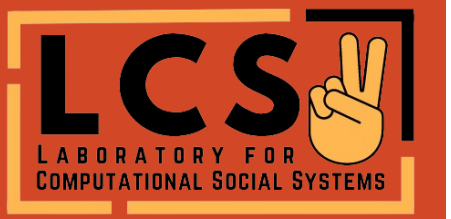
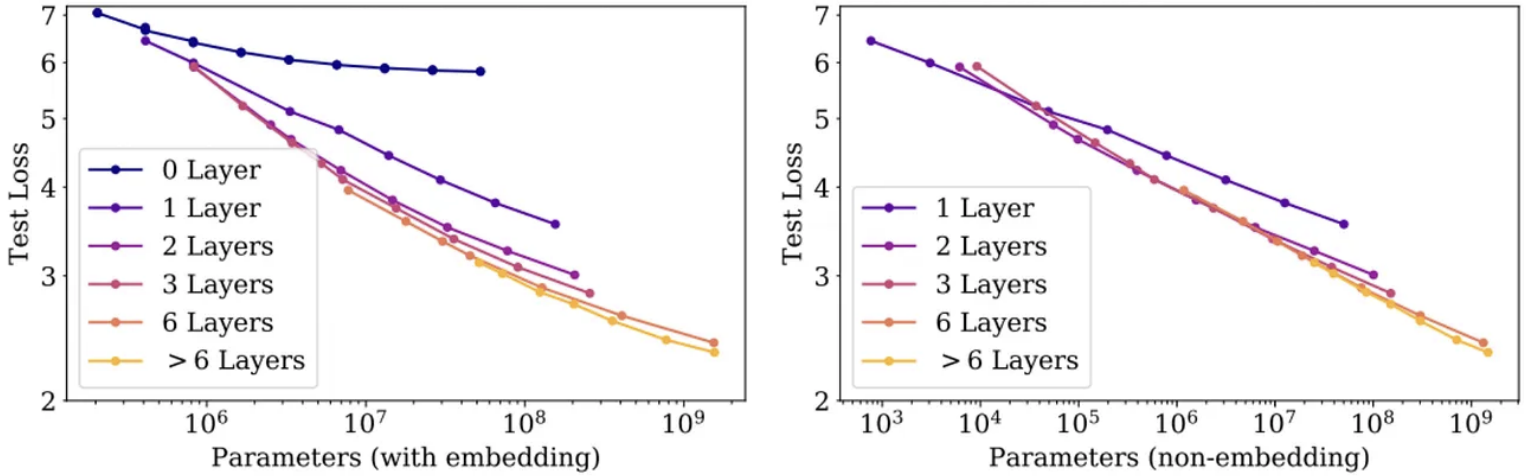

