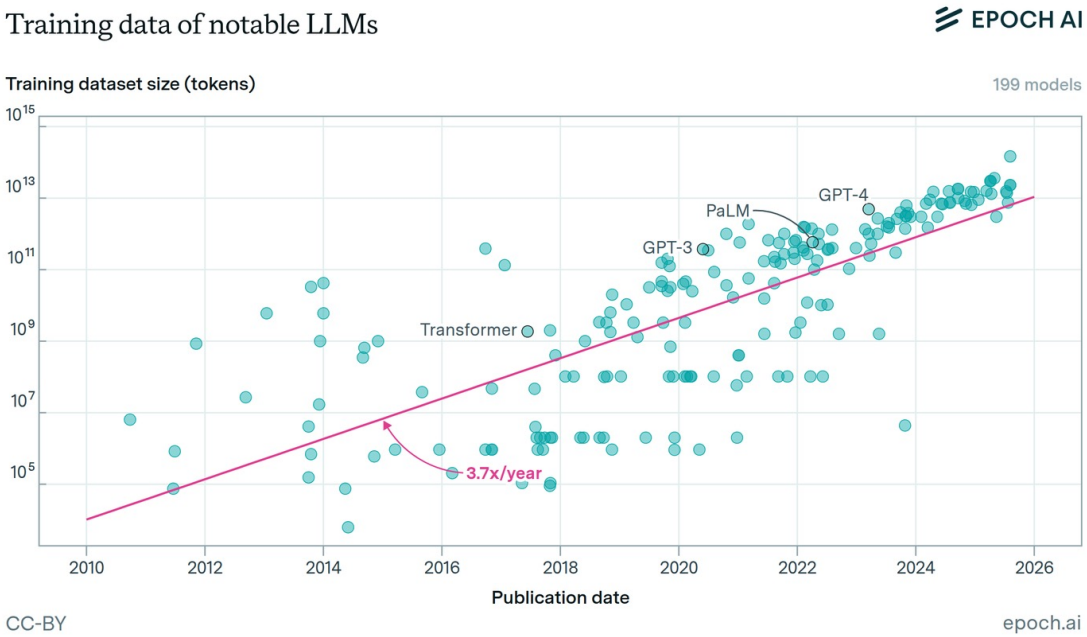
Larger Datasets Can Be Repeated More:
A Theoretical Analysis of Multi-Epoch Scaling in Linear Regression

Haodong Wen*

Joint work with Tingkai Yan*, Binghui Li*, Kairong Luo, Wenguang Chen, Kaifeng Lyu

We are running out of data

The size of datasets used to train LLMs doubles approximately every six months



Massive amount of data for pre-training!

- **GPT 2:** < 10 billion tokens
- **GPT 3:** 370 billion tokens
- **Llama 2:** 2 trillion tokens
- **Qwen 3:** 36 trillion tokens
- ...

Epoch.ai Report: Public data may be exhausted as early as 2028

One-Pass training → Multi-Epoch Training

One-Pass & Multi-Epoch Training at Scale

Question: How should we allocate the training budget?

Common Practice: One-pass training

Choose Model size (M); dataset size (N)

Data-constrained Regime: Multi-ecpoh training

Choose Model size (M); dataset size (N); epochs (K)

Chinchilla Scaling Law for (one-pass)

$$L(M, N) = L_0 + A\,M^{-\alpha} + B\,N^{-\beta}$$

Pretraining loss

Power law in N

Compute-optimal training: $M \sim N^a$

Muennighoff et al., 2023 (multi epoch):

$$L(M', N') = L_0 + A\,M'^{-\alpha} + B\,N'^{-\beta}$$

$$N' = (1 + R^*(1 - e^{\frac{K-1}{R^*}})) \cdot N$$

T: Number of training steps (or tokens)

The pre-training loss follows an power law wrt the training dataset size N

The pre-training loss follows an underlying power law wrt the effective dataset size N'

Heuristic Derivation of the Multi-Epoch Scaling Law

Effective dataset size: $N' = (1 + R^*(1 - e^{-(K-1)/R^*}) \cdot N$

Key Intuition: for every repeat of the data, the training process can extract $1 - \delta$ fraction of information contained in every data point.

