

Holistic scaling laws for optimal Mixture-of-Experts architecture optimization

Weilin Wan, Jingtao Han, Debing Zhang, Weizhong Zhang, Cheng Jin

Fudan University · Hi Lab, Xiaohongshu Inc. · Project Numina

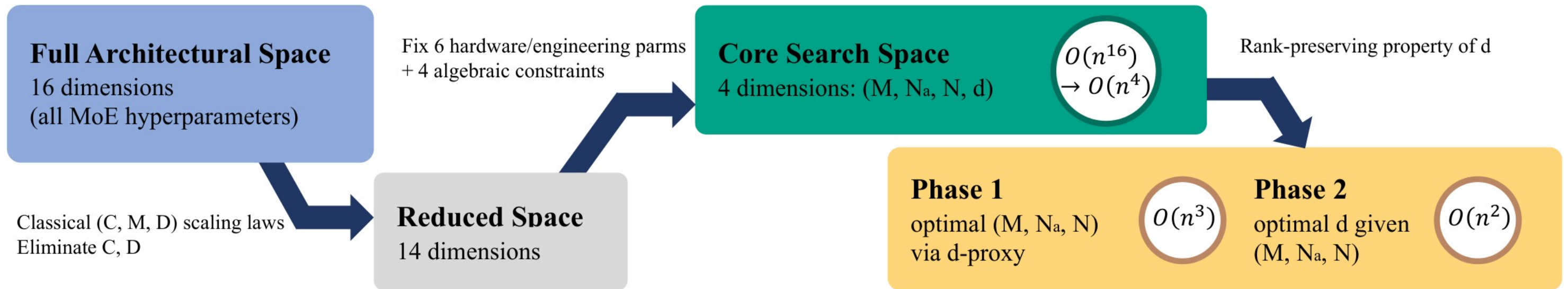
MoE design needs three resource measures

- M Computation per token
- N_a Active parameters: weight access
- N Total parameters: storage

Sparse routing separates these costs.

Similar compute can hide different parameter-storage requirements.

Two-phase architecture search

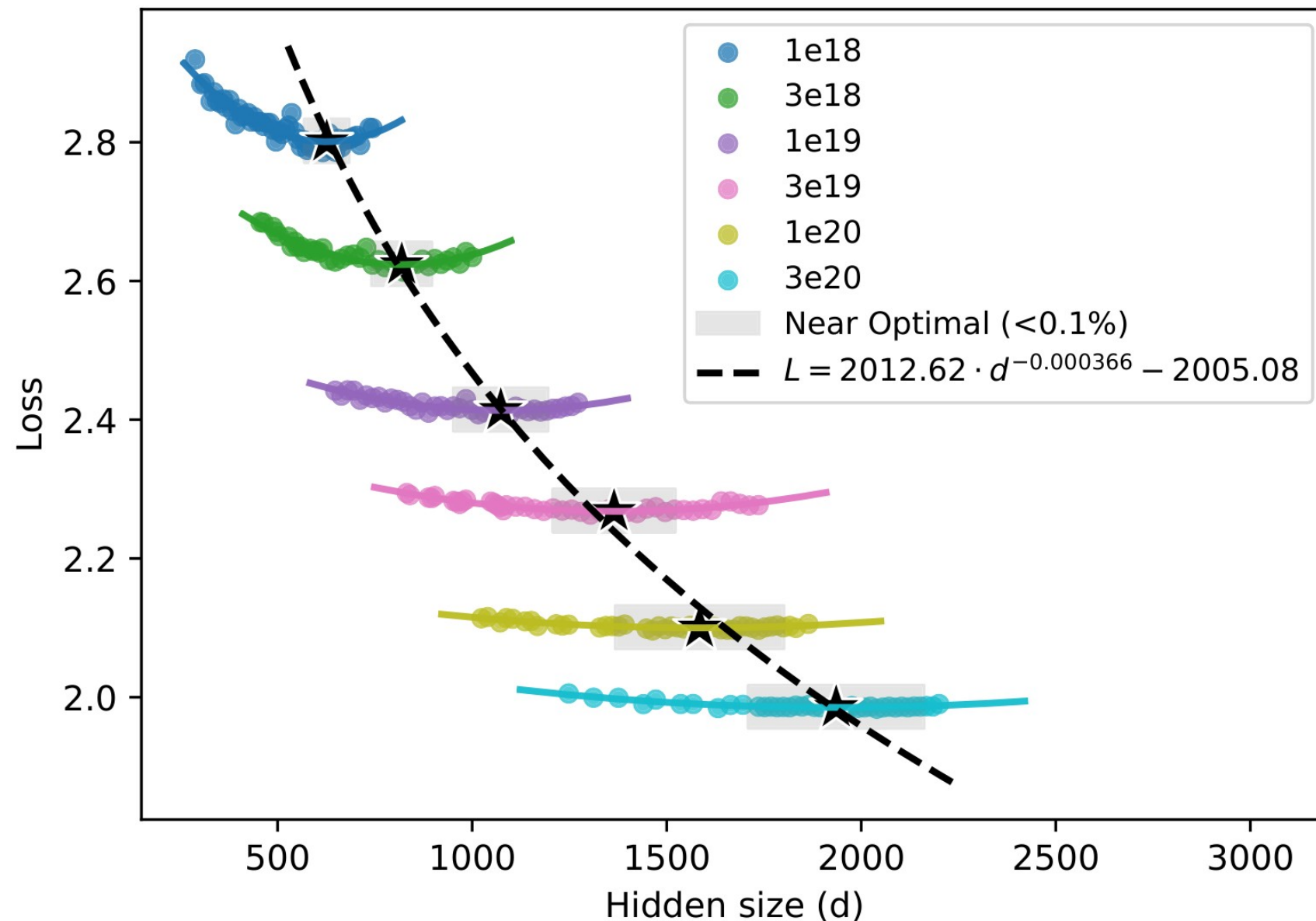


670+ MoE models across six compute scales: 10^{18} – 3×10^{20} FLOPs

Median-d proxy: Pearson $r = 0.9793$, Spearman $\rho = 0.9758$
10 configurations, 144 experiments at $C = 10^{18}$ FLOPs

Grid costs assume fixed architectural choices and empirical rank preservation.

Hidden size, loss and the near-optimal band



Shape matters

Within each curve, (M, N_a, N) are fixed. Hidden size still changes the loss.

Scale changes the optimum

Black stars mark fitted minima. Preferred hidden size increases across the tested budgets.

Grey bands: within 0.1% of minimum loss. Wider bands allow more design flexibility.

Using the fitted architecture

1. Set a training compute budget C .
2. Obtain (M, D) , N_a and d from the fits.
3. Choose $q = N/N_a$, then set $N = qN_a$.
4. Solve the structural constraints and satisfy hardware-alignment rules.

Scope of the evidence

Fixed $K = 8$, 288 routed + 1 shared expert.
Pre-training loss, tested up to 3×10^{20} FLOPs.
Extrapolation and downstream transfer need validation.

