Distillation Scaling Laws

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram,
Etai Littwin, Russ Webb

ICML 2025 | Apple | https://arxiv.org/abs/2502.08606

Small, capable models offer substantial benefits

Lower thermal output
Enables more device deployment

Lower latency
Enables real-time interactions

Lower carbon footprint
Enables everything

Inference cost $\equiv$ FLOPs per token $\propto N$

Model size (parameters)

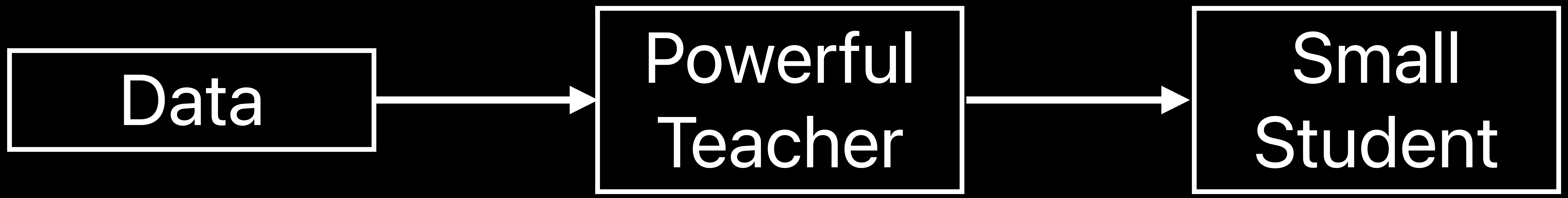
How can we make small and capable models?

Training *small* models *directly on data* is inefficient

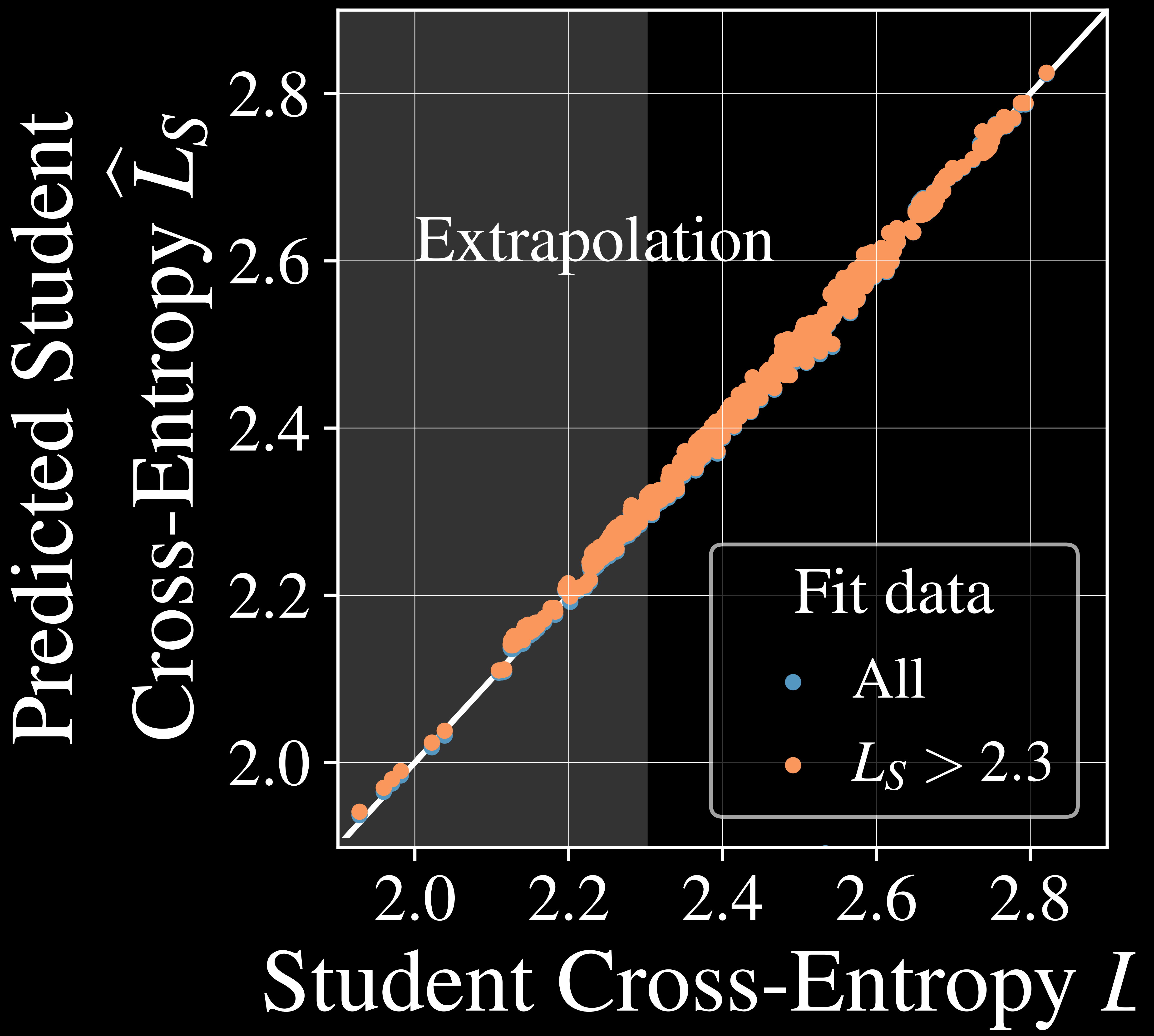
How do we maximize *data efficiency* for *small* models?