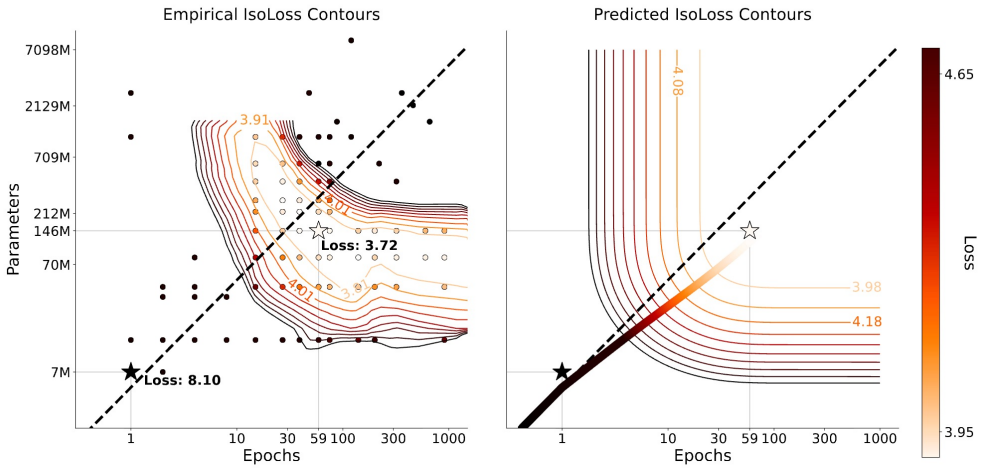
Effective dataset size: $N' = \underbrace{(1 + R^*(1 - e^{-\frac{K-1}{R^*}})}_{\text{Independent of N !}} \cdot N$

→ **Nice Property:** Effective reuse rate: $N'/N = (1 + R^*(1 - e^{-\frac{K-1}{R^*}})$

Heuristic Derivation of the Multi-Epoch Scaling Law

The decay ratio Independent of N!

Effective data reuse rate: $N'/N = (1 + R^*(1 - e^{-\frac{K-1}{R^*}})$



Conceptual concern: The heuristic that N'/N has an a universal decay independent of N may be too strong

how multi-epoch training affects the data scaling laws, theoretically.

Let's start with Linear regression!

- **Theorists love the Linear model, simple but as a mindset for more complex problems**
- A common testbed for theoretical explanations of **scaling laws (Lin et al., 2024)**

Our Work: Larger Datasets Can Be Repeated More

Our work: Larger datasets can be repeated more!

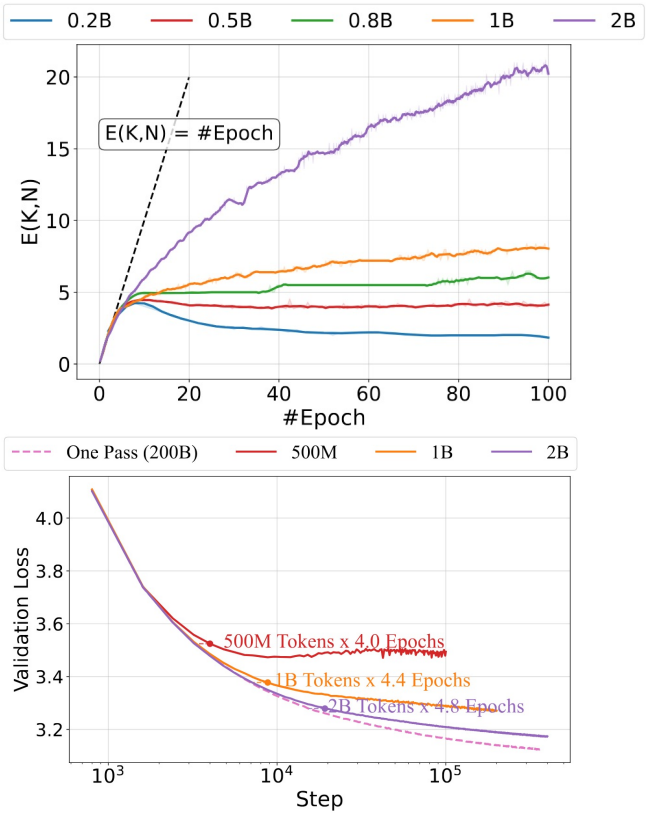
Definition 3.1 (Effective Reuse Rate). *Given K-epoch SGD trained with N fresh data samples, the effective reuse ratio is defined as:*

$$E(K,N) := \frac{1}{N}\min\{N' \geq 0 : \bar{\mathcal{R}}^*(1,N') \leq \bar{\mathcal{R}}^*(K,N)\}.$$

R(K,N): excess risk (loss - irreducible constant) with K-epochs and N-data

When K = o(log N): $E(K,N) = K\,(1+o(1))$
When K = ω(log N): $E(K,N) \propto \log N\,(1+o(1))$

