Data Selection via Optimal Control for Language Models

Yuxian Gu[1,2], Li Dong[2], Hongning Wang[1], Yaru Hao[2],
Qingxiu Dong[3], Minlie Huang[1], Furu Wei[2]

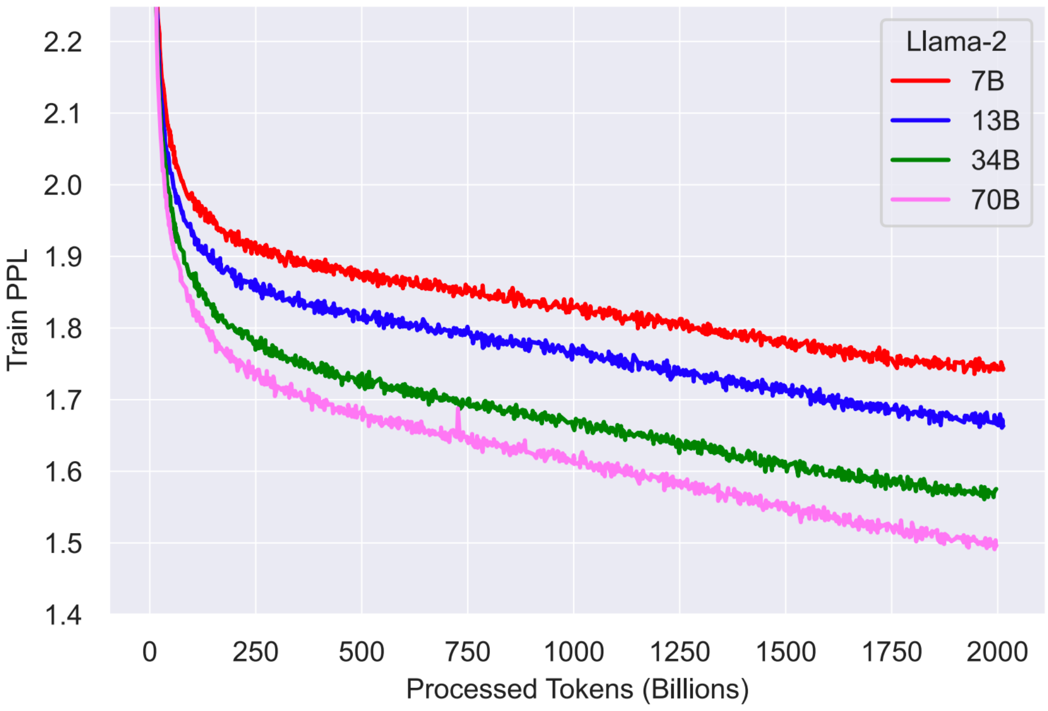
[1]The CoAI Group, Tsinghua University

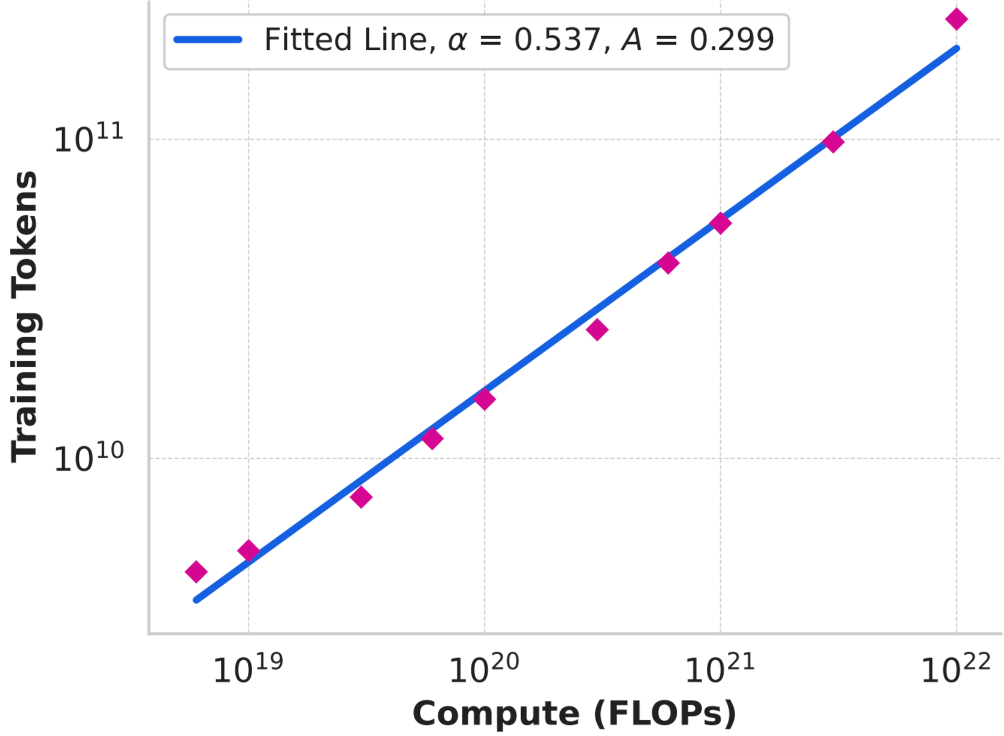
[2]Microsoft Research, [3]Peking Univeristy