Distillation!

Distillation transfers knowledge from a teacher to a student



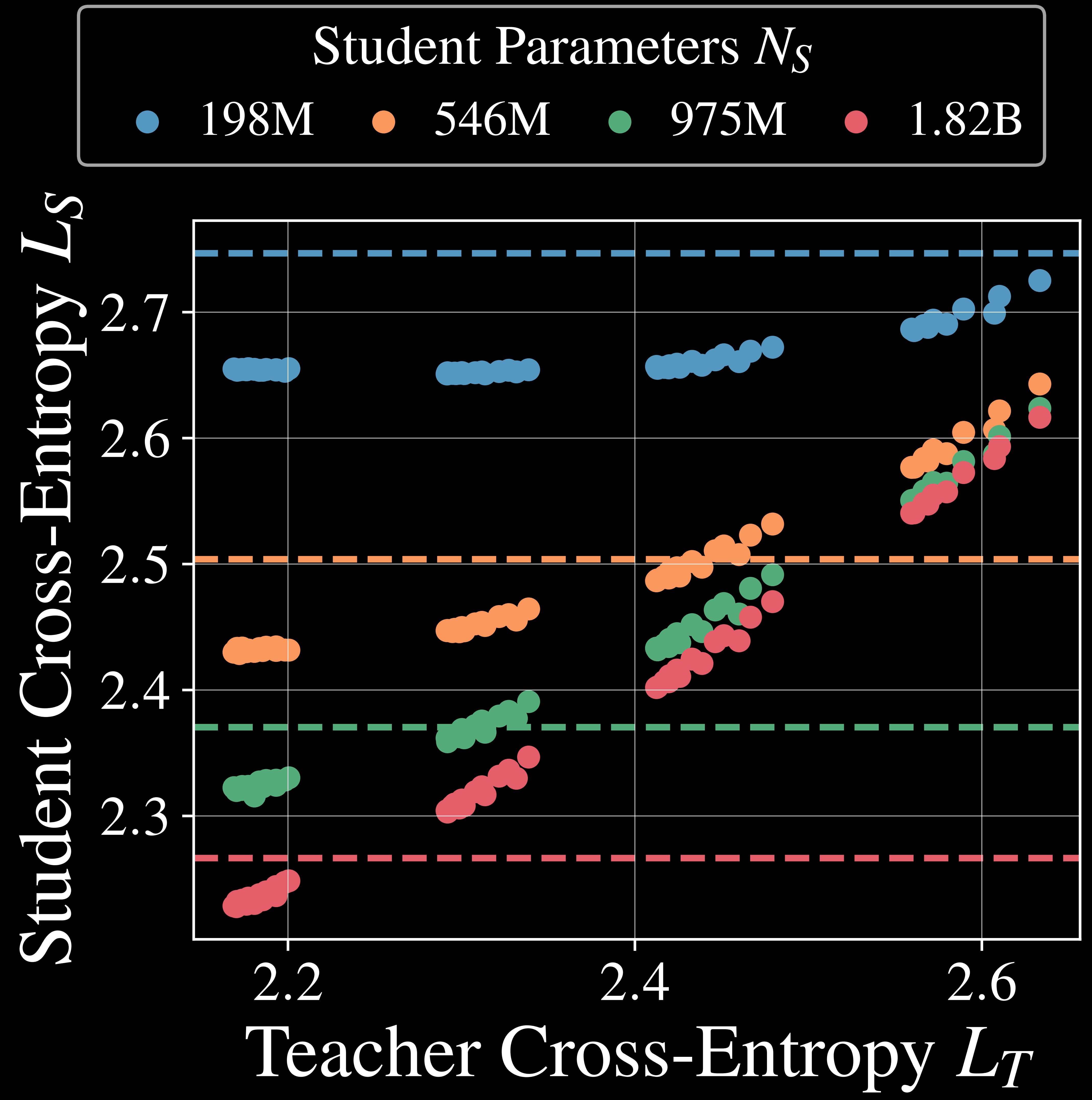
Our scaling law enables predictable distillation



