Tencent 腾讯

Scaling Laws for Floating Point Quantization Training

ICML 2025

Xingwu Sun*, Shuaipeng Li*, Ruobing Xie, Weidong Han, Kan Wu, Zhen Yang, Yixing Li, An Wang, Shuai Li, Jinbao Xue, Yu Cheng, Yangyu Tao, Zhanhui Kang, Chengzhong Xu, Di Wang, Jie Jiang

Pipeline

1. Classical Scaling Law
2. Our Capybara Scaling Law
3. Findings

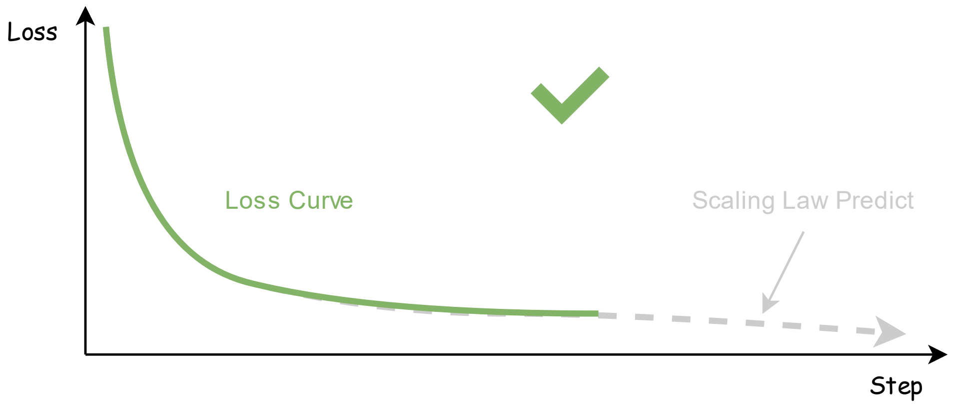
1 Classical Scaling Law

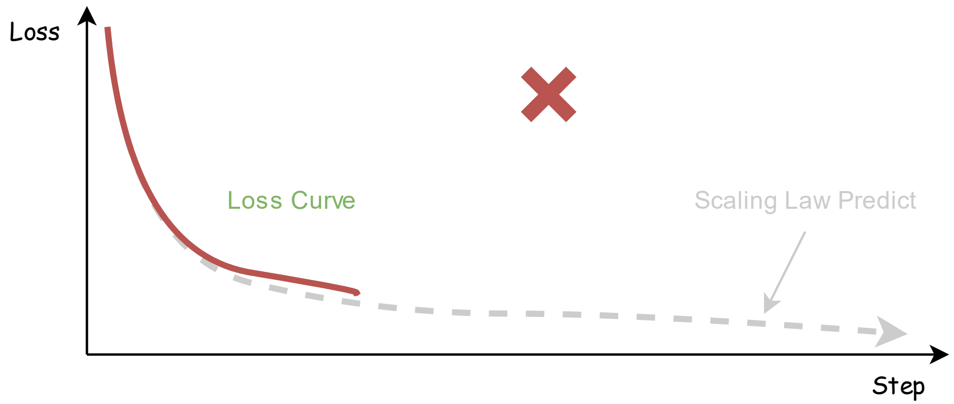
Scaling laws could guide the model training of LLMs

Typical scaling laws adopt model size (N) and data size (D) to predict the final loss of LLMs:

$$Loss(N, D)$$

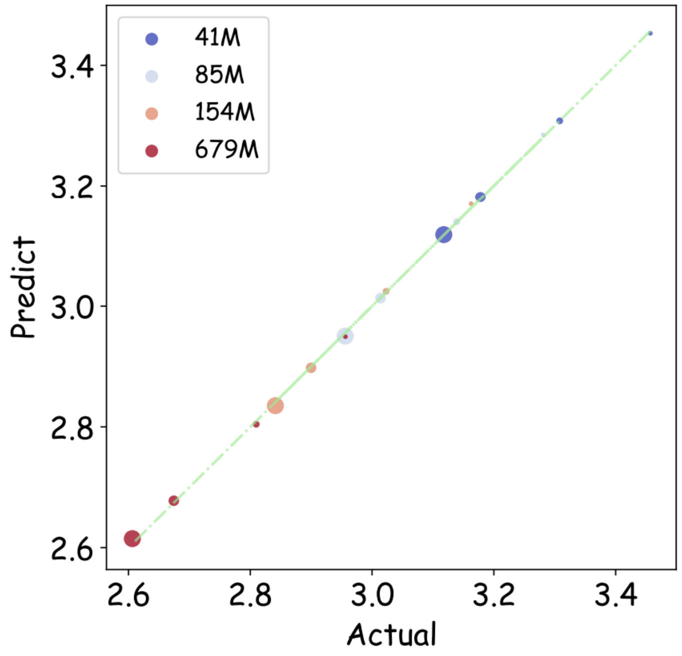
Through lots of experiments based on LLMs with smaller D and N, we could build the scaling law that could well predict the loss of larger models