Key Insight: Effective reuse rate should depend on dataset size N



Theoretical Setup: Linear Regression

Task:

- Linear regression: $y = \langle w^*, x \rangle + \xi$
- $w^* \in \mathbb{R}^d$: ground truth weight
- ξ: label noise with zero mean, $\mathbb{E}[\xi^2] = \sigma^2$
- Strongly convex: $\lambda_{min}(\mathbb{E}[xx^T]) \geq constant > 0$

Model:

- Linear model: $f(x, w) = \langle x, w \rangle$
- MSE loss: $\ell(w, x, y) = \frac{1}{2}(f(x, w) - y)^2$, $L(w) = \mathbb{E}_{(x,y)}[\ell(w, x, y)]$
- Excess risk: $R(w) = L(w) - \frac{1}{2}\sigma^2$

Theoretical Setup: Linear Regression

Training:

- Multi-epoch SGD with random shuffling, with constant LR η:

$$w_{t+1} = w_t - \eta \nabla \ell_t\left(\mathrm{w_t}, \mathrm{x_{j(t)}}, \mathrm{y_{j(t)}}\right) = \left(I - \eta \mathrm{x_{j(t)}} \mathrm{x_{j(t)}}^T\right) w_t + \xi_{j(t)} \mathrm{x_{j(t)}}$$

Training metrics:

- Expected excess risk $\bar{R}(K, N) = \mathbb{E}[R(w)]$ (expectation over training process)
- Optimal expected excess risk: $R^*(K, N) = \min_\eta \bar{R}(K, N)$.

Recap **the effective reuse rate**:

$$E(K, N) \coloneqq \frac{1}{N} \min\{N' > 0 : R^*(1, N') \leq R^*(K, N)\}$$

Our results

Theorem 1 (multi-epoch scaling law).

$$R^*(K,N) = \begin{cases} \frac{C_1 \log KN}{KN} \cdot \big(1 + o(1)\big) & \text{For } K = o(\log N) \\ C_2 \frac{1}{N} \cdot (1 + o(1)) & \text{For } K = \omega(\log N) \end{cases}$$

Theorem 2 (effective reuse rate).

$$E(K,N) = \begin{cases} K \cdot \big(1 + o(1)\big) & \text{For } K = o(\log N) \\ C \log N \cdot (1 + o(1)) & \text{For } K = \omega(\log N) \end{cases}$$

Phase transition happens when K is large

There is two phases for the multi-epoch training

Effective reuse regime: Reusing data works approximately like fresh data

Limited reuse regime: additional epochs yield only log gains,

Larger datasets can be repeated more

Why Can Larger Datasets Can Be Repeated More

Consider a fixed number of epochs K

Regime 1.

When the dataset size N is small → **overfitting happens** → $E(K,N) < K$

Regime 2.

When the dataset size N is large → **Random matrix product concentration**

$$A := \prod_{i=1}^{N}\left(I - \eta z_i z_i^T\right) \to \mathbb{E}A$$

→ $E(K,N) \approx K$

Now vary number of epochs from 1 to K , and fix N as we do in training:

Regime 2 → Regime 1: $E(K,N) \approx K \to E(K,N) = \Theta(\log N) < K$

Beyond Strongly Convex

In strongly convex case:

The effective reuse rate saturates as $E(K,N) \to \Theta(\log N)$

- Q1: Does the log factor universally hold?
- Q2: If not, how would $E(K,N)$ behave in other problems

Zipf-law data distribution

- Natural data distributions often exhibit power law structures
- **Problem setup:**
  - d one-hot data points, the i-th data point with $\mathrm{x}_i = \mu_i e_i$ (with norm μ_i)
  - The i-th data point is sampled with probability p_i
  - Zipf-law: $\mathbb{E}[xx^T]_{ii} \sim i^{-a}$.

Effective Reuse Rate for Zipf-distributed data

Power spectrum results

Theorem 1.

$$E(K,N) = \begin{cases} K \cdot \left(1 + o(1)\right) & \text{For } K = o(N^{\frac{b}{a-b}}) \\ \Theta(N^{\frac{b}{a-b}}) & \text{For } K = \omega(N^{\frac{b}{a-b}}) \end{cases}$$

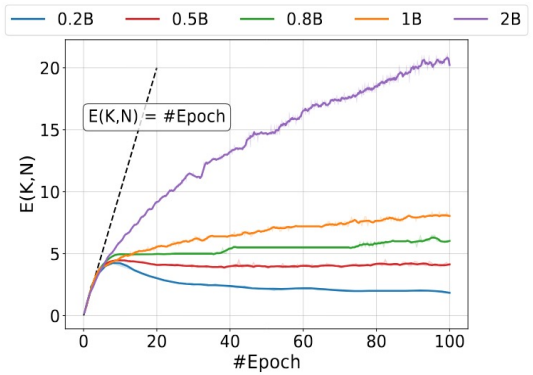
Log-Power spectrum results

Theorem 2.

