Scaling Laws for Uncertainty in Deep Learning

Mattia Rosso[1], Simone Rossi[2], Giulio Franzese[2], Markus Heinonen[3], Maurizio Filippone[1]
[1]KAUST (Saudi Arabia), [2]EURECOM (France), [3]Aalto University (Finland)

Introduction and Motivation

Why Uncertainty Quantification?

- Critical for safety-critical applications: medical diagnosis, autonomous driving, financial decision-making
- Enables reliable AI systems that know when they don't know
- Essential for trustworthy predictions beyond point estimates

Uncertainty Decomposition:

- **Aleatoric (AU):** Irreducible noise inherent in data
  $\mathrm{AU}(x) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{H}[p(y \mid x, \boldsymbol{\theta}^{(k)})]$
- **Epistemic (EU):** Model uncertainty from limited data/parameters
  $\mathrm{EU}(x) = \mathbb{H}\left[\frac{1}{K}\sum_{k=1}^{K} p(y \mid x, \boldsymbol{\theta}^{(k)})\right] - \mathrm{AU}(x)$
- EU can be reduced with more data or better models

Research Questions

Neural scaling laws for loss well-established (Kaplan et al., 2020), but:

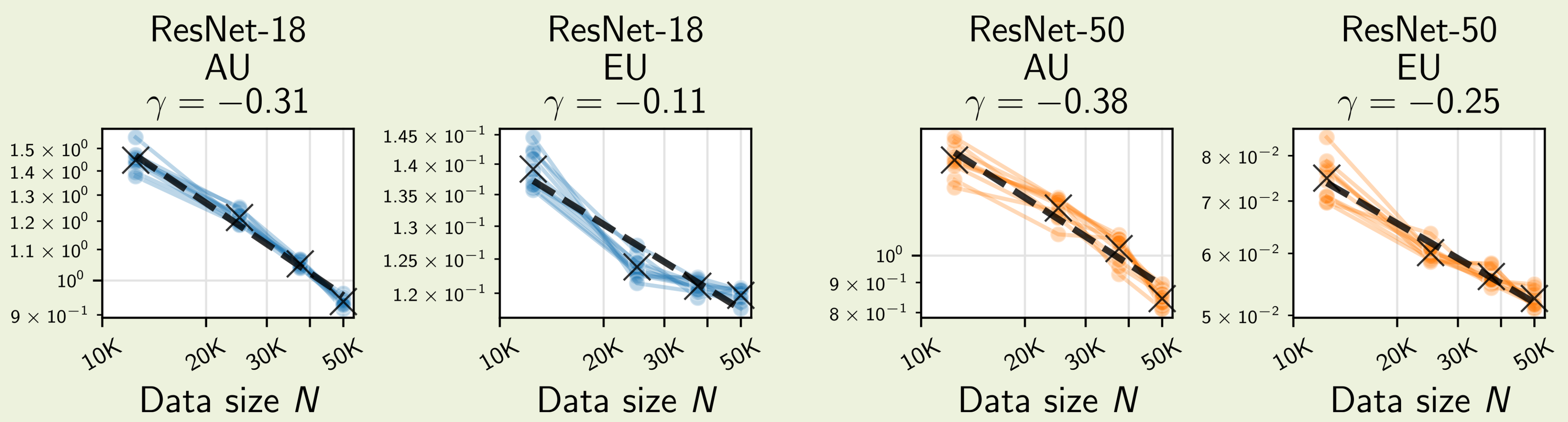
- *How do uncertainties scale with data (and model) size?*
- Can we predict uncertainty behavior at scale?
- Do different UQ methods exhibit universal scaling patterns?