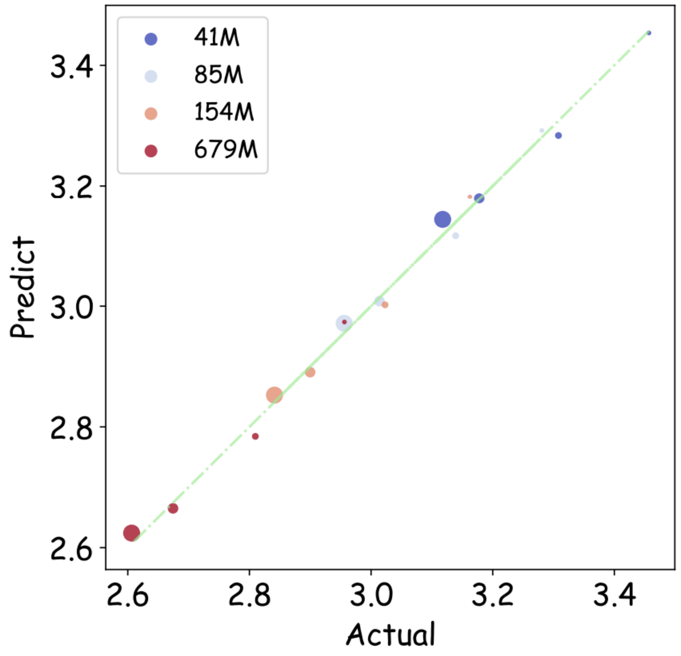
Chinchilla Scaling Law (Hoffmann er al., 2022)

$$L(N, D) = \frac{n}{N^{\alpha}} + \frac{d}{D^{\beta}} + \varepsilon$$



OpenAI Scaling Law (Kaplan er al., 2020)

$$L(N, D) = \left[\left(\frac{n}{N}\right)^{\frac{\alpha}{\beta}} + \frac{d}{D}\right]^{\beta} + \varepsilon$$



We select the Chinchilla scaling law as the base form of our scaling law for floating-point quantization training.

2 Our Capybara Scaling Law

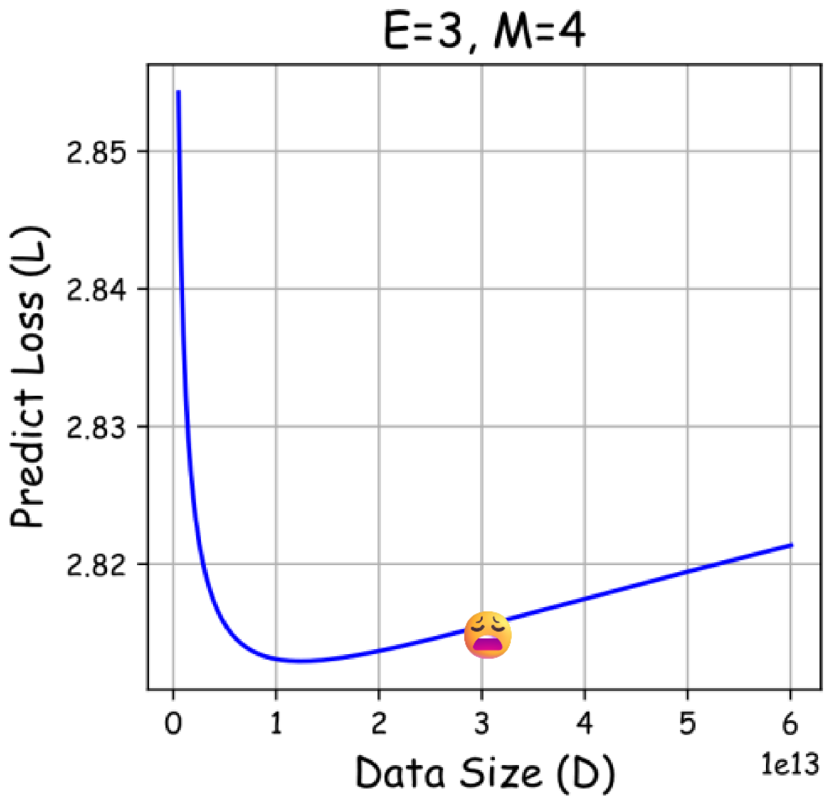
Motivation of Exploring Scaling Laws for FPQT

