[ICML Oral] OPUS: Towards *Efficient* and *Principled* Data Selection in Large Language Model Pre-training in *Every* Iteration

Shaobo Wang *‡ Xuan Ouyang * Tianyi Xu * Yuzheng Hu Jialin Liu Guo Chen Tianyu Zhang
Junhao Zheng Kexin Yang Xingzhang Ren † Dayiheng Liu † Linfeng Zhang †

Presenter: Shaobo Wang
Shanghai Jiao Tong University

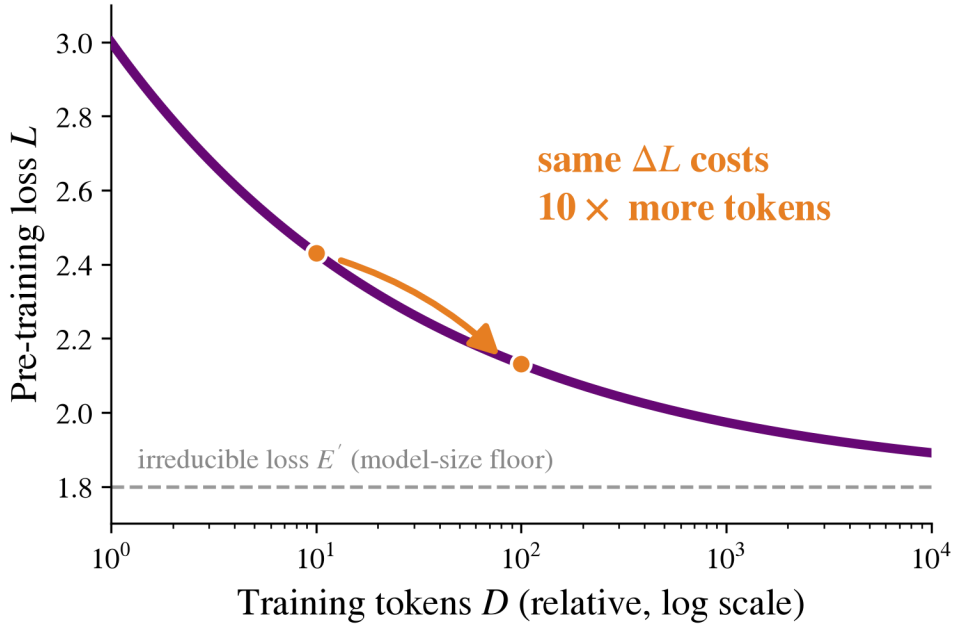
Background1 — LLM Scaling laws are driven by *data* and *model size*

- Scaling laws: pre-training loss is driven by **model size *N*** and **data volume *D***

$$\hat{L}(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

- Once *N* is fixed, the model term becomes a constant floor — **data is the only lever left**

$$L(D) = E' + \frac{B}{D^{\beta}}, \qquad \beta \approx 0.28$$



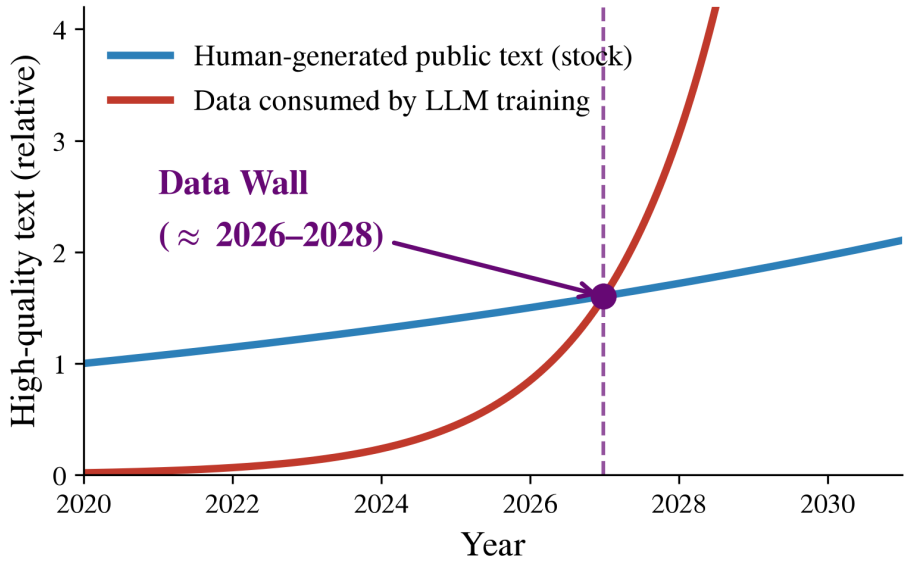
Diminishing returns: driving loss down by volume alone requires exponentially more tokens

Hoffmann et al. "Training compute-optimal large language models." NeurIPS 2022.

Background2 — The Data Wall Is Approaching: Quantity Is Running Out

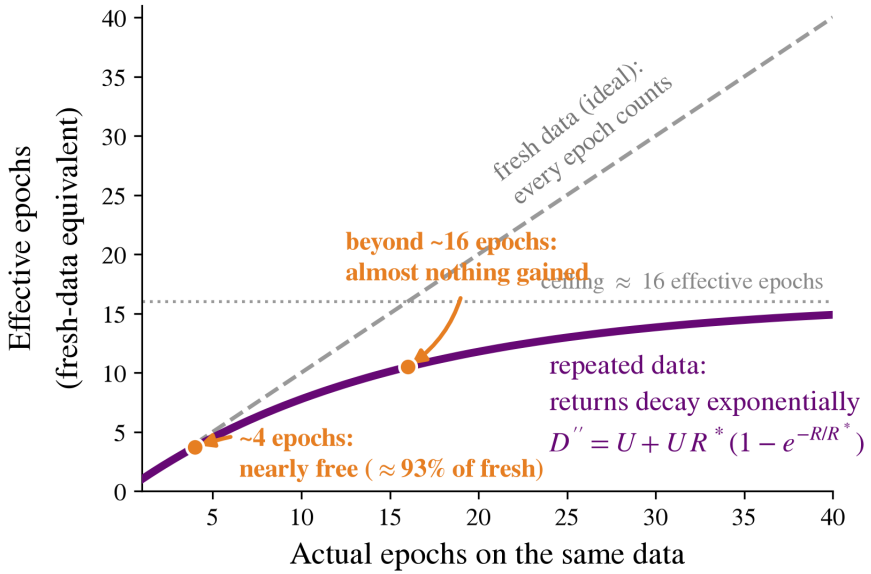
❑ LLMs consume data far faster than humans create it

- Frontier training sets grow ~2.4×/year
- High-quality public text projected to be exhausted in 2026–2028



❑ Repetition cannot fake more data

Returns from repetition decay exponentially: ~4 epochs are nearly free, beyond ~16 epochs adds almost nothing



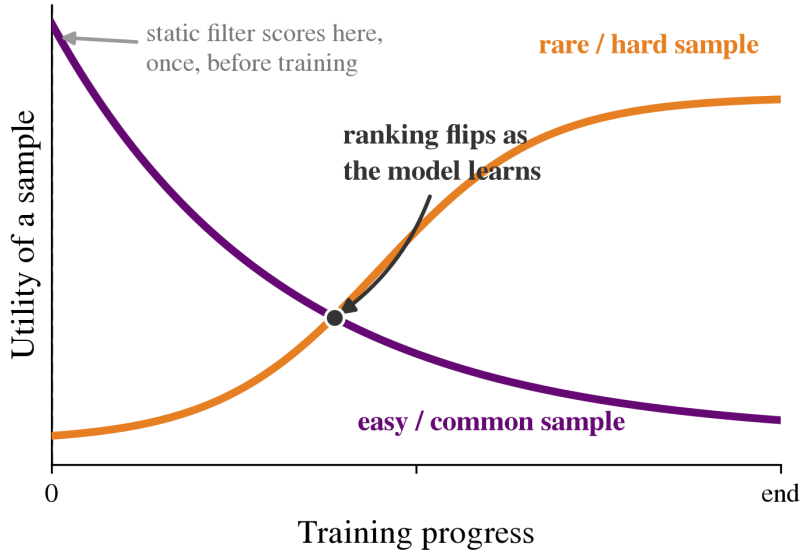
Pre-training must shift from *more tokens* to ***better tokens***