清華大學
Tsinghua University

Data challenges for pre-training LMs

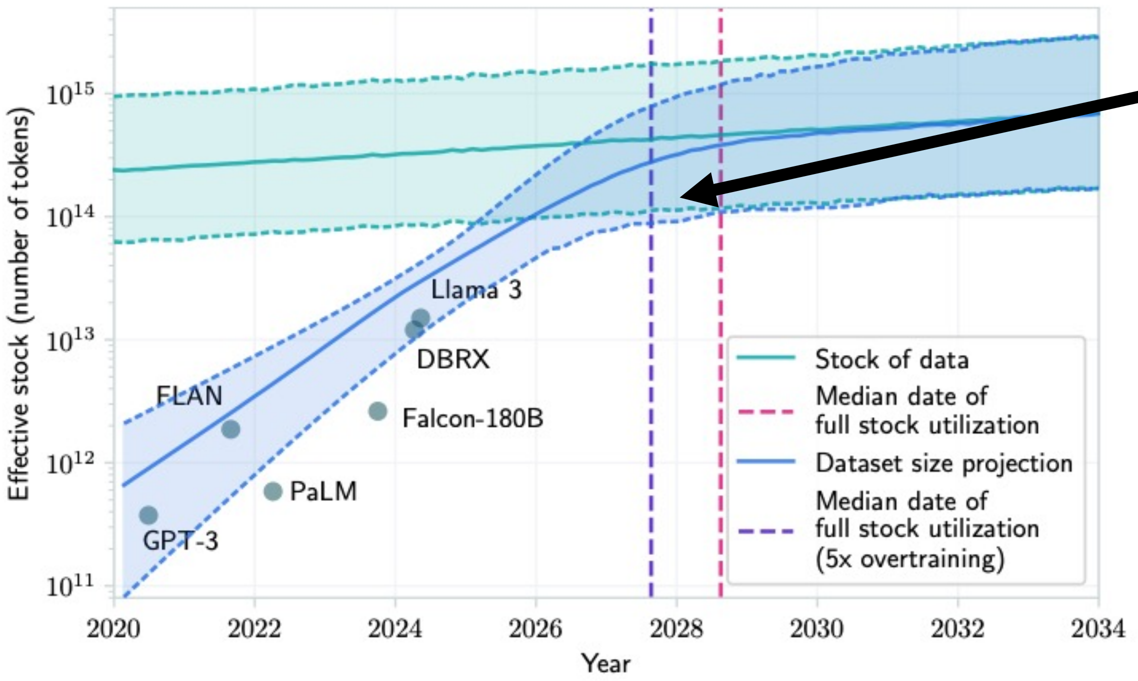
- Large amount of data makes pre-training quite inefficient.
- High-quality pre-training data is running out.
- Data selection/cleaning is a heuristic-based tricky task.





Motivation

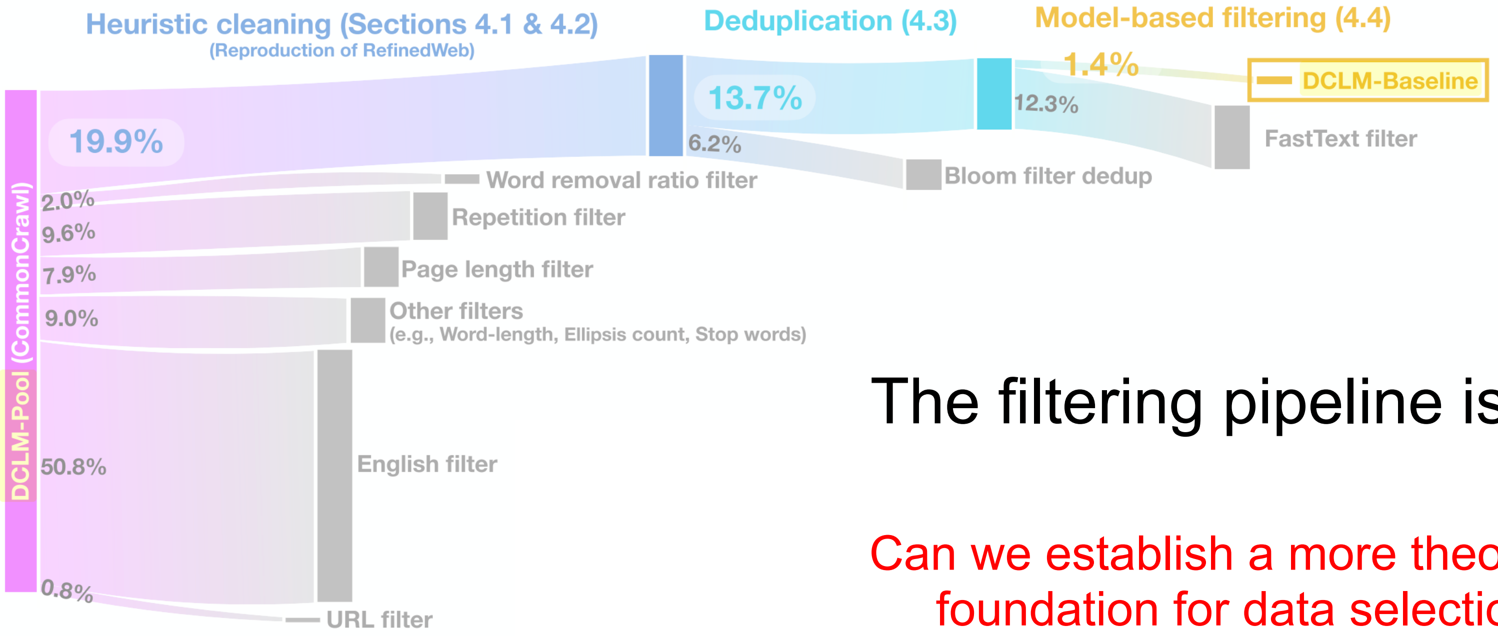
- Large amount of data makes pre-training quite inefficient.
- High-quality pre-training data is running out.
- Data selection/filtering is a heuristic-based tricky task.



