1 MB	6.45s
Deep Ensemble	49.61M	670.1 MB	18.99s

arradiat.github.io/projects/singular-bnn · github.com/arradiat/SBNN · arXiv:2602.00387

The quality-efficiency frontier changes.





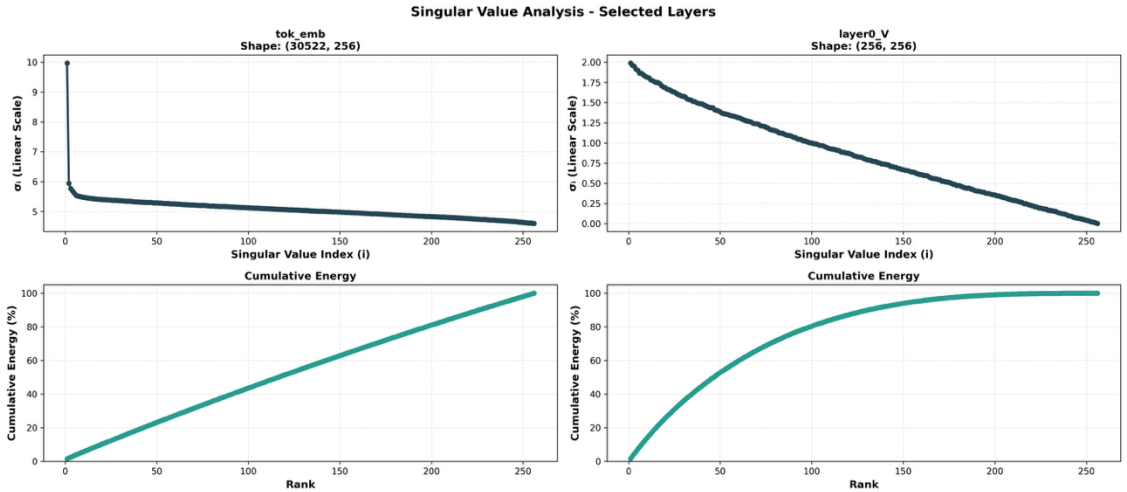
The result is not "low-rank always wins every metric"; it is that useful Bayesian uncertainty becomes plausible at modern parameter scales.

Observed singular value decay supports the rank-r constraint.

When layer spectra decay quickly, the manifold is not an arbitrary bottleneck. It is a structured approximation class.

Theory connection

EYM tail singular values determine rank approximation error; rank ablations and spectra guide practical rank selection.