- Scaling laws of large language models (LLMs) could help developers effectively select superior parameter settings before experiments and accurately predict the model performance under different configurations.
- Training and serving with lower precision becomes a popular solution. Compared to integer quantization, floating-point (FP) quantization can better maintain LLMs' accuracy at extremely lower bit rates and thus is often equipped in low-precision LLMs.
- Currently, there is no systematic exploration on the scaling laws for floating–point quantization, which is widely used in practical LLM systems.

The Capybara Scaling Law

$$\mathrm{L}(N, D, E, M, B) = \frac{n}{N^{\alpha}} + \frac{d}{D^{\beta}} + \varepsilon + \frac{D^{\alpha}}{N^{\beta}} \frac{log_2 B}{\gamma(E+0.5)^{\delta}(M+0.5)^{\upsilon}}$$

Going beyond certain limit is as bad as falling short.



Under constrained resources and space, increasing the number of capybaras can significantly reduce their survival rate and quantity once a certain density threshold is surpassed.

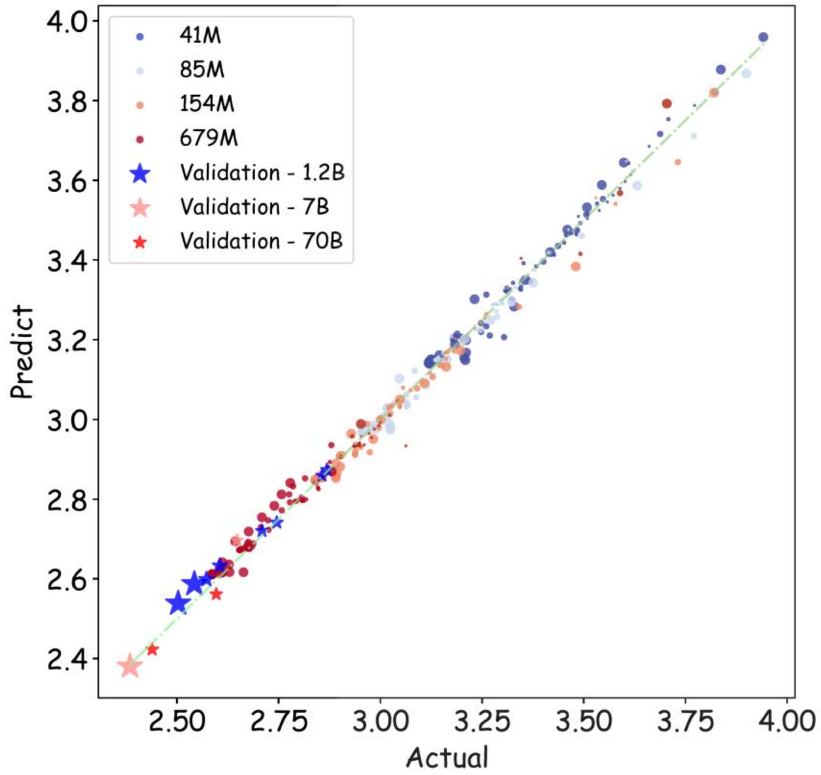
We observe a similar phenomenon in our scaling law: with a fixed model size, expanding the data size does not consistently yield improvements when the "knowledge density" becomes too high under the pressure of low-precision training --- just like Capybaras.

The Capybara Scaling Law

$$L(N, D, E, M, B) = \frac{n}{N^{\alpha}} + \frac{d}{D^{\beta}} + \epsilon$$

$$+ \frac{D^{\alpha}}{N^{\beta}} \frac{\log_2 B}{\gamma (E+0.5)^{\delta} (M+0.5)^{\nu}}$$

Going beyond certain limit is as bad as falling short.