# Enough of Scaling LLMs! Let's Focus on Downscaling

Yash Goel, Ayan Sengupta, Tanmoy Chakraborty

# What is Model Scaling?

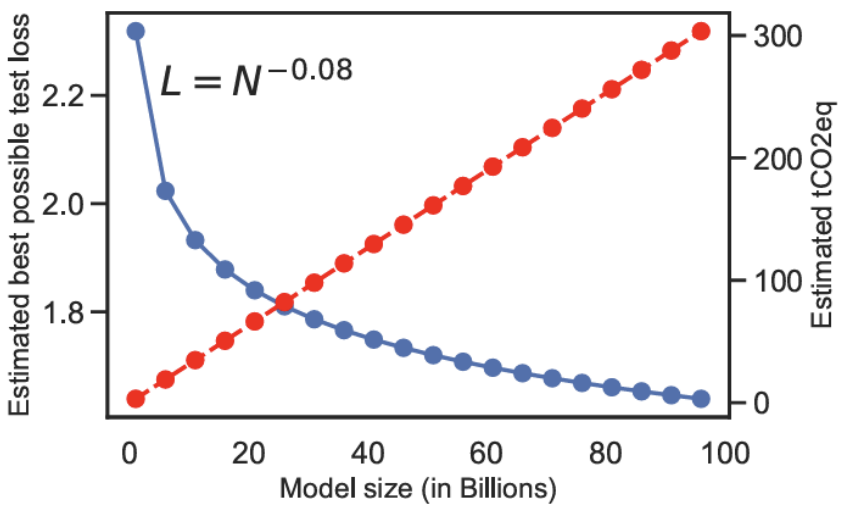


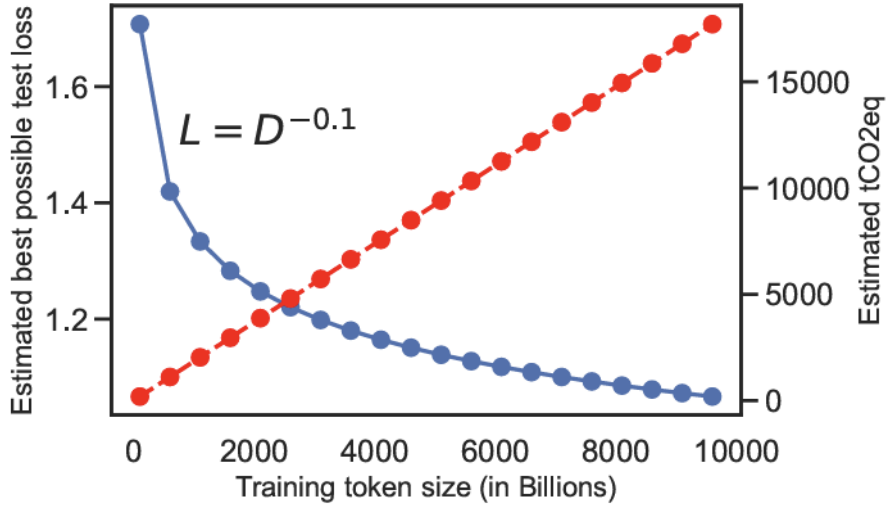

Scaling laws provide an analytical framework supporting that training a big neural network on a large dataset can lead to significantly low test error.

Past 4-5 years have seen an exponential growth in the number of scaling laws studies, predominantly for LLMs pretraining and fine-tuning.

# Can We Scale Up Scaling Laws?

Widely adopted LLM pre-training scaling laws (Kaplan, Chinchilla) follow power-law $\approx D^{\alpha}N^{\beta}$, with the model size $N$ and pre-training data size $D$.




Even for a sub-linear performance improvement, scaling laws suggest exponential increase in compute. 10% reduction in test-loss incurs >300% increase in compute.

# Scaling Laws Beyond Chinchilla

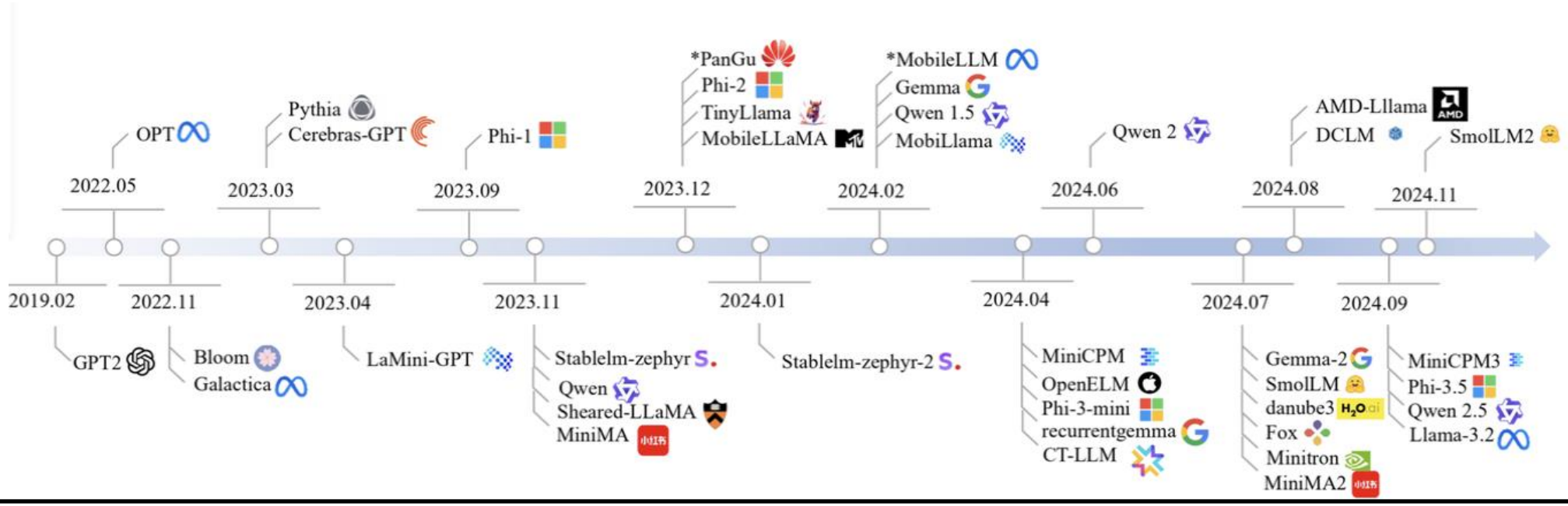
Recent works show that Kaplan/Chinchilla laws, which assume uniform scaling of model size and data, fail to account for:

Mixture granularity – Small, fine-grained experts outperform monolithic scaling

Inference-level scaling – more model calls don’t always improve accuracy; optimal call count matters

Task-specific saturation – downstream performance saturate early, even if model scale increases

# Rise of Small and Efficient Language Models



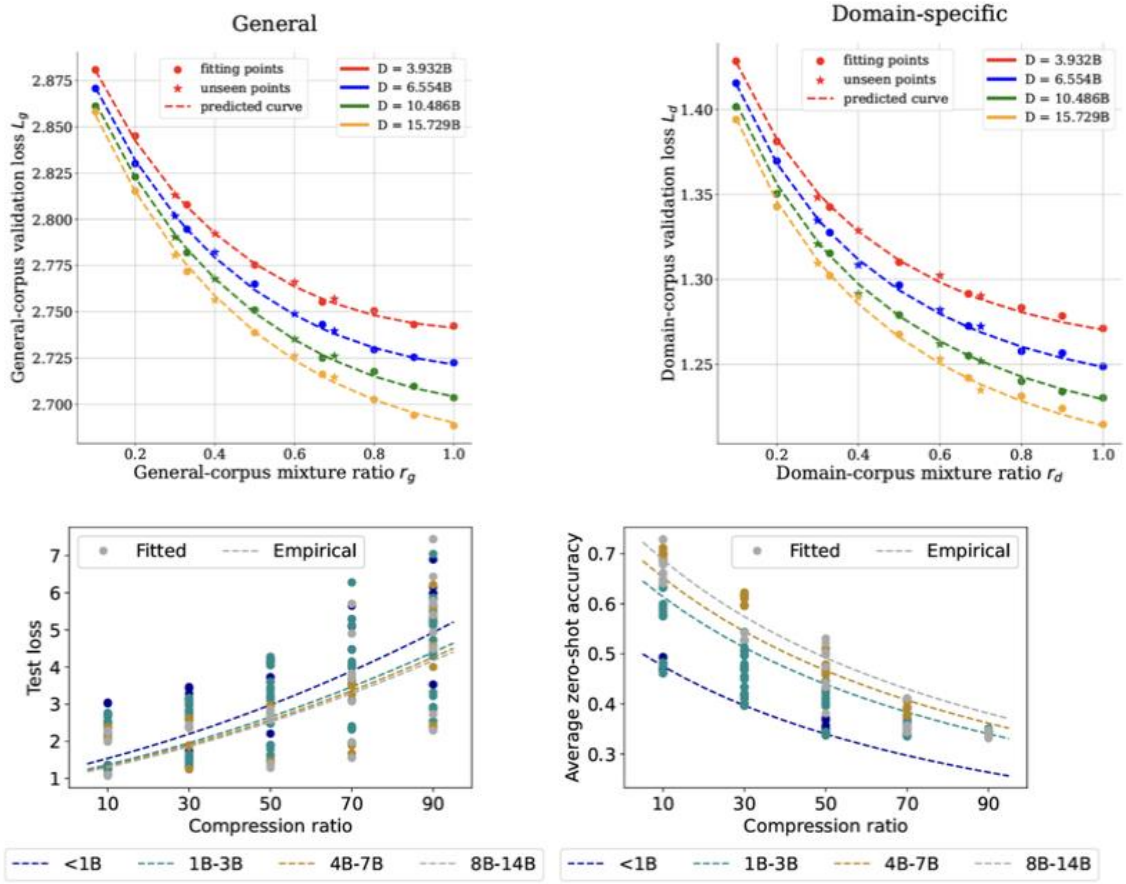

Development of small and efficient large pre-trained models have dramatically accelerated in recent 1-2 years, with SLMs exhibiting better performance-speedup trade-off.

Chinchilla scaling law suggests that SLMs can achieve better scaling with more pre-training data.

# Why Do Small & Efficient Models Work?

High-quality data is better than bigger data volume. Domain-specific continual pre-training is better than generic pre-training.