The question is no longer *how much* data, but *which* data

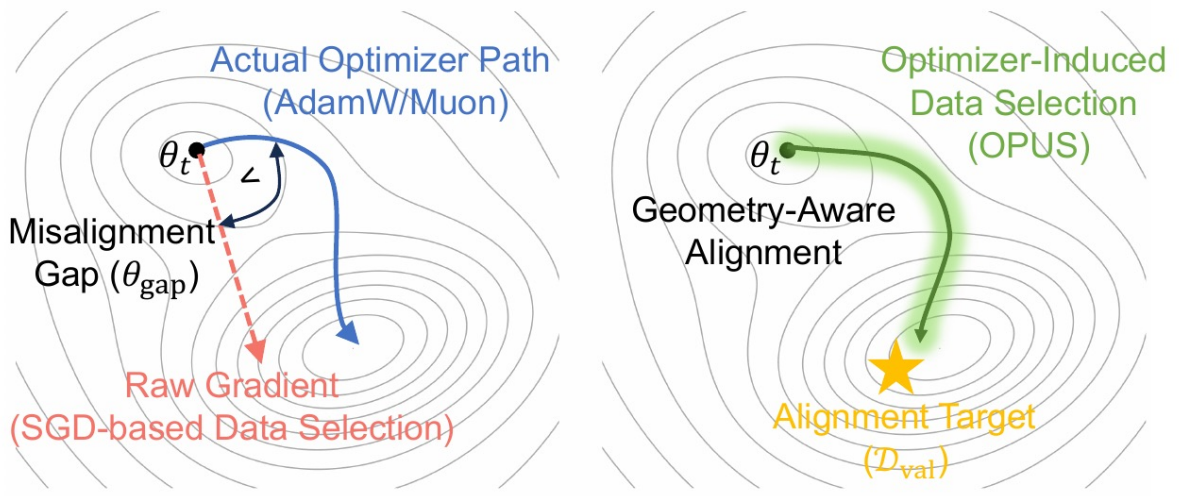
Villalobos et al. "Will we run out of data? Limits of LLM scaling based on human-generated data." 2022.
Muennighoff et al. "Scaling data-constrained language models." NeurIPS 2023.

Background3 — Today's Data Selection Relies on Heuristics

- **Static filters** (e.g., DCLM, FineWeb-Edu) are *training-agnostic*: they assume a sample's utility never changes as the model evolves



- **Dynamic selectors** score in *raw-gradient (SGD) space*, misaligned with the update geometry of modern optimizers (AdamW, Muon)



(a) Previous Data Selection (b) Optimizer-Induced Data Selection

We need a principled *dynamic data recipe* — deciding which tokens should shape the model at this specific *optimizer step*

What Data Selection Needs at Pre-Training Scale

- **Principled:** define data utility w.r.t. the update the optimizer actually applies — not the raw gradient.
- **Reliable:** build a *validation* signal that reflects downstream ability.
- **Efficient:** make data selection scalable and fast enough at pre-training data scales.

Principled Data Utility

- Data utility for a single datapoint z :

$$U_z^{(t)} := \mathcal{L}(\mathcal{D}_{\text{val}};, \theta_t) - \mathcal{L}\big(\mathcal{D}_{\text{val}}; \theta_t + \Delta\theta_t(z)\big)$$

where $\Delta\theta_t(z) = -\eta_t \mathbf{P}_t \nabla_\theta \mathcal{L}(z; \theta_t)$, $\mathbf{P}_t$ is the *optimizer-induced preconditioner*.

SGD

the reference point

$g_t = \nabla L(B_t; \theta_t)$

$\Delta\theta_t = -\eta_t g_t$

"just follow the gradient"

$\mathbf{P_t \approx I}$ **(identity)**

no reshaping — raw-gradient scoring is exact only here

AdamW

per-coordinate rescaling

$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$

$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$

$\Delta\theta_t = -\alpha_t \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \varepsilon)$

$\mathbf{P_t \approx \alpha_t\, Diag(\sqrt{\hat{v}_{t-1}} + \varepsilon)^{-1}}$

diagonal — every coordinate gets its own step size

Muon

whole-matrix orthogonalization

$M_t = \mu M_{t-1} + (1-\mu) g_t$

$\Delta W_t \propto -\text{NewtonSchulz}(M_t)$

freeze NS at step t $\Rightarrow$ $NS(Z) \approx S_t Z$

$\mathbf{P_t = \kappa_t\, S_t}$ **(dense, layerwise)**

matrix-valued — rotates and mixes coordinates of the update

Data utility must be measured on the **effective update** $\mathbf{u_t(z) = P_t\, \nabla L(z; \theta_t)}$, not on the raw gradient — raw-gradient scores are only correct for SGD

Principled Data Utility

- Data utility for a single datapoint z :

$$U_z^{(t)} := \mathcal{L}(\mathcal{D}_{\text{val}};, \theta_t) - \mathcal{L}\big(\mathcal{D}_{\text{val}}; \theta_t + \Delta\theta_t(z)\big)$$

where $\Delta\theta_t(z) = -\eta_t \mathbf{P}_t \nabla_\theta \mathcal{L}(z; \theta_t)$, $\mathbf{P}_t$ is the *optimizer-induced preconditioner*.

- **First-order Taylor expansion** $(\mathbf{H}_{\text{val}} \approx \mathrm{I})$ gives the final tractable score:

$$U_z^{(t)} \approx \eta_t \langle \mathbf{P}_t \nabla_\theta \mathcal{L}(z; \theta_t), \nabla_\theta \mathcal{L}(\mathcal{D}_{\text{val}}, \theta_t) \rangle - \eta_t^2 \left\langle \mathbf{P}_t \nabla_\theta \mathcal{L}(z; \theta_t) \sum_{z_j \in \widehat{\mathcal{B}}_t} \mathbf{P}_t \nabla_\theta \mathcal{L}\big(z_j; \theta_t\big) \right\rangle$$

Importance (Alignment) Diversity (Redundancy)

where $\widehat{\mathcal{B}}_t$ is the already-selected subset at step t.

Principled Data Utility

- Data utility for a single datapoint z :

$$U_z^{(t)} := \mathcal{L}(\mathcal{D}_{\text{val}};, \theta_t) - \mathcal{L}\big(\mathcal{D}_{\text{val}}; \theta_t + \Delta\theta_t(z)\big)$$

where $\Delta\theta_t(z) = -\eta_t \mathbf{P}_t \nabla_\theta \mathcal{L}(z; \theta_t)$, $\mathbf{P}_t$ is the *optimizer-induced preconditioner*.