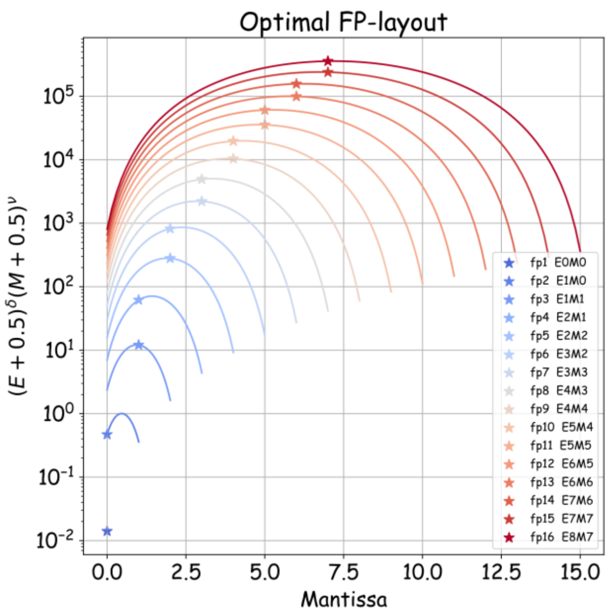


We have trained more than 360+ LLMs with different N, D, E, M, B settings to learn our scaling law.

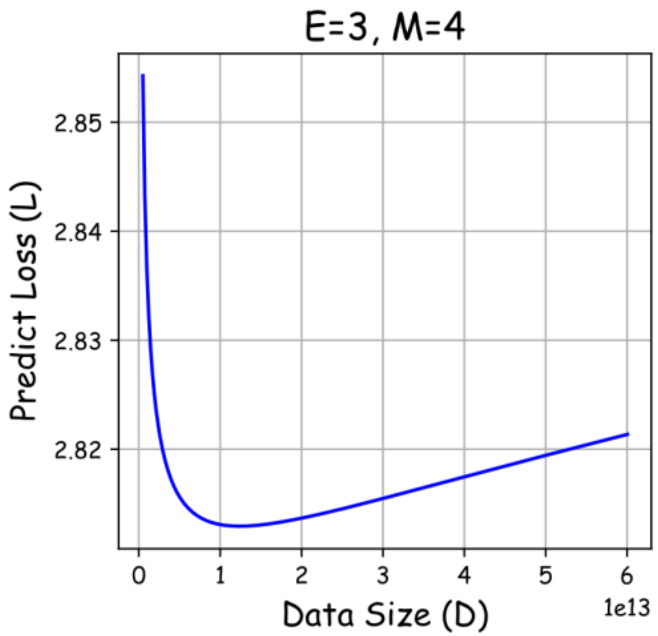
The proposed scaling law functions well when predicting larger LLMs (1.2B, 7B, and 70B) under floating-point quantization training.

3 Findings

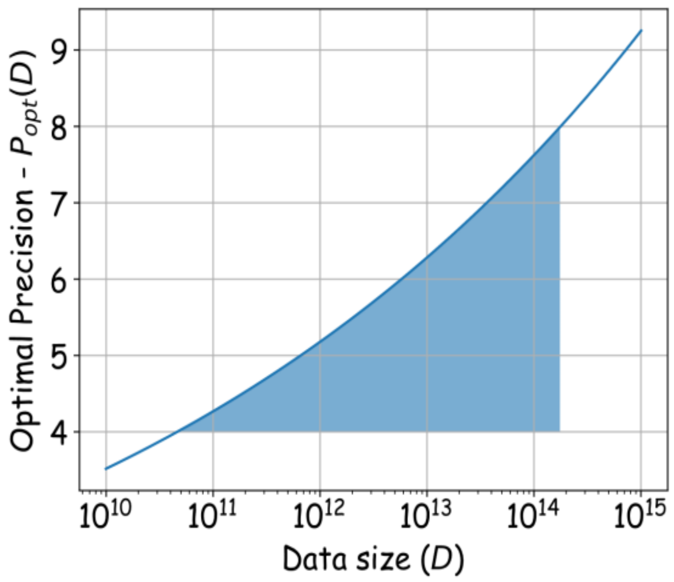
Implications found by our scaling law



We conduct the optimal float layout analysis. The optimal float layouts of FP4, FP8, and FP16 are E2M1, E4M3, and E8M7



We find the critical data size for optimal performance. Sometimes increasing the data size will result in performance decline under FPQT



We discuss the compute-optimality with fixed configurations. Luckily, under practical compute budget, the optimal cost-performance ratio precision lies between 4 and 8 bits

Thanks

Tencent腾讯