Models consume faster than humans produce.

Data challenges for pre-training LMs

- Large amount of data makes pre-training quite inefficient.
- High-quality pre-training data is running out.
- Data selection/filtering is a heuristic-based tricky task.



The filtering pipeline is complex!

Can we establish a more theoretical foundation for data selection?

Data Selection based on Optimal Control

- Data challenges for pre-training LMs
  - Large amount of data makes pre-training quite inefficient.
  - High-quality pre-training data is running out.
  - Data selection/cleaning is a heuristic-based tricky task.
- Our data selection method **PDS** addresses the above problems

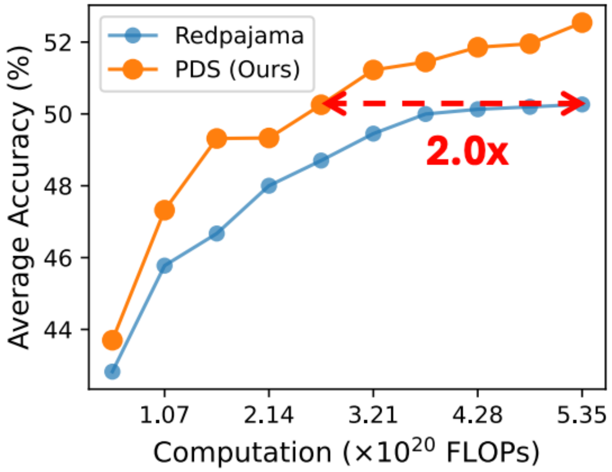
Theorem 2.1 (PMP Conditions for Data Selection)

$$\boldsymbol{\theta}^*_{t+1} = \boldsymbol{\theta}^*_t - \eta \nabla L(\boldsymbol{\theta}^*_t, \boldsymbol{\gamma}^*), \quad \boldsymbol{\theta}^*_0 = \boldsymbol{\theta}_0,$$

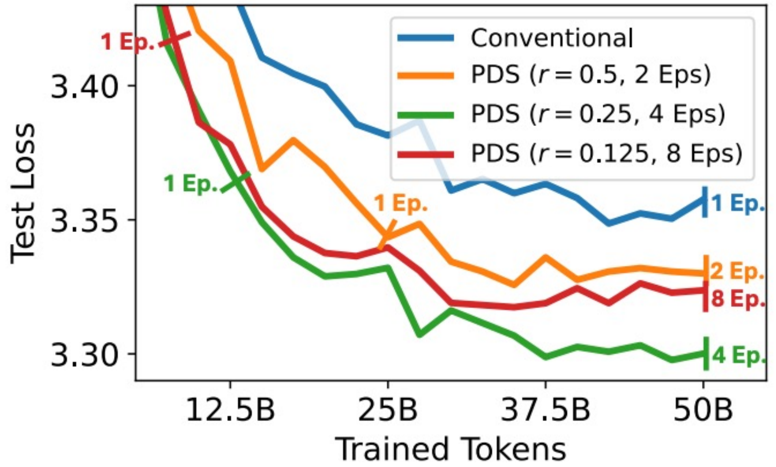
$$\boldsymbol{\lambda}^*_t = \boldsymbol{\lambda}^*_{t+1} + \nabla J(\boldsymbol{\theta}^*_t) - \eta \nabla^2 L(\boldsymbol{\theta}^*_t, \boldsymbol{\gamma}^*) \boldsymbol{\lambda}^*_{t+1},$$

$$\boldsymbol{\gamma}^* = \arg\max_{\boldsymbol{\gamma}} \sum_{n=1}^{|\mathcal{D}|} \gamma_n \left[\sum_{t=0}^{T-1} {\boldsymbol{\lambda}^*_{t+1}}^\top \nabla l(x_n, \boldsymbol{\theta}^*_t) \right]$$