- **First-order Taylor expansion** $(\mathbf{H}_{\text{val}} \approx \mathrm{I})$ gives the final tractable score:

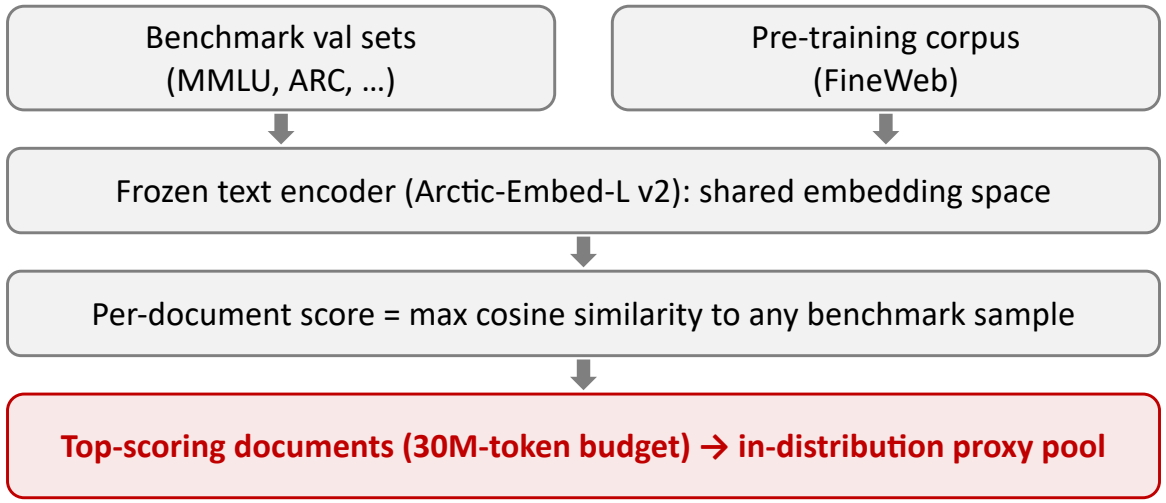
$$U_z^{(t)} \approx \eta_t \langle \mathbf{P}_t \nabla_\theta \mathcal{L}(z; \theta_t), \nabla_\theta \mathcal{L}(\mathcal{D}_{\text{val}}, \theta_t) \rangle - \eta_t^2 \left\langle \mathbf{P}_t \nabla_\theta \mathcal{L}(z; \theta_t) \sum_{z_j \in \widehat{\mathcal{B}}_t} \mathbf{P}_t \nabla_\theta \mathcal{L}\big(z_j; \theta_t\big) \right\rangle$$

How to determine a good validation set?

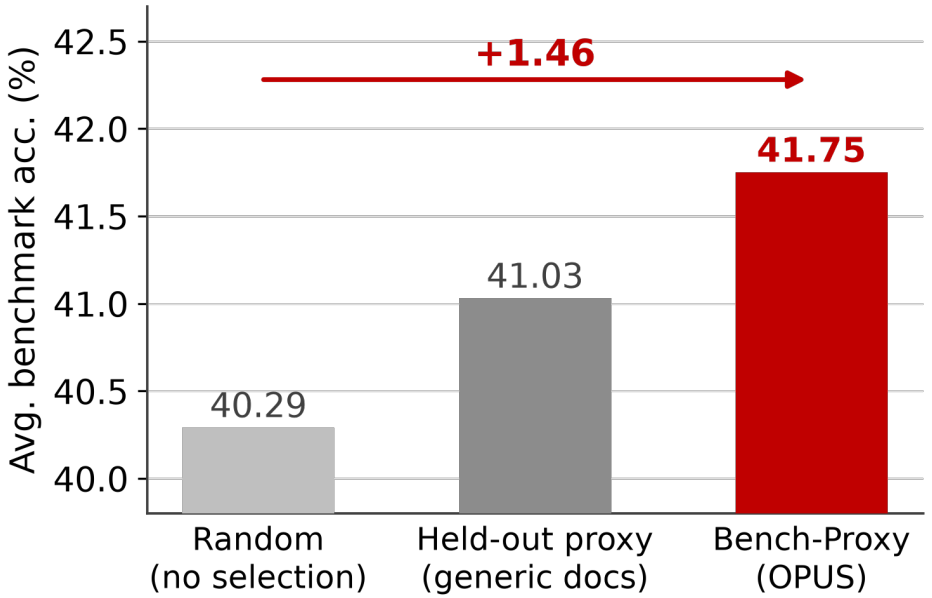
where $\widehat{\mathcal{B}}_t$ is the already-selected subset at step t.

Build a reliable validation set

- **The dilemma:** raw benchmark data as validation ⇒ **distribution shift + gradient noise**; a random held-out set is stable but blind to downstream ability
- **Bench-Proxy:** **retrieve benchmark-aligned documents from the pre-training corpus itself** — aligned with target tasks, yet inside the pre-training manifold

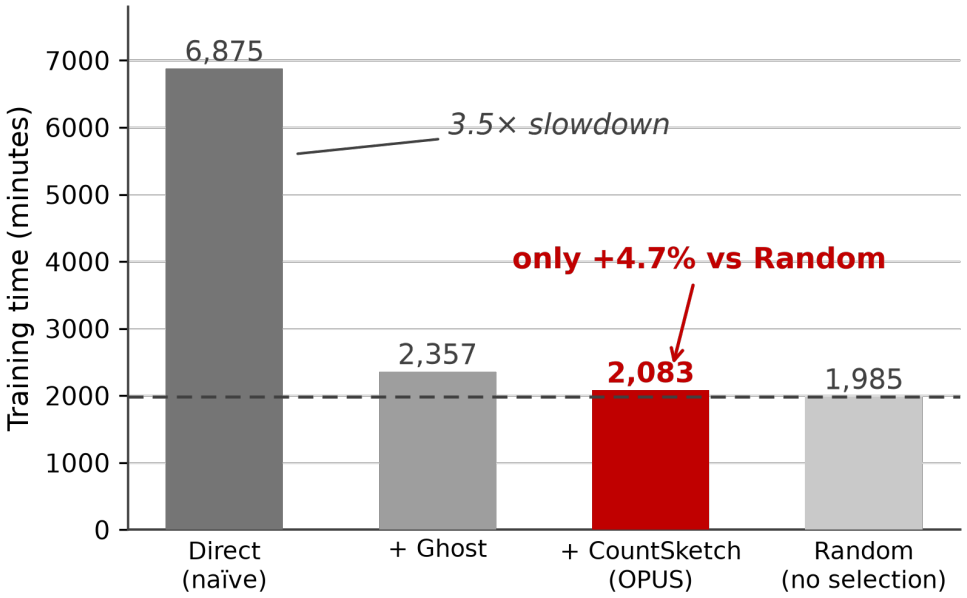


Each step: sample a mini-batch from the proxy pool → proxy gradient direction



Make everything efficient and scalable

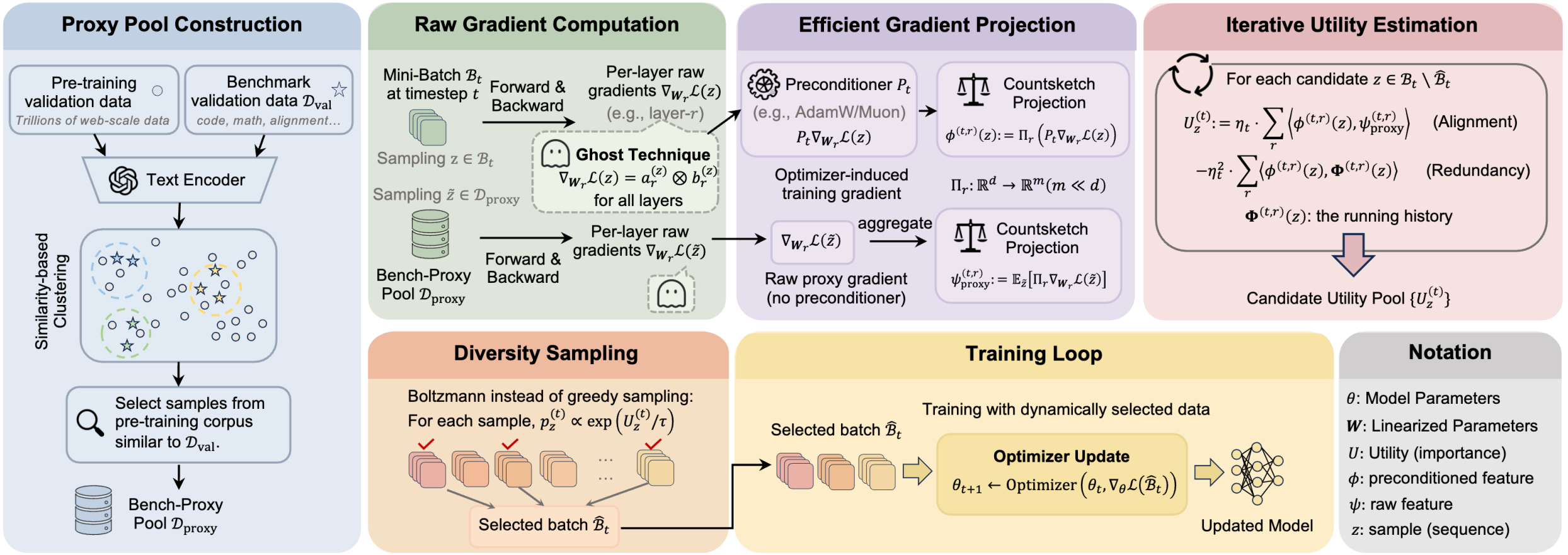
- **Naïve online scoring:** per-sample forward/backward + materializing full per-sample gradients ⇒ 3.5× slowdown
- **Ghost technique:** a linear layer's per-sample gradient is rank-1 (a ⊗ b) — reuse activations & output grads from the standard pass; discard layer-by-layer, never materialize full gradients
- **CountSketch projection:** all utility inner products computed in an $m \ll d$ sketch space; AdamW's diagonal preconditioner applied on the fly at $O(d_{\text{in}} + d_{\text{out}})$