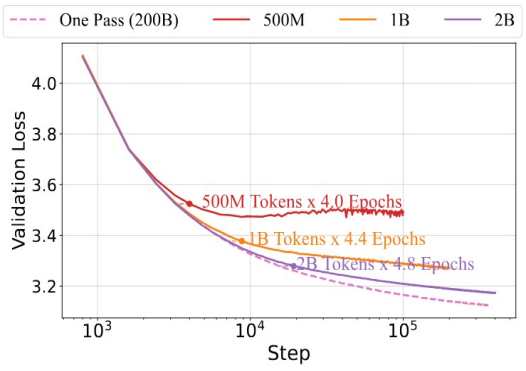
$$E(K,N) = \begin{cases} K \cdot \left(1 + o(1)\right) & \text{For } K = o((\log N)^b) \\ \Theta((\log N)^b) & \text{For } K = \omega((\log N)^b) \end{cases}$$

Insights: Larger dataset can still be repeated more; the **phase transition point differs** with the distribution of data (e.g., from $\log N$ to $N^{\frac{b}{a-b}}$ or $(\log N)^b$)

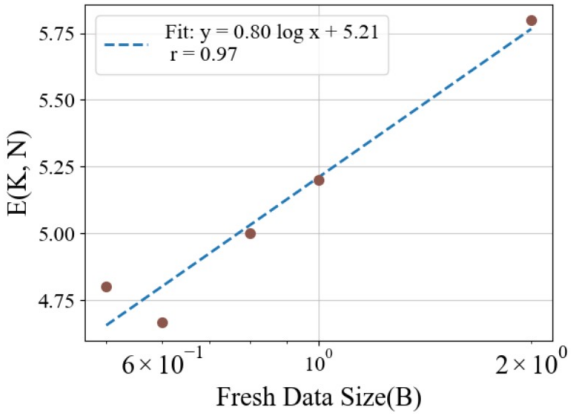
Recap: Large Language Model Experiments



(a) The effective reuse rate $E(K, N)$ as a function of the epoch number K.



(b) Training loss as a function of training steps for differe fresh data sizes.



There is two phases for the multi-epoch training in LLM training

Effective reuse regime: Reusing data works approximately like fresh data!

Limited reuse regime: As #epochs increases, E(K,N) saturates to some dataset-dependent point, and the benefits of reusing data vanish.

A log-law in our experiments! $K(N) = A \cdot \log N + B$

Summary

- A Theoretical Analysis of Multi-Epoch Scaling in Linear Regression
  - **Insight for multi-epoch training:** Larger dataset can be repeated more.
  - reuse your data up to log N times (not too many times)

- **Limitation and Future works:**

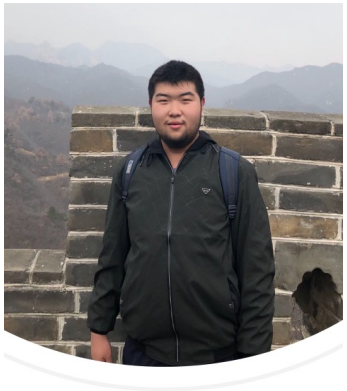
1. Constant learning rate
    1. Can we extend the analysis to training with learning rate decay?
2. Linear model
    - Can we extend the analysis to a more realistic regime with feature learning?
3. How does the effective reusing rate change with model size?

Extension: Can we extend our method to order data processing paradigm?

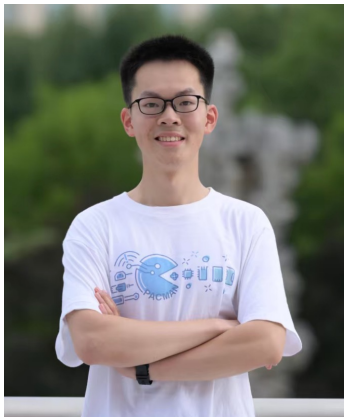
- curriculum learning; data schedule; synthetic data...

Thanks for your attention!

Tingkai Yan*



Binghui Li*



Kairong Luo

Wenguang Chen

Kaifeng Lyu