Good theoretical guarantees



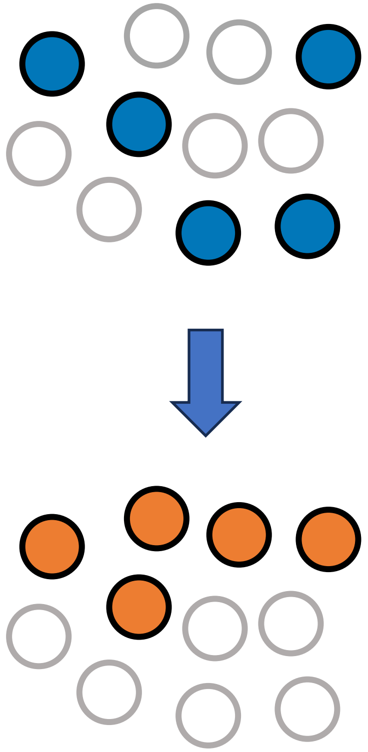
2x acceleration on selected data



Perf. improvement on limited data

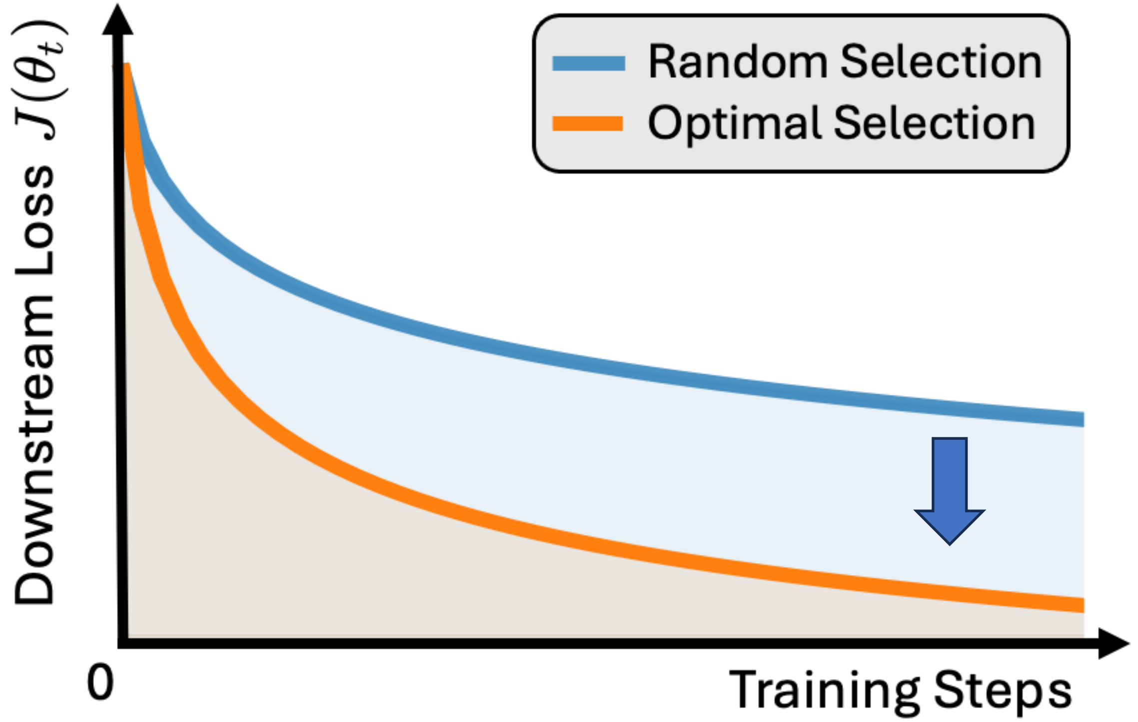
Data Selection as a Control Problem

- The Data selection strategy is the control signal to optimize



$$\gamma = [1,0,1,\cdots 0] \in R^N$$

$$\gamma^* = [1,1,1 \cdots 0] \in R^N$$



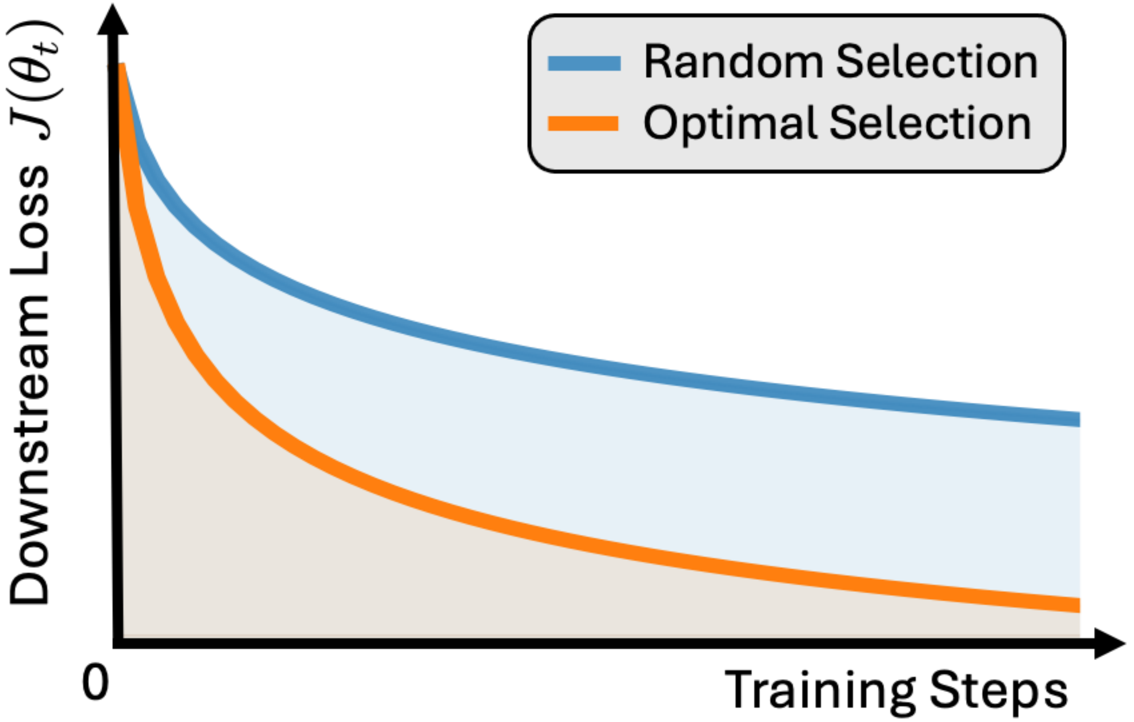
Formulation

- The Data selection strategy is the control signal to optimize

$$\min_{\gamma} \sum_{t=1}^{T} J(\boldsymbol{\theta}_t),$$

$$\text{s.t. } \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla L(\boldsymbol{\theta}_t, \boldsymbol{\gamma})$$