Only +4.7% overhead vs Random (2,083 vs 1,985 min)
every-step selection at pre-training scale

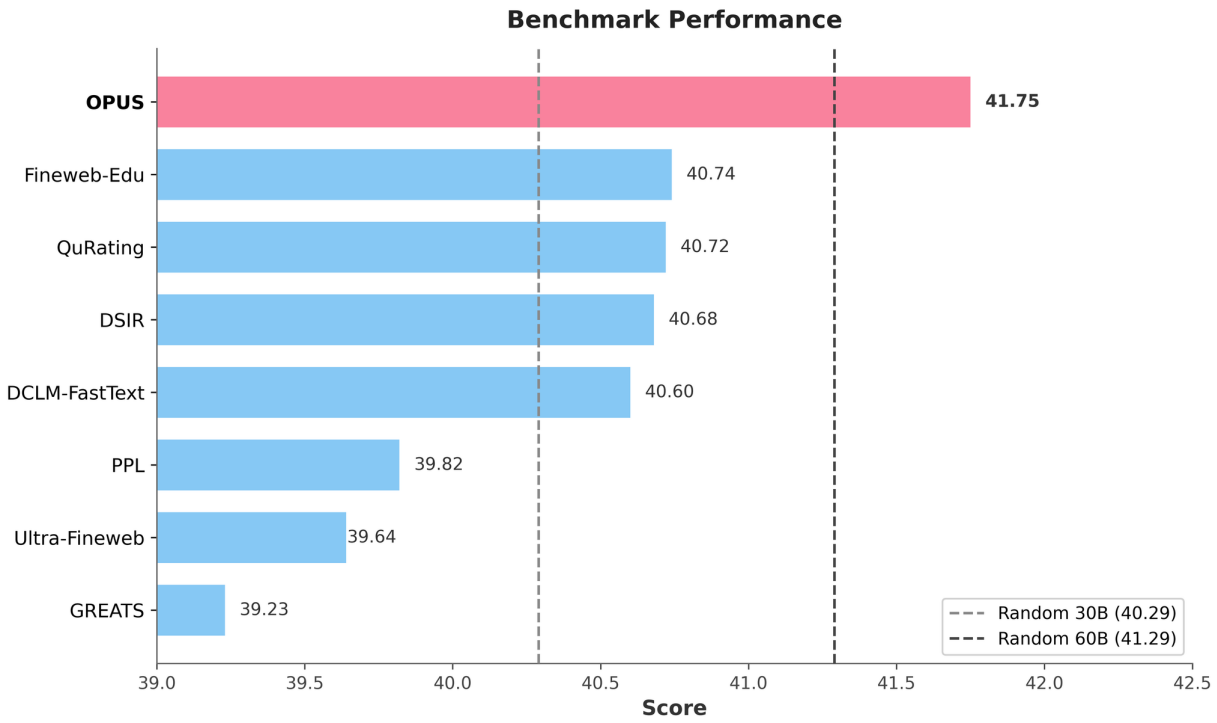
Put everything together

Optimizer-induced Projected Utility Selection

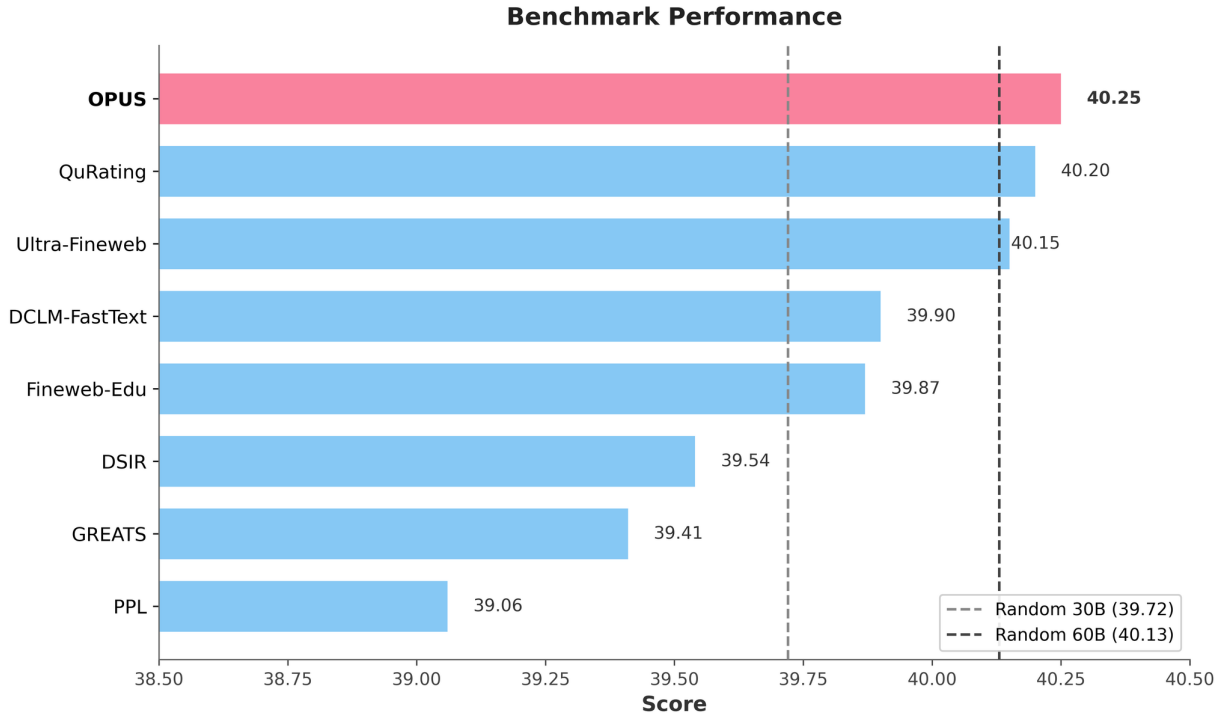


Results on the Muon Optimizer

❑ Across 10 benchmarks (including knowledge, reasoning, and commonsense), OPUS consistently achieves the best avg. performance under the ***Muon*** optimizer.



