AI Progress Should Be Measured by Capability-Per-Resource, Not Scale Alone

A Framework for Gradient-Guided Resource Allocation in LLMs

David McCoy, Yulun Wu, Zachary Butzin-Dozier

UC Berkeley, Division of Biostatistics

NeurIPS 2025

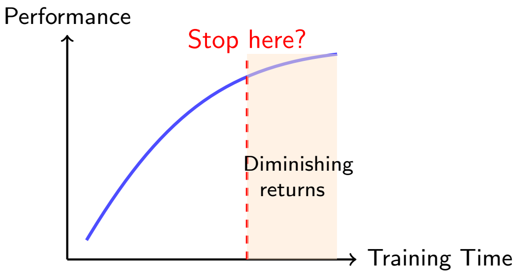
The Problem: Scaling Fundamentalism

Current AI Development:

- GPT-3: 552 tons CO_2
- LLaMA 65B: Final 15% of training $\rightarrow$ ¡0.01 improvement
- Growing resource inequality

Our Solution:

- Measure by $\Delta\Psi/\Delta\Gamma$
- Stop training when returns diminish
- Use gradient blueprints for adaptation



Why This Works: Heavy-Tailed Gradients

Key Observation:

Gradients follow power-law:

$$\|\nabla_{\theta_{(r)}}\| \approx C \cdot r^{-\alpha}, \quad \alpha \in (1,2)$$

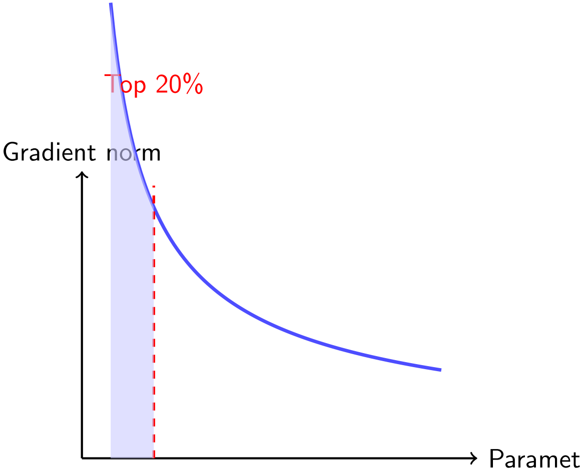
Implication:

- Top 10% params $\rightarrow$ 50% gradient mass
- Small fraction matters most

Theorem (Partial-Update Advantage)

Under power-law gradients, $\exists\, k^* \in (0,1)$ where:

$$\frac{\Delta_{k^*}(\Psi)}{\mathcal{C}(\Delta_{k^*})} > \frac{\Delta_{full}(\Psi)}{\mathcal{C}(\Delta_{full})}$$



Gradient Blueprints: Bridging Labs & Adapters

What are Blueprints?

Foundation labs publish metadata:

- Submodule gradient norms
- Recommended update fractions k^*
- Domain-specific weightings

Layer	**Grad**	k^*
attn.0	0.052	0.15
ffn.0	0.033	0.25
attn.1	0.045	0.13

Adapters: 60-80% memory reduction

Multiplicative Gains:

Parameter $\times$ Data Selection

20% params $\rightarrow$ 80% perf

30% data $\rightarrow$ 90% perf

Combined: 72% perf at 6% cost

Result: 12$\times$ efficiency gain

Real-World Impact & Key Results

For Foundation Labs:

- Save 10-20% compute
- Reduce carbon footprint
- Publish blueprints with models

For Smaller Labs:

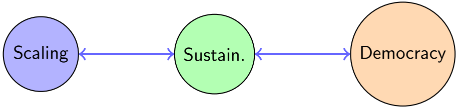
- Fine-tune on consumer GPUs
- $8\times$ memory reduction
- Democratized access

Example (Biomedical):

- Tune 13% of mid-layers
- Match full-model performance
- Enable research at scale

Theoretical Contributions:

1. **Prop 3.1**: Partial updates optimal
2. **Thm 3.2**: Gradient norms approximate influence
3. **Data selection**: Extends to training data
4. **Cross-influence**: Multiplicative gains



Capability-Per-Resource

Key Takeaways

1. **Challenge scaling fundamentalism**: Resource efficiency must be first-class
2. **Gradient blueprints**: New standard for model releases enabling efficient adaptation
3. **Theoretical foundations**: Heavy-tailed gradients justify selective updates as optimal
4. **Practical impact**:
    - Foundation labs: 10-20% compute savings, reduced emissions
    - Smaller labs: consumer GPU fine-tuning, democratized access
    - Combined approach: $10\times+$ efficiency gains

Thank You! See you at the poster!

david_mccoy@berkeley.edu