Only teacher cross-entropy influences student performance

Distillation Scaling Law

$$L_S \approx L_T + \underbrace{f(L_T)}_{\text{Approx. Error}} \times \underbrace{L(N_S, D_S)}_{\text{Power Law}}$$



Our distillation scaling law enables compute optimal-distillation

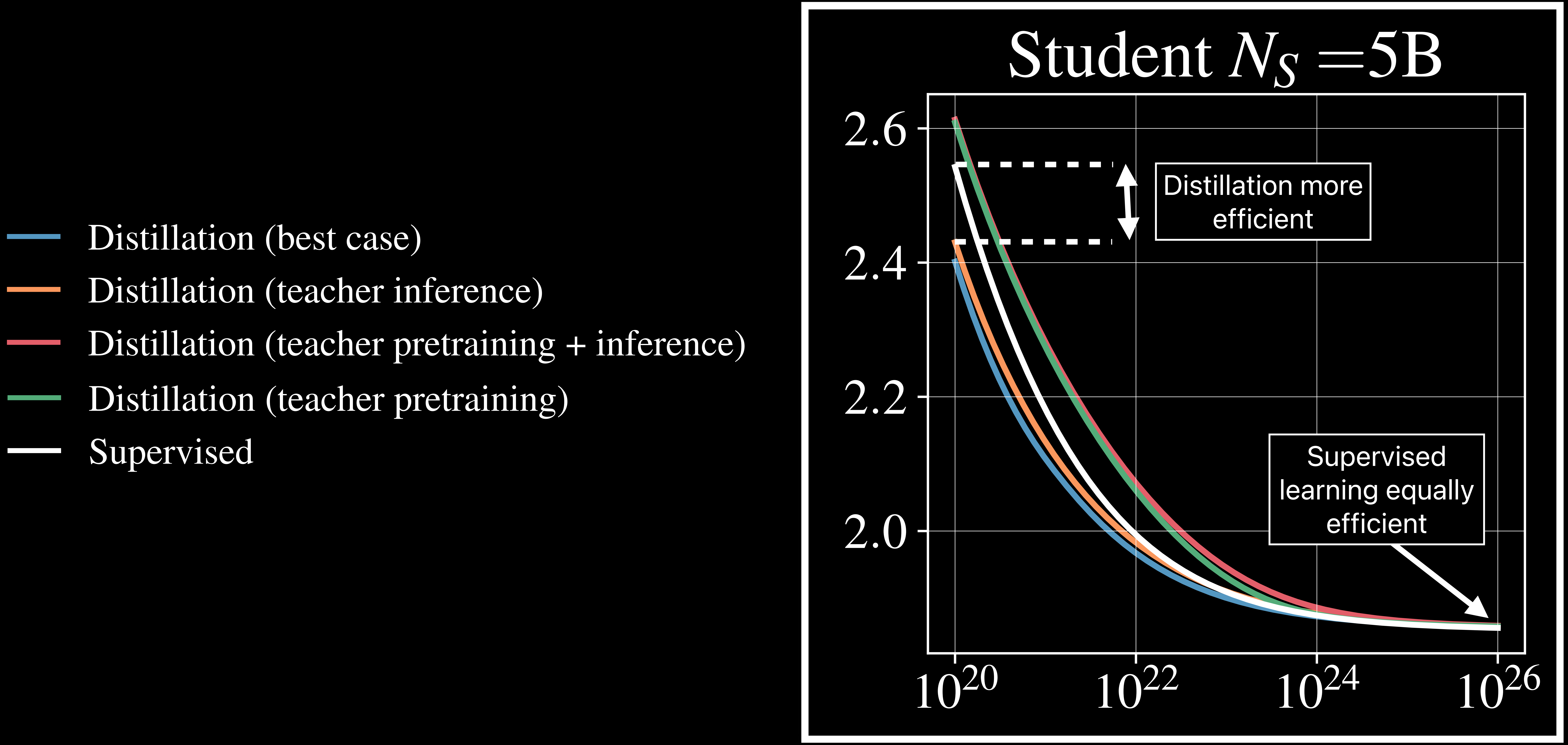
Compute-Optimal Distillation

The student (size + tokens), and teacher (size + tokens) producing the best student subject to a compute budget

We produced recipes that are *3x more data and compute efficient* that optimal supervised learning on data

Distillation is more efficient when discounting teacher training

This efficiency gap disappears at large compute and token budgets



Summary of Distillation Scaling Laws

1. We developed a distillation scaling law to predict student model performance

2. Using this law, we discovered training recipes that are up to 3x more efficient than optimal supervised learning

3. We also ran the largest distillation study to date, uncovering key guidelines to maximize student performance