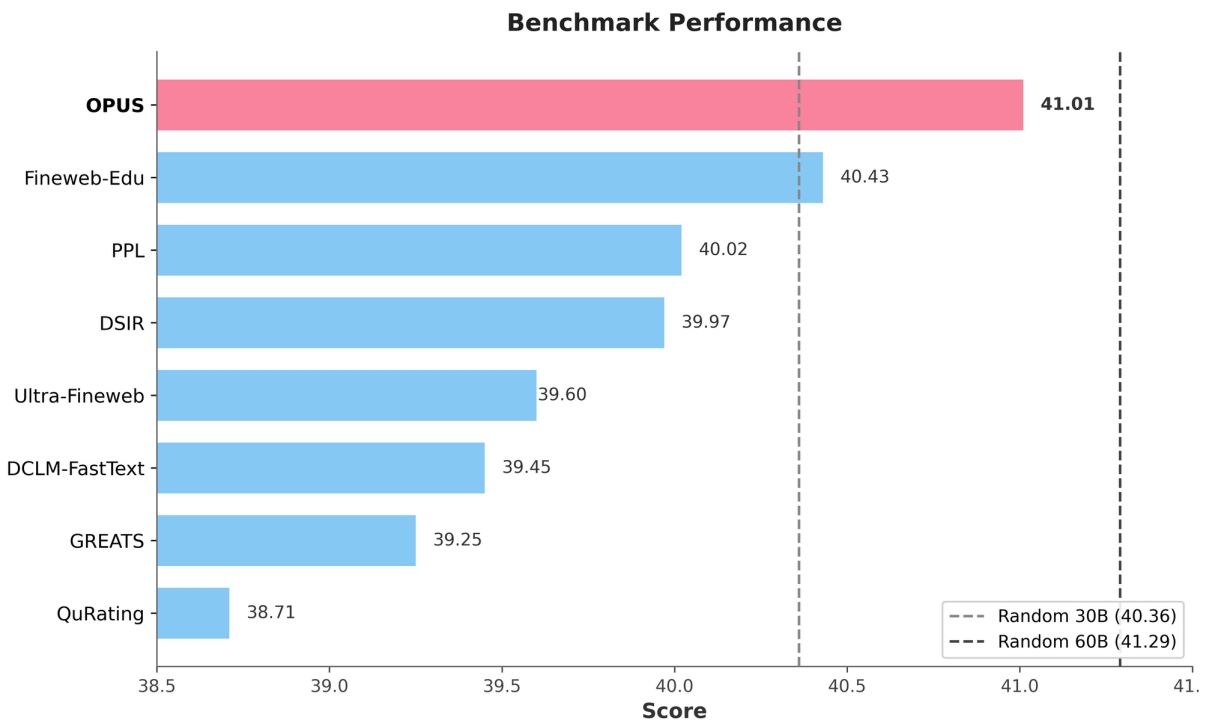
GPT2-XL on FineWeb (30B tokens)



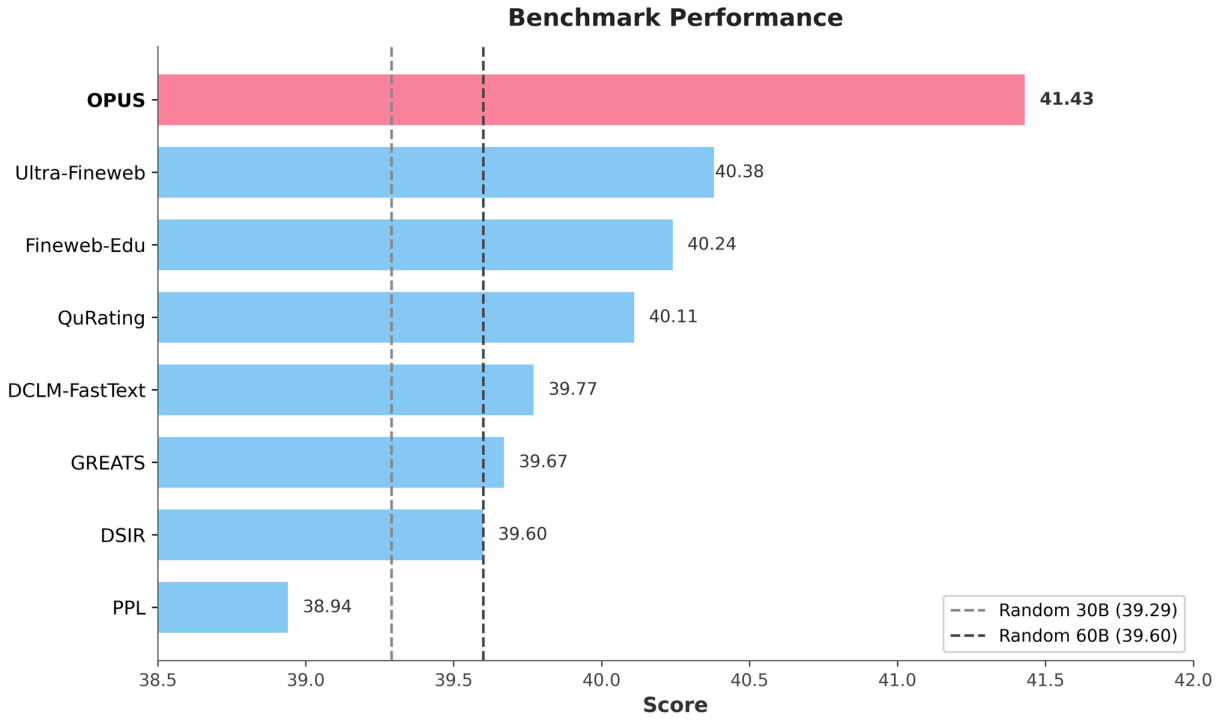
GPT2-Large on FineWeb (30B tokens)

Results on the AdamW Optimizer

❑ Across 10 benchmarks (including knowledge, reasoning, and commonsense), OPUS consistently achieves the best avg. performance under the ***AdamW*** optimizer.



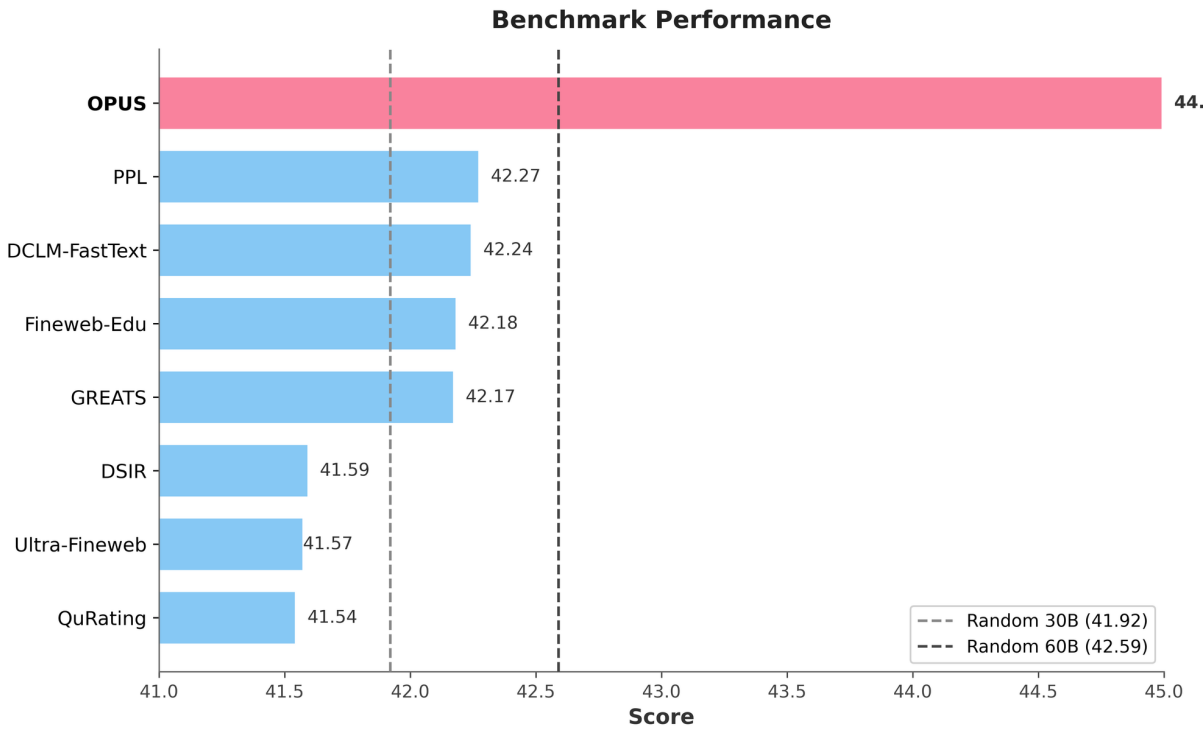
GPT2-XL on FineWeb (30B tokens)