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{n=1}^{|\mathcal{D}|} \gamma_n l(x_n, \boldsymbol{\theta})$$



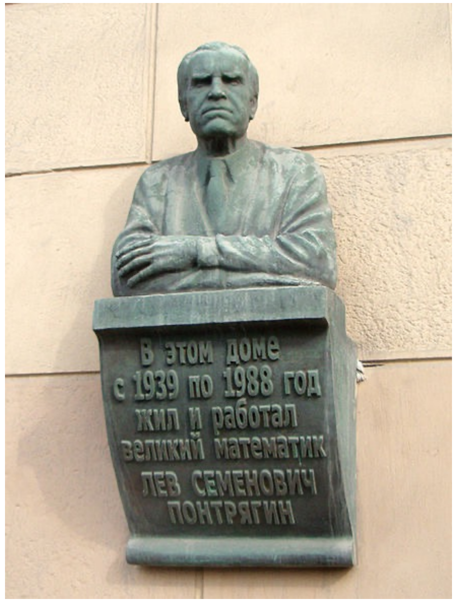
Solving the Problem

- Pontryagin's Maximum Principle (PMP)
  - Gives a necessary condition for the optimality of the problem

$$\min_{\boldsymbol{\gamma}} \sum_{t=1}^{T} J(\boldsymbol{\theta}_t),$$

$$\text{s.t.}\ \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla L(\boldsymbol{\theta}_t, \boldsymbol{\gamma})$$

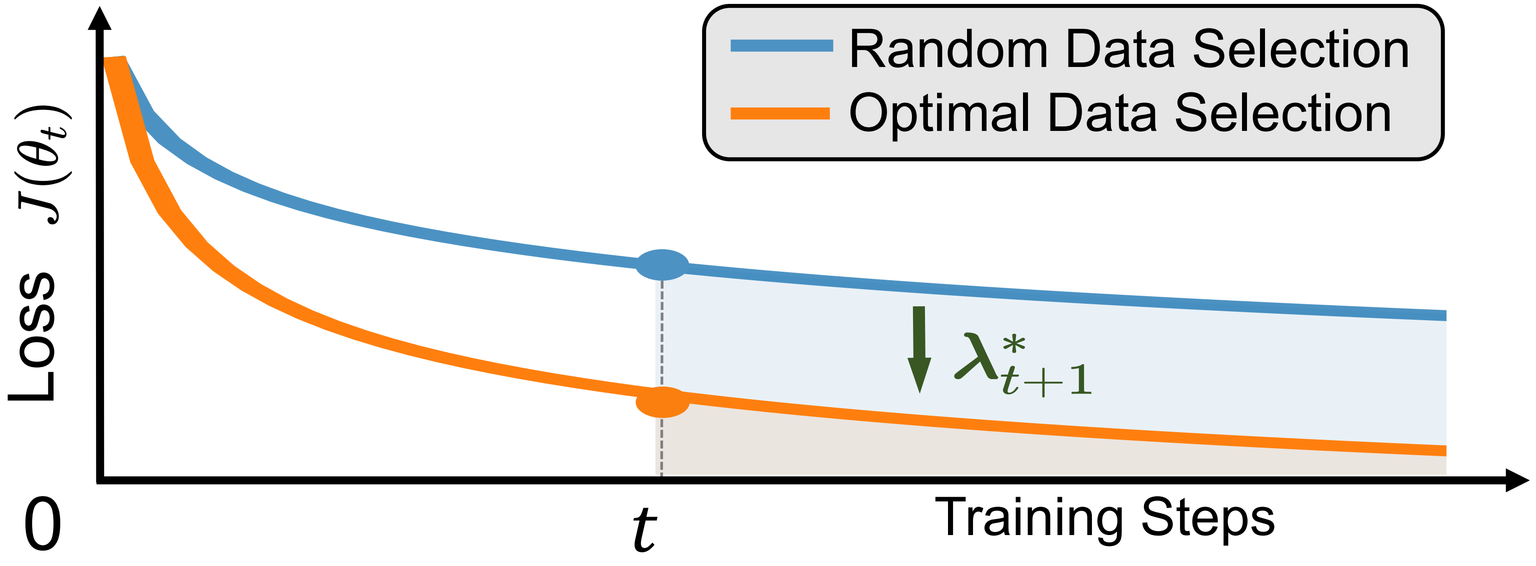
$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{n=1}^{|\mathcal{D}|} \gamma_n l(x_n, \boldsymbol{\theta})$$