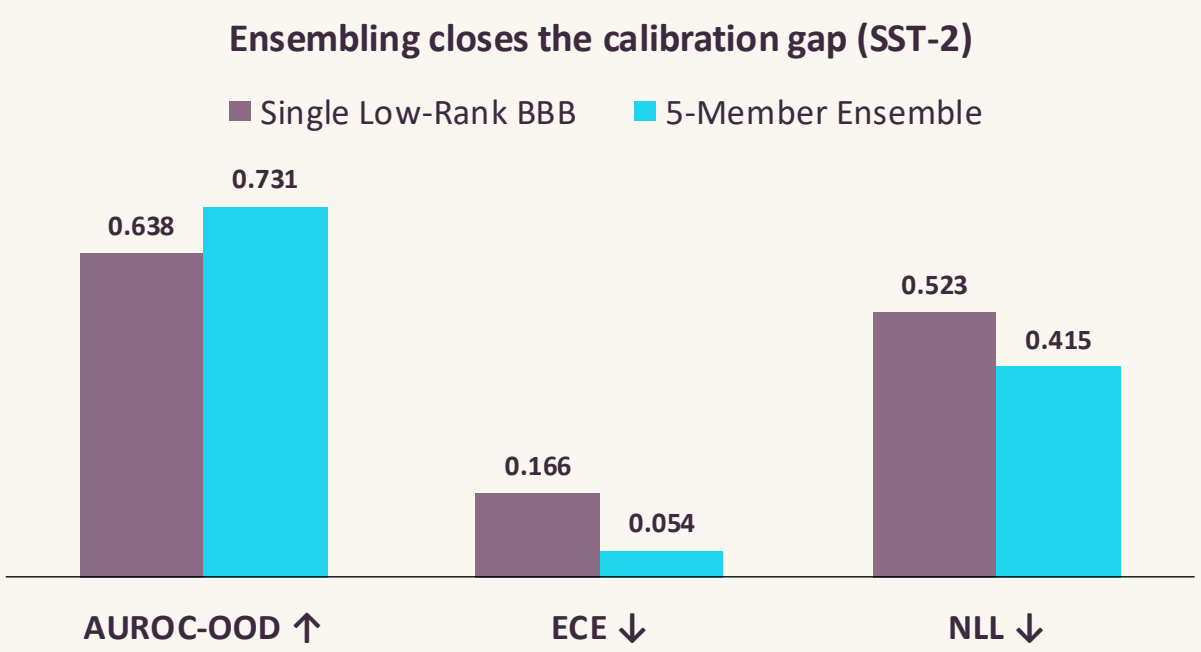
COMPARATOR: SWAG

A strong posterior baseline does not overturn the central tradeoff.

Setting	SWAG strength	Low-rank advantage
SST-2	Accuracy tied at 0.808 vs 0.806	Better NLL/OOD; 1.47M params vs 208.37M
MIMIC	Higher in-domain AUROC/NLL	OOD MI metrics: 0.802 / 0.788 vs 0.634 / 0.680
Beijing	Higher coverage	Much narrower/costlier tradeoff; stronger OOD for low-rank

SWAG is a credible baseline. However, SBNN's geometry gives a better quality-efficiency path in the paper's target regimes.

Key Insight: Calibration – OOD Detection Tradeoff



MIMIC-III

Low-rank improves OOD despite weaker NLL than Deep Ensembles

SST-2

Deep Ensemble stronger on both NLL and OOD; Low-rank competitive at 33× fewer params

Beijing

Low-rank best calibration, coverage, selective prediction — OOD advantage secondary

Hypothesis: The rank constraint on $\mathcal{R}\_r$ enforces structured weight correlations (Lemma 3.2) that maintain broader epistemic uncertainty distributions. This benefits abstention, coverage, and OOD awareness at the cost of predictive sharpness (NLL). The tradeoff is task-dependent.

✦ **Ensembling closes the gap** — *5 low-rank members, still cheaper than one full-rank BBB*

AUROC-OOD **0.638 → 0.731** | ECE **0.166 → 0.054** | NLL **0.523 → 0.415**

HONEST TAKEAWAY

The empirical message is not compression alone.

Where low-rank shines

OOD separation on MIMIC; coverage/selective prediction on Beijing; efficiency-adjusted uncertainty on SST-2.

Where ensembles remain strong

In-distribution likelihood and sharpness can favor Deep Ensembles; calibration gaps remain in some settings.

Why that matters

Trustworthy AI often needs useful uncertainty under shift, abstention, and deployment constraints, not only marginal NLL.

SBNNs make that goal more practical: structured uncertainty, lower cost, and scalable to modern architectures.

A platform for scalable Bayesian posteriors in modern architectures.

Adaptive rank selection

Ranks chosen by spectra, validation tradeoffs, or sparse priors rather than fixed grids.

Bayesian Transformers

Structured uncertainty for large sequence models without ensemble-scale cost.

Beyond Gaussian factors

Laplace, spike-and-slab, and richer factor distributions while keeping singular support.

Trustworthy AI

Uncertainty that supports abstention, coverage, OOD awareness, and constrained deployment.

Singular posterior geometry for scalable Bayesian deep learning.

Singular Bayesian Neural Networks

arradiat.github.io/projects/singular-bnn

arxiv.org/abs/2602.00387

github.com/arradiat/SBNN

mame.toure@mail.mcgill.ca

Thank you