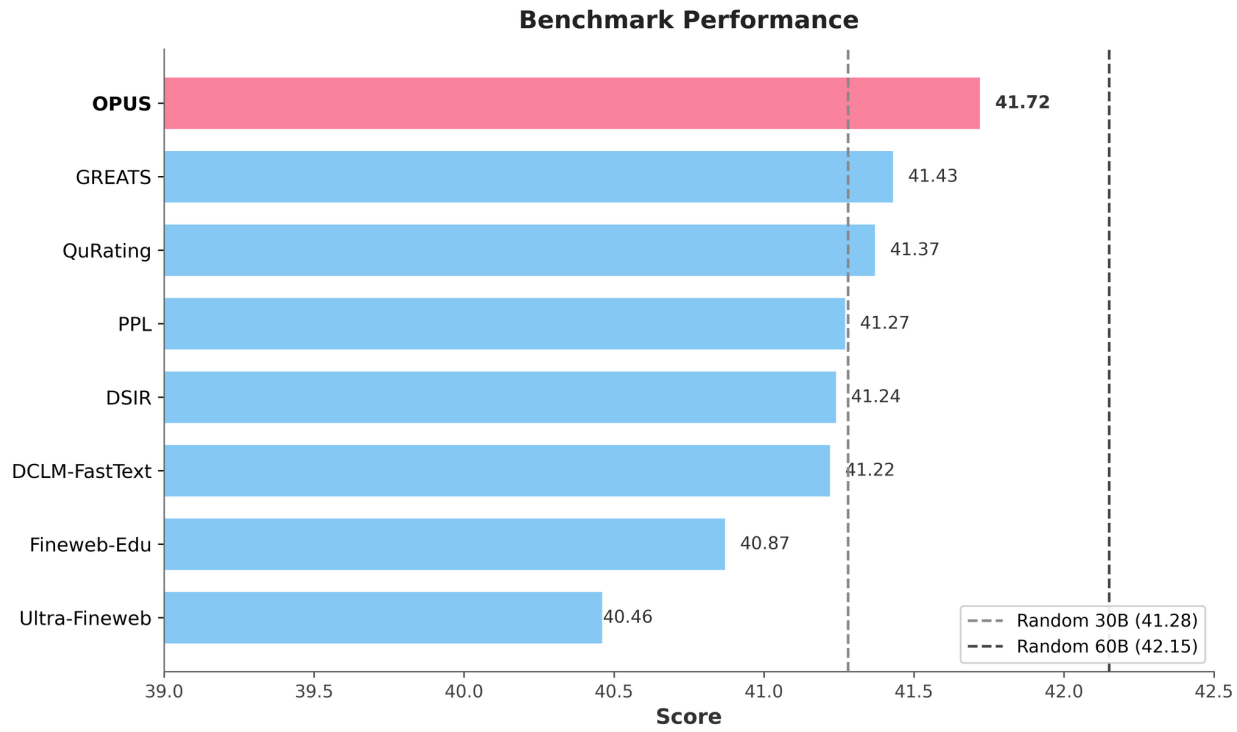
GPT2-Large on FineWeb (30B tokens)

Combined with static filters

- **Strict regime:** OPUS selects only from the mid-quality (score-3), while baselines use the superior score-4+5 subset
- Even from worse data OPUS wins: **44.99 on GPT-2 XL** vs best baseline 42.27 trained on higher-quality data



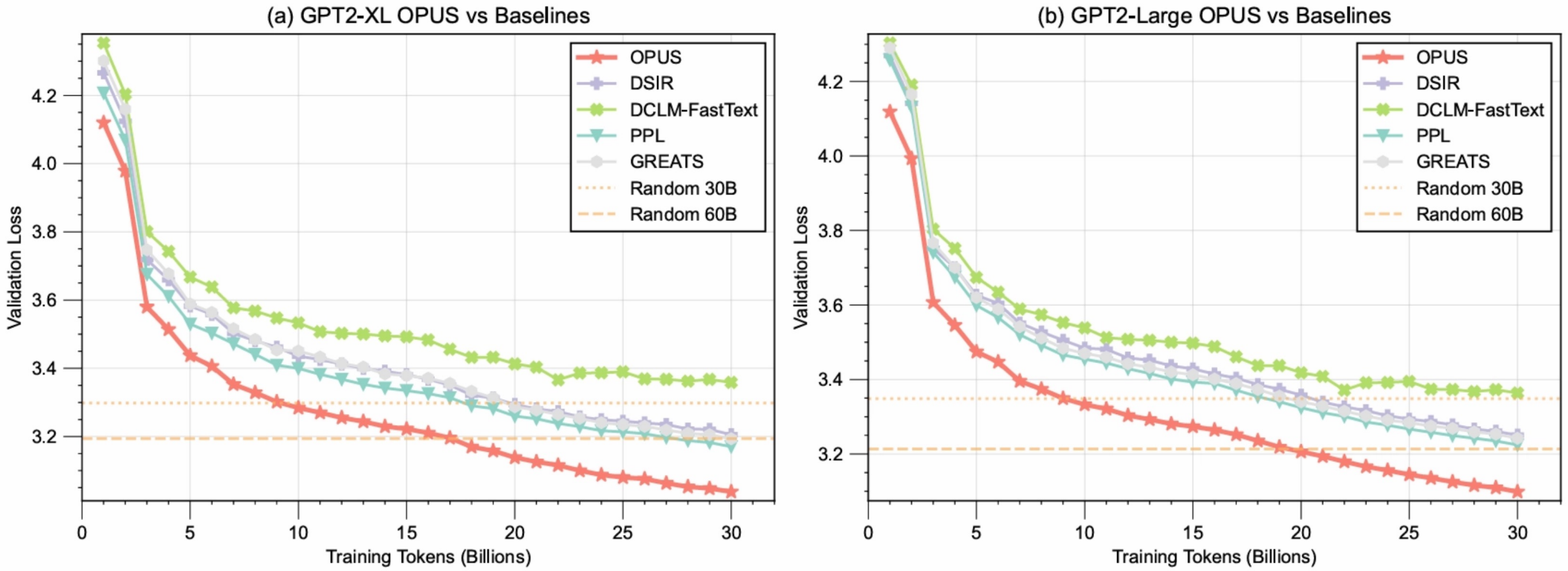
GPT2-XL on Muon (30B tokens)



GPT2-Large on Muon (30B tokens)

Validation loss

- OPUS achieves a **faster and consistently lower** loss trajectory — while selecting only from the score-3 pool
- GPT-2 XL: **reaches Random @ 60B's loss with only ~17B tokens** — ≈3.5× faster convergence