Lev Pontryagin, 1908 - 1988

PDS: PMP-Based Data Selection

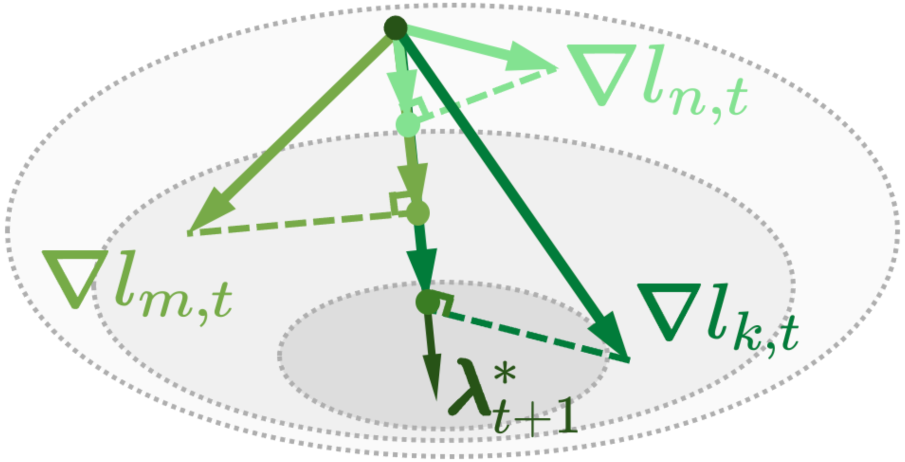
- PMP gives the ideal gradient direction for optimal data selection



$$\boldsymbol{\lambda}^*_t = \boldsymbol{\lambda}^*_{t+1} + \nabla J(\boldsymbol{\theta}^*_t) - \eta \nabla^2 L(\boldsymbol{\theta}^*_t, \gamma^*) \boldsymbol{\lambda}^*_{t+1}$$

PDS: PMP-Based Data Selection

- Select data whose gradient aligns with the optimal direction



$$\sum_t {\boldsymbol{\lambda}^*_{t+1}}^\top \nabla l_{n,t} < \sum_t {\boldsymbol{\lambda}^*_{t+1}}^\top \nabla l_{m,t} < \sum_t {\boldsymbol{\lambda}^*_{t+1}}^\top \nabla l_{k,t}$$

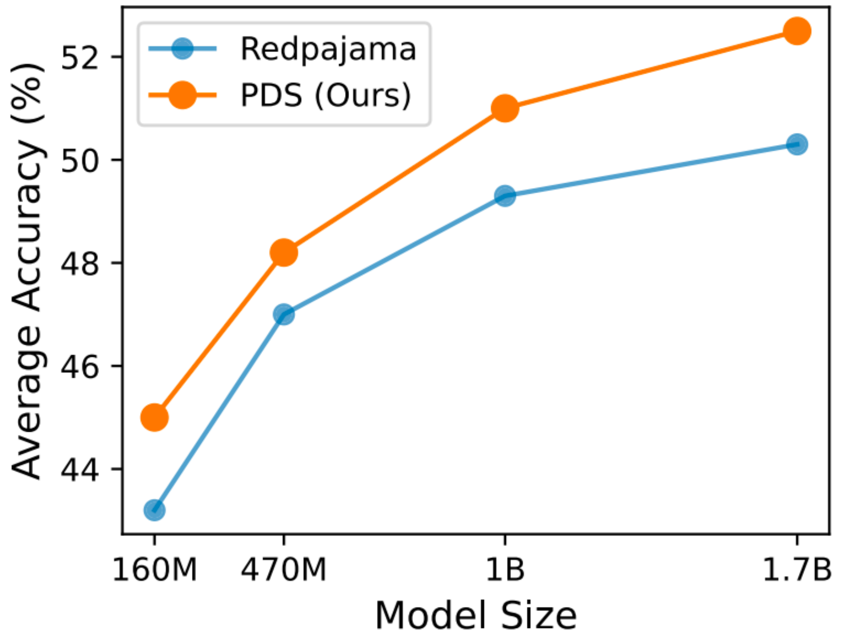
Select 30%: $\gamma_n = 0$ $\gamma_m = 0$ $\gamma_k = 1$

$$\boldsymbol{\gamma}^* = \arg\max_{\boldsymbol{\gamma}} \sum_{n=1}^{|\mathcal{D}|} \gamma_n \left[\sum_{t=0}^{T-1} {\boldsymbol{\lambda}^*_{t+1}}^\top \nabla l(x_n, \boldsymbol{\theta}^*_t) \right]$$