Key Contributions

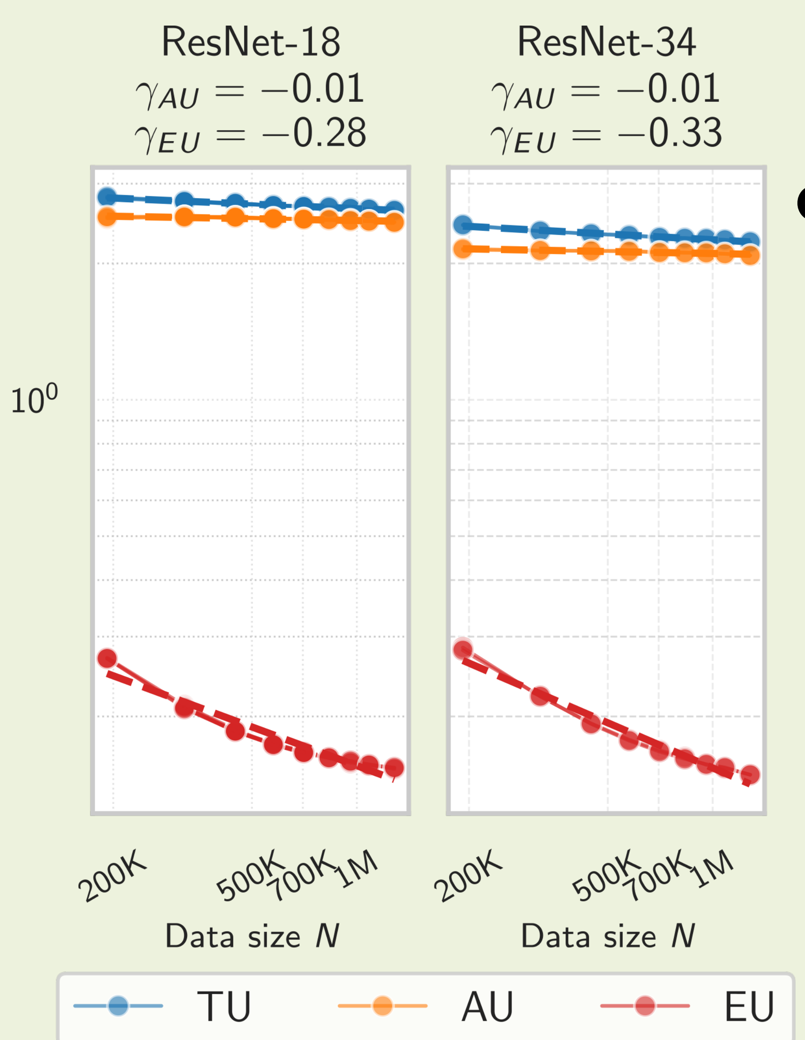
- **Large systematic study** of uncertainty scaling laws across architectures and datasets
- **Power-law behavior** consistently observed: Uncertainty $\propto N^{\gamma}$
- **Different scaling exponents** for different Bayesian/ensemble methods
- **Open questions** on the impact of DL optimization strategies (Sharpness-Aware Minimization, LoRA fine-tuning)

Does uncertainty predictably follow scaling laws?

Setup: CIFAR100 with Monte Carlo Dropout (MCD)



Setup: ImageNet with MCD



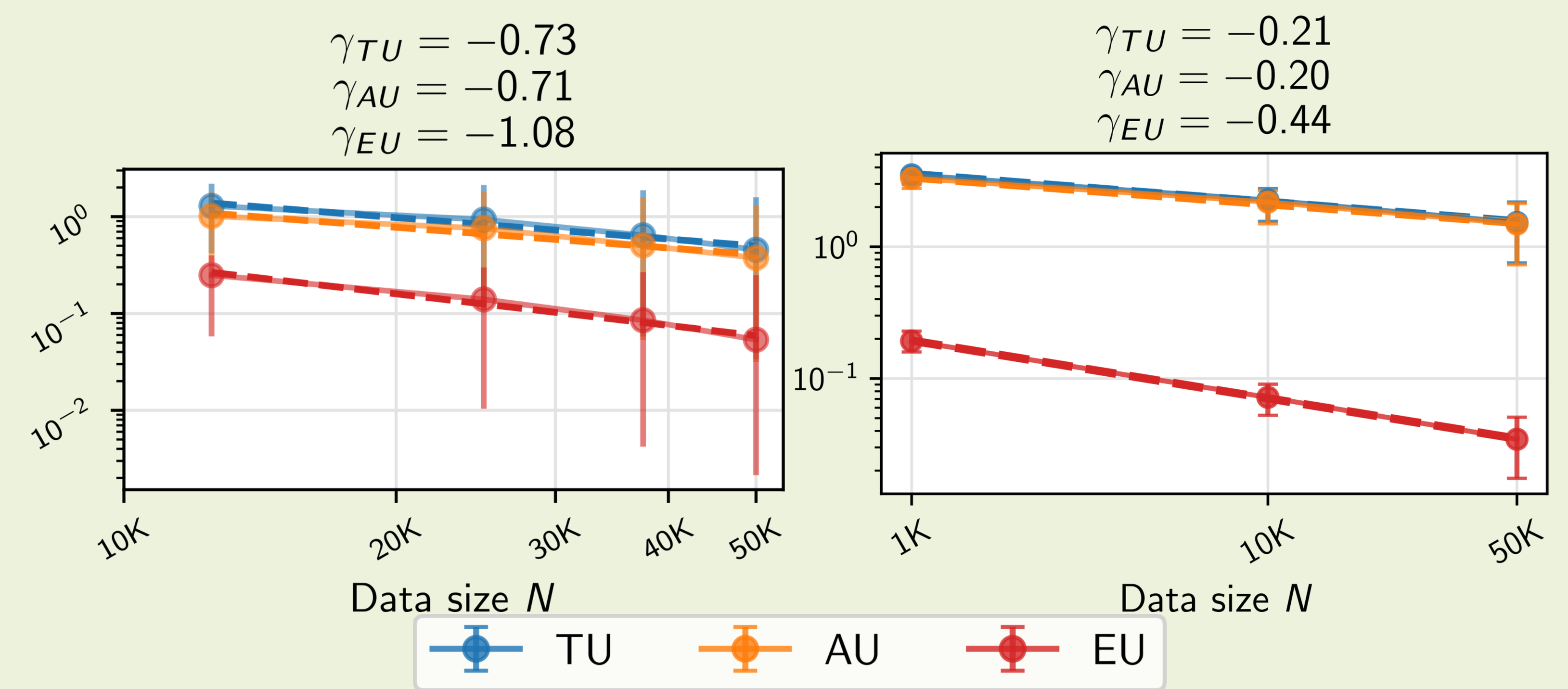
Observations:

- Clear power-law scaling for AU and EU
- EU decreases *slower* than AU: for large data, model uncertainty remains significant
- On larger dataset (ImageNet), AU doesn't decrease (expected due to inherent noise floor)

Take a picture to download the full paper

Are scaling laws universal across UQ methods?

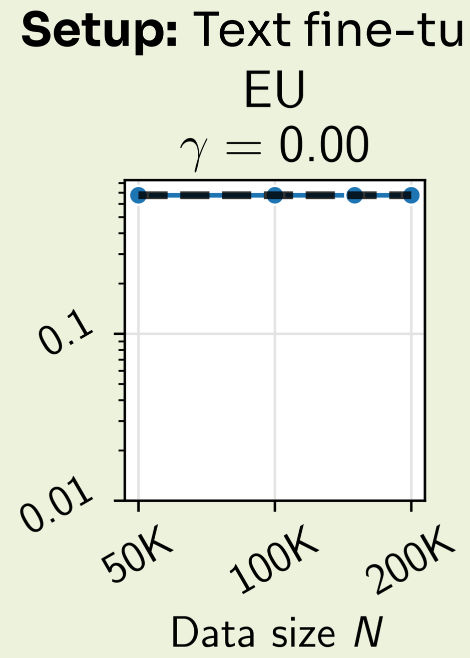
Setup: CIFAR100 with Deep Ensembles *(left)* and MCMC *(right)*



- Power-law scaling holds across UQ methods ...
- but exponents vary: DE shows fastest EU decrease (ensemble diversity diminishes)

Are scaling laws affected by optimization strategies?

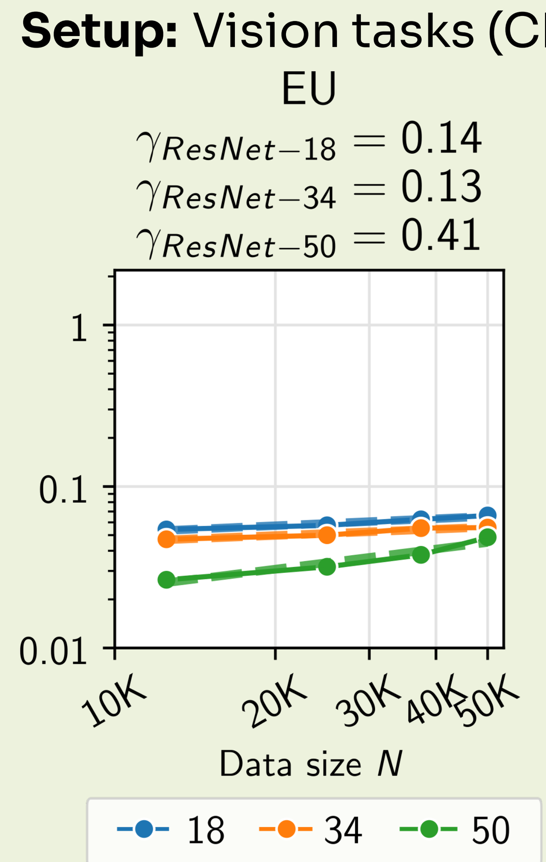
Setup: Text fine-tuning with Bayesian LoRA (Laplace approximation)



Observations:

- No meaningful scaling observed ($\gamma \approx 0$)
- Pre-training data dominates uncertainty characteristics also during fine-tuning

Setup: Vision tasks (CIFAR100) with MCD and SAM optimizer



Observations:

- Contrary to standard training, EU *increases* with data!
- Flatter minima preserve functional diversity

Concluding Remarks

- Uncertainty scaling is **predictable** in deep learning, but **method-dependent**
- Epistemic uncertainty **persists even with large datasets**
- Optimization strategy fundamentally affects uncertainty landscape

Open questions

- Theoretical characterization of scaling exponents for different UQ methods
- Applying uncertainty scaling laws to active learning and data acquisition