Domain PPL on FineWeb and FineWeb-Edu

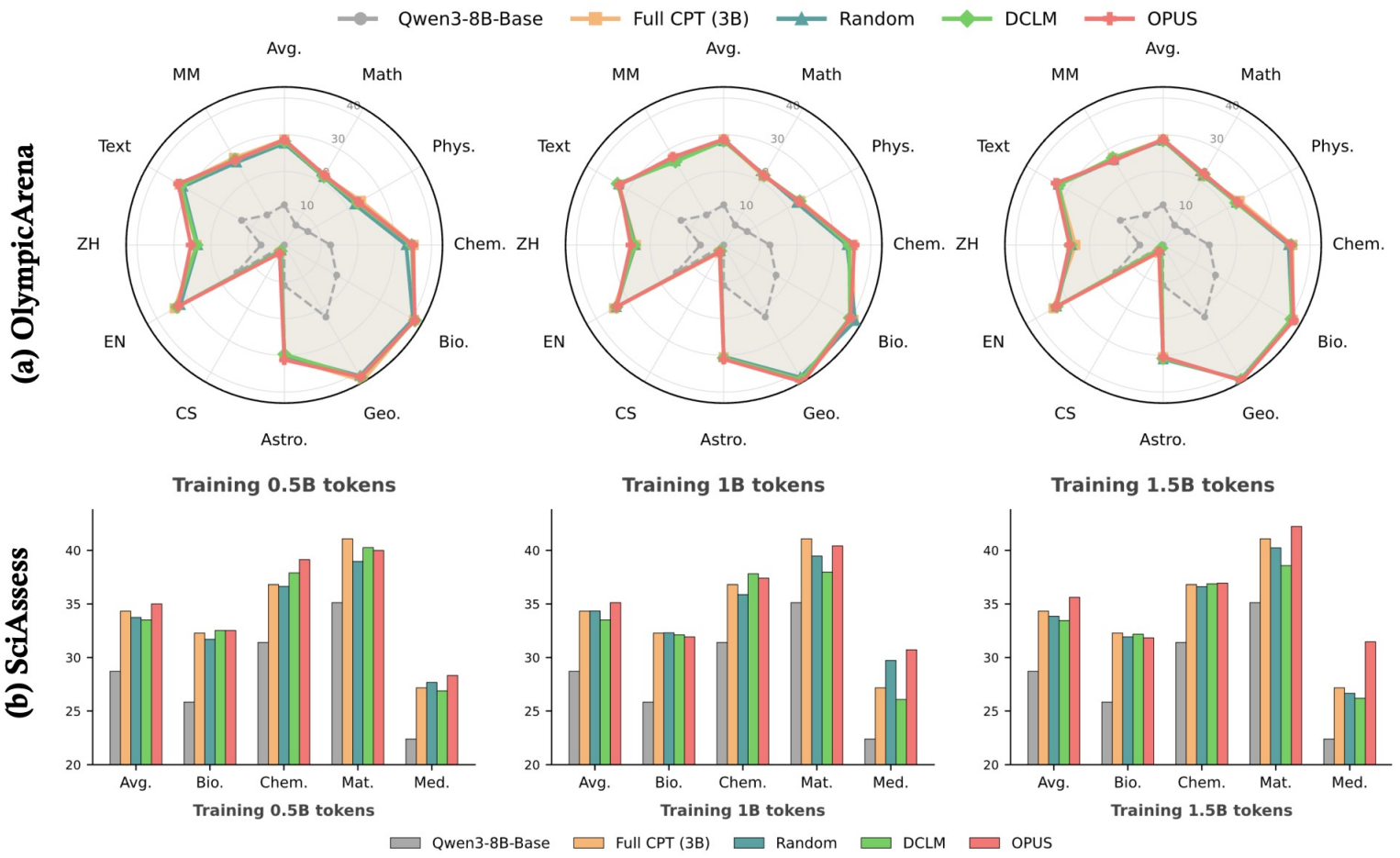
OPUS outperforms all dynamic and static methods.

Method	Health	Business	Politics	Education	History	Lifestyle	Science	Arts & Lit.	Entertainment	Computing	Avg.
GPT-2 Large with Muon optimizer on 30B update tokens of FineWeb											
Random (30B)	3.21	3.26	3.28	3.31	3.32	3.37	3.40	3.49	3.56	3.62	3.38
DSIR	3.21	3.26	3.28	3.31	3.32	3.38	3.40	3.49	3.57	3.63	3.39
DCLM-FastText	**3.17**	3.24	3.26	3.30	3.36	3.37	3.36	3.46	**3.54**	3.60	3.37
FineWeb-Edu	**3.17**	3.24	**3.25**	**3.28**	3.26	3.41	3.34	3.48	3.58	3.61	3.36
QuRating	3.40	3.60	3.79	3.57	3.68	4.05	3.61	3.92	4.27	4.11	3.80
UltraFineweb	3.19	3.29	3.30	3.32	**3.30**	3.43	3.38	3.50	3.59	3.62	3.39
PPL	3.22	3.26	3.28	3.31	3.32	3.37	3.39	3.49	3.56	3.61	3.38
GREATS	3.25	3.31	3.33	3.36	3.38	3.42	3.46	3.55	3.62	3.66	3.43
OPUS (Ours)	3.18	**3.23**	**3.25**	**3.28**	**3.30**	**3.34**	**3.37**	**3.47**	**3.54**	**3.58**	**3.35**
GPT-2 XL with Muon optimizer on 30B update tokens of FineWeb											
Random (30B)	3.18	3.25	3.26	3.29	3.30	3.35	3.40	3.49	3.56	3.61	3.37
DSIR	3.15	3.22	3.23	3.26	3.25	3.32	3.35	3.44	3.52	3.56	3.33
DCLM-FastText	3.15	3.23	3.25	3.31	3.25	3.36	3.34	3.45	3.53	3.60	3.35
FineWeb-Edu	3.16	3.23	3.24	3.28	3.25	3.40	3.34	3.47	3.62	3.60	3.36
QuRating	3.27	3.53	3.67	3.47	3.59	3.91	3.51	3.83	4.14	3.96	3.69
UltraFineweb	3.10	3.20	3.19	3.24	3.21	3.33	3.29	3.41	3.50	3.53	3.30
PPL	3.11	3.17	3.18	3.21	3.22	3.27	3.30	3.40	3.46	3.50	3.28
GREATS	3.22	3.29	3.29	3.33	3.32	3.39	3.42	3.51	3.58	3.66	3.40
OPUS (Ours)	**3.08**	**3.15**	**3.16**	**3.18**	**3.21**	**3.23**	**3.29**	**3.39**	**3.45**	**3.44**	**3.26**

Method	Health	Business	Politics	Education	History	Lifestyle	Science	Arts & Lit.	Entertainment	Computing	Avg.
GPT-2 Large with Muon optimizer on 30B update tokens of FineWeb-Edu Subset (score $\geq$ 3)											
Random (30B)	3.27	3.52	3.58	3.49	3.48	3.81	3.43	3.75	4.03	3.82	3.62
DSIR	3.29	3.55	3.61	3.52	3.49	3.84	3.46	3.77	4.05	3.86	3.64
DCLM-FastText	3.34	3.61	3.67	3.59	3.58	3.89	3.5	3.82	4.09	3.89	3.70
FineWeb-Edu	3.41	3.67	3.72	3.62	3.60	3.97	3.57	3.87	4.17	3.98	3.76
QuRating	3.46	3.76	3.90	3.65	3.79	4.13	3.70	4.00	4.36	4.16	3.89
UltraFineweb	3.42	3.72	3.87	3.66	3.77	4.05	3.58	3.96	4.26	4.00	3.83
PPL	3.25	3.49	3.54	3.46	3.44	3.78	3.41	3.71	3.99	3.80	3.59
GREATS	3.29	3.55	3.62	3.52	3.50	3.84	3.46	3.77	4.06	3.86	3.65
OPUS (Ours)	**3.14**	**3.34**	**3.44**	**3.37**	**3.37**	**3.63**	**3.38**	**3.63**	**3.87**	**3.71**	**3.49**
GPT-2 XL with Muon optimizer on 30B update tokens of FineWeb-Edu Subset (score $\geq$ 3)											
Random (30B)	3.25	3.51	3.55	3.48	3.45	3.79	3.42	3.73	4.00	3.83	3.60
DSIR	3.24	3.50	3.54	3.47	3.44	3.78	3.41	3.72	4.00	3.81	3.59
DCLM-FastText	3.36	3.64	3.70	3.62	3.61	3.94	3.52	3.86	4.13	3.94	3.73
FineWeb-Edu	3.29	3.55	3.58	3.50	3.49	3.82	3.45	3.75	4.02	3.83	3.63
QuRating	3.50	3.79	3.93	3.70	3.83	4.18	3.73	4.04	4.39	4.24	3.93
UltraFineweb	3.43	3.74	3.90	3.68	3.80	4.07	3.59	3.99	4.28	4.02	3.85
PPL	3.22	3.47	3.50	3.44	3.40	3.74	3.39	3.69	3.96	3.77	3.56
GREATS	3.29	3.55	3.60	3.52	3.49	3.84	3.45	3.77	4.05	3.88	3.64
OPUS (Ours)	**3.11**	**3.31**	**3.37**	**3.34**	**3.31**	**3.59**	**3.33**	**3.58**	**3.83**	**3.69**	**3.45**