Performance Improvement

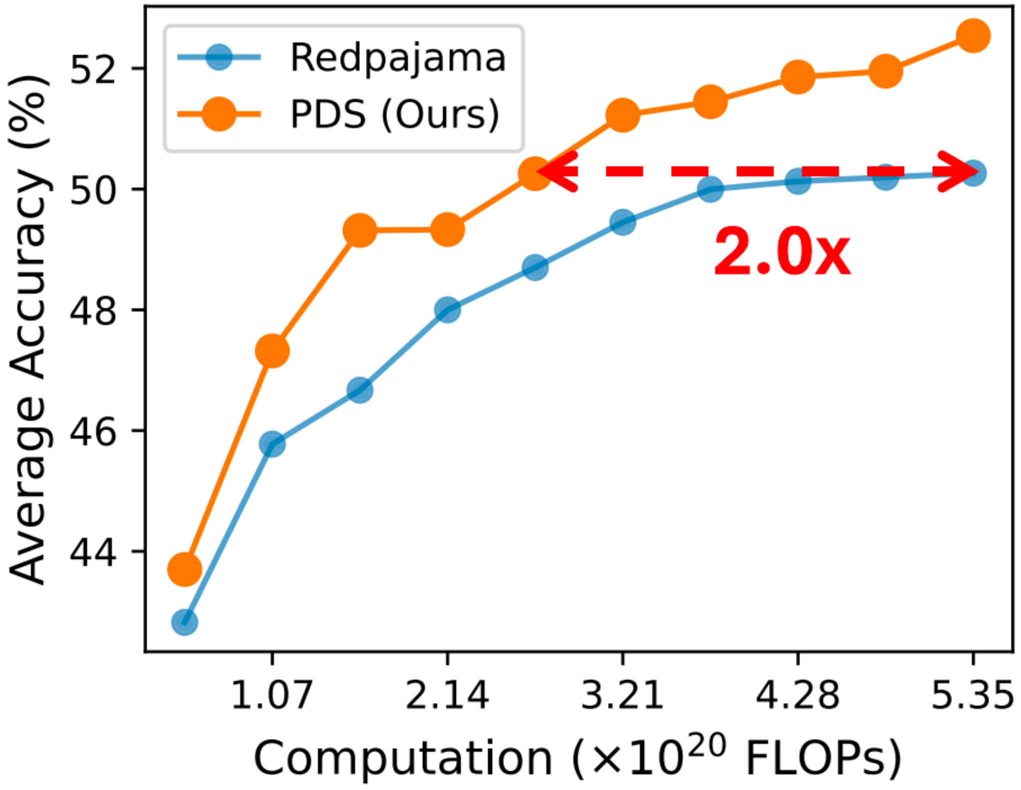
- Select 50B-token corpus from 125B-token corpus.
- Match the total training steps with the baselines (training computation)

	HS	LAMB	Wino.	OBQA	ARC-e	ARC-c	PIQA	SciQ	BoolQ	Avg.
					Model Size = 470M					
Conventional	36.7	41.4	52.4	**30.4**	44.8	25.2	61.0	70.6	60.4	47.0
RHO-Loss	36.6	42.4	53.0	29.4	43.7	25.2	60.4	72.8	59.8	47.0
DSIR	36.4	42.6	51.7	29.8	46.0	24.7	61.0	72.0	55.8	46.7
IF-Score	36.6	41.8	**53.4**	29.6	44.7	25.1	60.8	68.8	58.7	46.6
PDS	**37.9**	**44.6**	52.3	29.8	**46.5**	**25.8**	**61.8**	**73.8**	**61.4**	**48.2**
					Model Size = 1B					
Conventional	39.9	47.6	52.4	30.6	49.3	26.4	63.1	73.7	60.9	49.3
RHO-Loss	39.8	47.0	53.0	30.8	48.0	26.4	62.9	71.1	**61.0**	48.9
DSIR	40.8	47.8	53.0	31.2	49.8	26.8	62.7	76.6	58.0	49.6
IF-Score	39.4	47.0	52.6	28.6	49.4	26.4	63.5	74.0	60.5	49.0
PDS	**42.1**	**48.8**	**54.0**	**33.4**	**51.3**	**28.0**	**64.1**	**78.5**	58.7	**51.0**