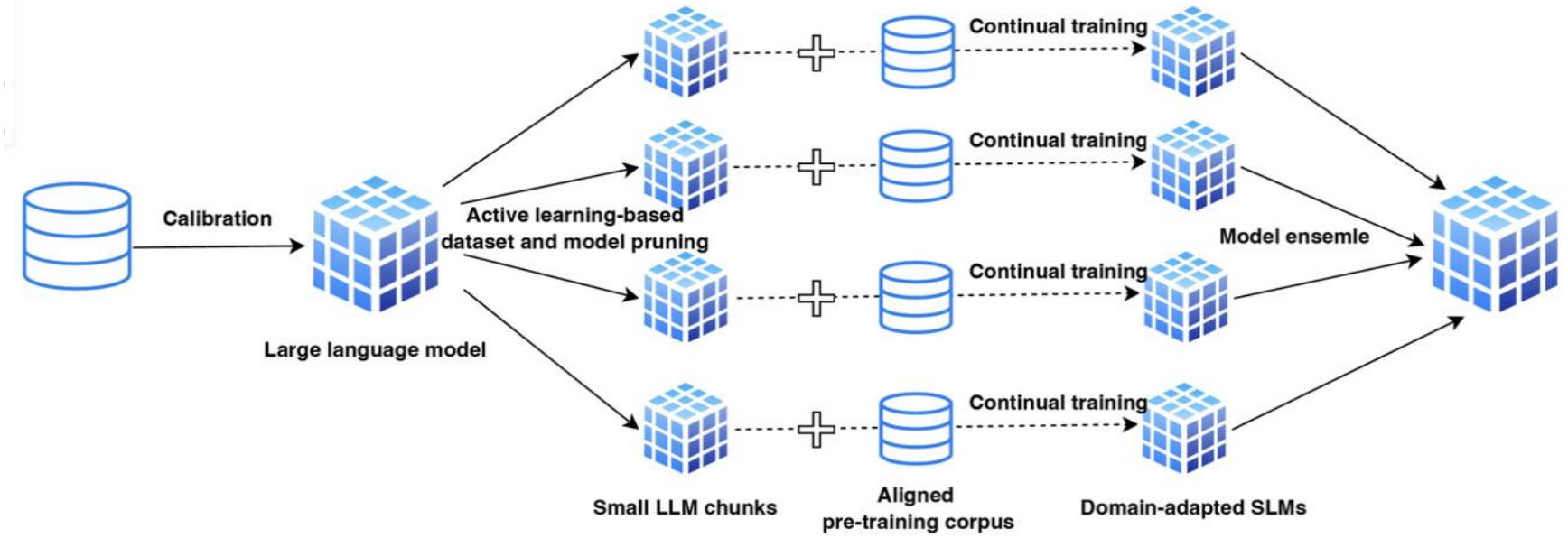
Better compression of knowledge




Knowledge distillation outperforms supervised pre-training when total compute stays below a model-size dependent threshold.

# *Downscaling* for Better Trade-off

Compress → Specialize → Combine




1. Prune models + data
2. Train on aligned domain data
3. Ensemble into unified system

# Theoretical Guarantee of Downscaling Law

Theoretical guarantee can be derived from model compression, model ensemble and domain-adapted pre-training laws -

For $n \in \mathbb{Z}_+$ satisfying $\left(\frac{n^a - 1}{n^{a+\gamma}}\right)(n-1)^\gamma \geq \frac{1}{bN_0{}^\delta}$, the expected ensemble loss $L_0 - b + \frac{b}{n^a} < L_0$, with $N_0$ being the model size of the original (uncompressed) model, $L_0$ being the test loss of the uncompressed model. $a, b, \gamma, \delta$ are fitting constants could be obtained from Lobacheva et al., and Chen at al.,

**Example –** Putting $a = 0.83, b = 0.83, \gamma = 1.08, \quad \delta = 0.29$ for LLaMA-3-8B model, we obtain $n > 7$. i.e., for with 8 or more compressed models each with size $< 1B$, we can **achieve better test loss than the original LLaMA-8B model.**

# Our Position: Why Downscaling Deserves the Spotlight

- **Scaling is ecologically and economically unsustainable:** The carbon cost for incremental performance gains grows nonlinearly with scale, threatening environmental goals and making large-scale models increasingly inaccessible.
- **Inequitable compute access limits research diversity**: Continued scaling amplifies disparity; smaller, resource-constrained institutions are left behind unless we prioritize efficient small models.
- **Overparameterization harms performance generality:** Larger models often overfit niche domains, while carefully downscaled models can be more robust and adaptable to target tasks.
- **Downscaling accelerates research iteration:** Smaller models train faster and consume less compute, enabling rapid experimentation, debugging, and reproducibility.
- **Encourages methodological innovation:** Focusing on downscaling drives progress in pruning, quantization, KD, routing, and dataset efficiency, leading to richer modeling techniques than raw scale.