Continued Pre-training on 8B models

- Qwen3-8B-Base + SciencePedia: improves **domain knowledge (SciAssess)** & **scientific reasoning (OlympicArena)**
- **Best performance at 0.5B tokens — beats Random CPT @ 3B** ⇒ 6× data efficiency



Case Study: OPUS as an Online and Post-hoc Data Filter

- **No training at all:** a single scoring pass with the unmodified checkpoint — a buffer of 32 mixing clean FineWeb-Edu documents with real toxic text (**Jigsaw toxic comments + Enron spam**)
- Selected: **well-formed educational prose**; Rejected: **pharmacy ads, mail-server bounces, etc.**

OPUS-Selected: clean candidates — *mean score = +5.293*

S1 Selected +6.274 JOHN BAKER UNRAVELS HISTORY Genealogy Expert John F. Baker Jr., has written the most accessible and exciting work of African American...	**S2** Selected +5.153 Ayurveda traces its origins to the Rig Veda, the world's oldest surviving book in an Indo-European language. The Rig Veda, 3000 B.C., is a c...
S3 Selected +5.153 The R/V David Folger's route home takes it north along the Hudson River and then through the Champlain Canal to Lake Champlain. On July 30, ...	**S4** Selected +5.153 Within a library collection, materials are typically organized by subject. Librarians assign a call number based on a work's subject and sou...
S5 Selected +5.153 NY1's "Making Census Of It" series continues this week with a spotlight on the Bronx, the borough that experienced the largest growth in the...	**S6** Selected +5.153 Developed in the 1920s by the legendary physical trainer, Joseph H. Pilates, "The Pilates Method" is an exercise system focused on improving...
S7 Selected +5.153 The State of Food Insecurity in the World 2011 highlights the differential impacts that the world food crisis of 2006-08 had on different co...	**S8** Selected +5.153 In diachronic comparison of languages, say PIE to Latin to Romance, it is a classic recognition that the later languages strictly lose some ...

OPUS-Rejected: toxic candidates — *mean score = +0.548*

S1 Rejected −0.025 greece based professionals and organizations greece based professionals and organizations you areinvited to register on training consortiumj...	**S2** Rejected +0.037 confidence confidence attn : manager hello dear , i am well confidence of your capability to assist me in a transaction for mutual benefit f...
S3 Rejected +0.074 fantastic investors portfoiio fantastic investors portfoiio shirley , investor alert - lrcj - brand new stock for your attention lauraan cor...	**S4** Rejected +0.310 failure notice failure notice hi . this is the qmail - send program at harrisburg . villagetech . com . i ' m afraid i wasn ' t able to deli...
S5 Rejected +0.438 preferred non - smoker rates for smokers preferred non - smoker rates for smokers case study # 1 male - 63 $ 5 , 000 , 000 face good health ...	**S6** Rejected +0.571 reminder reminder psychologists bottled overflowing deeply bloomfield singers hepburn odder hopkinsian reassembles stupidity coolies picasso...
S7 Rejected +0.601 impress your girl with a huge c*****t ! impress your girl with a huge c*****t ! heya ! has your c*m ever dribbled and you wish it had shot o...	**S8** Rejected +2.375 business assistance business assistance dear sir . first , i must solicit your confidence in this transaction , this is by virture of its na...

cjadams et al. Toxic Comment Classification Challenge, Kaggle, 2017.
Metsis et al. Spam Filtering with Naive Bayes — Which Naive Bayes? CEAS, 2006.

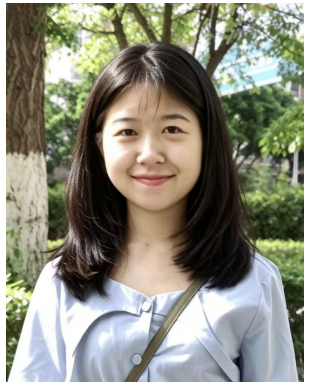
Thanks to all the collaborators!

Shaobo Wang
Ph.D@SJTU
On the job market

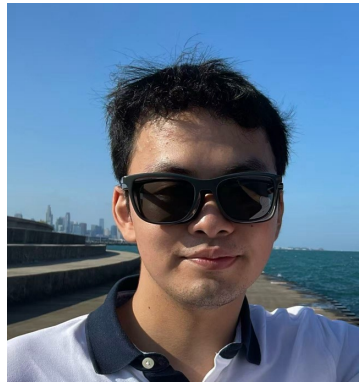
Xuan Ouyang
Ph.D@UW-Madison
Seeking summer Intern

Tianyi Xu
Ohio State University

Jialin Liu
Undergrad @ SJTU

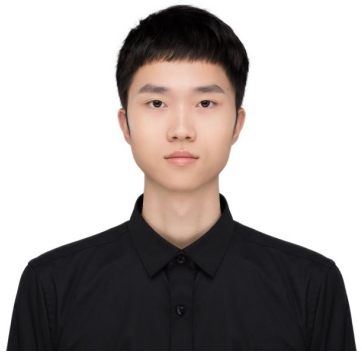


Guo Chen
Intern @ SJTU



Yuzheng Hu
Ph.D@UIUC
Google Research

Tianyu Zhang
Ph.D@MILA
Mistral AI



Junhao Zheng
Alibaba Qwen Team

Kexin Yang
Alibaba Qwen Team

Xingzhang Ren
Alibaba Qwen Team

Dayiheng Liu
Alibaba Qwen Team

Linfeng Zhang
Assistant Prof.@SJTU

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Paper Code WeChat

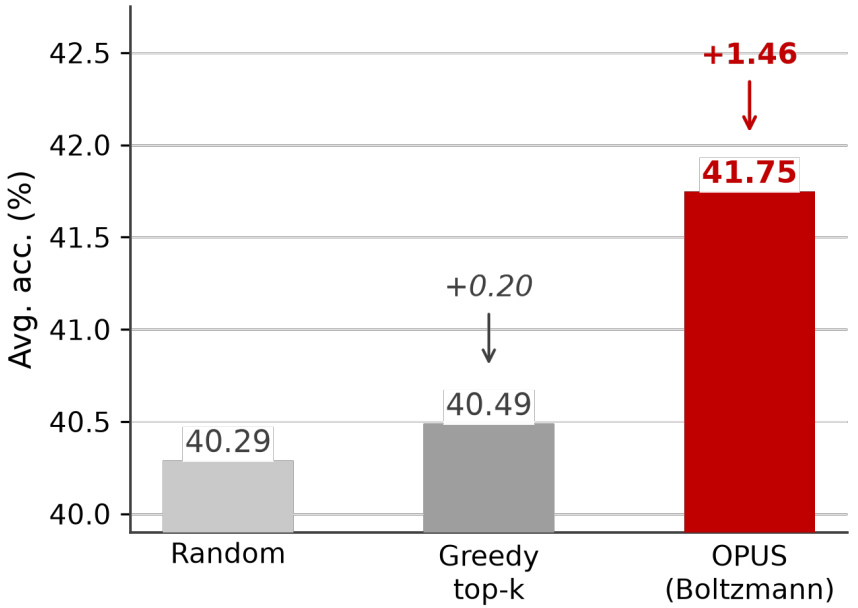
Thanks!

ICML
International Conference
On Machine Learning

Presenter: Shaobo Wang
Shanghai Jiao Tong University

Ablation — why Boltzmann sampling over greedy top-k?

- Greedy fully trusts a noisy signal: utility is a small-batch estimate, and top-K **over-concentrates on overlapping candidates**
- Boltzmann sampling $p(z) \propto \exp(U_t(z)/\tau)$ favors high utility while retaining complementary samples ($\tau = 0.9$)



Greedy buys almost nothing (+0.20); sampling unlocks +1.46 over Random
Diversity under noise is part of the objective, not a heuristic add-on

Ablation — hyperparameter sensitivity

- OPUS is stable across settings and beats Random in most configurations, including buffer sizes, temperatures, and sketch dimension in CountSketch.