Computation Saving

- 2.0x acceleration on 1.7B models



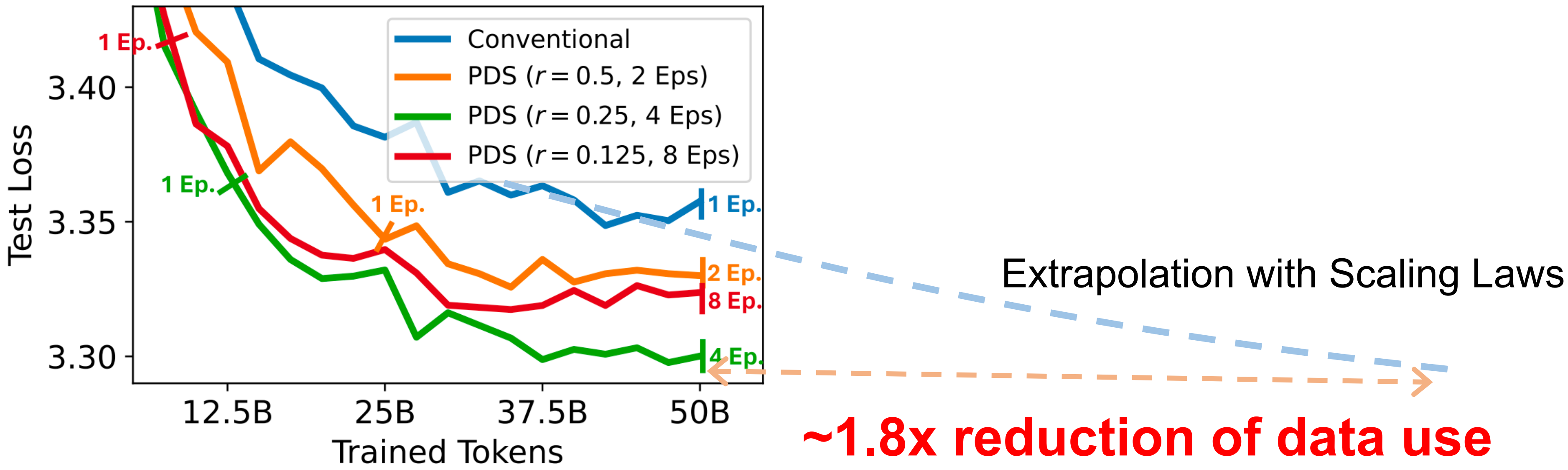
- PDS is efficient and offline
  - Select data once for all models

		FLOPs ($\times 10^{20}$)	Actual Time
PDS	Proxy γ-solver	0.49	15.2 Hours
	Data Scorer	0.063	1.50 Hours
	Data Selection	0.0	10.2 Minutes
Pre-Training		5.1	144 Hours

Data Utilization Improvement

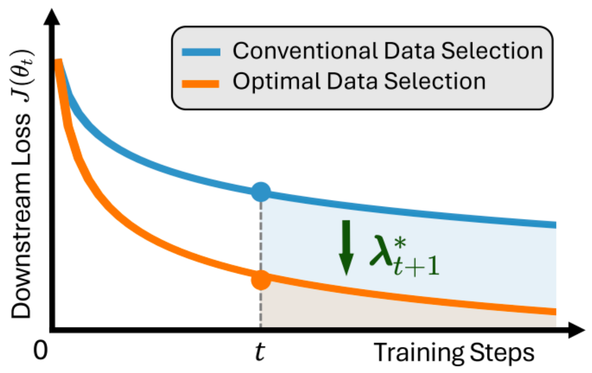
- Performance improvement with limit data (50B tokens)

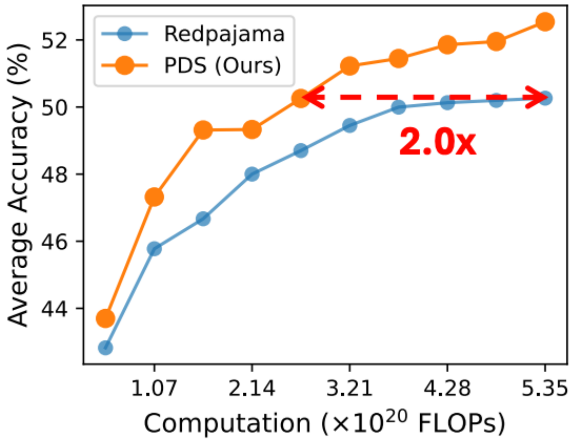
1 Ep.	Pre-Training (w/o Data Selection)
1 Ep. | 2 Ep.	Select 50% data, train for 2 epochs
1 Ep. | 2 Ep. | 3 Ep. | 4 Ep.	Select 25% data, train for 4 epochs
1 Ep. | 2 Ep. | 3 Ep. | 4 Ep. | 5 Ep. | 6 Ep. | 7 Ep. | 8 Ep.	Select 12.5% data, train for 8 epochs

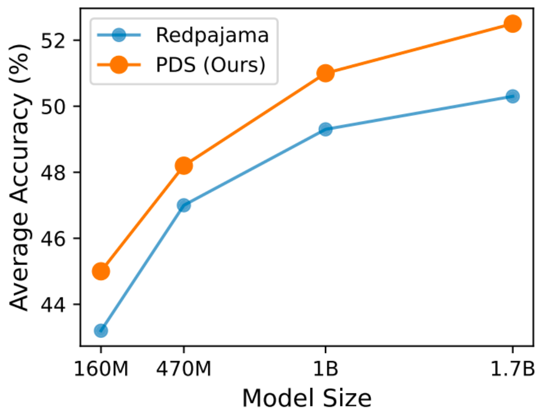


Conclusion

- A novel perspective for Data selection: Optimal Control problem
  - Good theoretical guarantees ✅
  - Efficient Implementation ✅
  - Sound empirical results ✅







A **rigorous, theory-driven alternative** to the ad-hoc practices that currently dominate LM pre-training

Thanks!

Paper: https://arxiv.org/abs/2410.07064

GitHub: https://github.com/microsoft/LMOps/tree/main/data_selection

HuggingFace: https://huggingface.co/Data-Selection

Paper:

Code:

清華大學 Tsinghua University

HuggingFace: