Programming every example: Lifting pre-training data quality like experts at scale

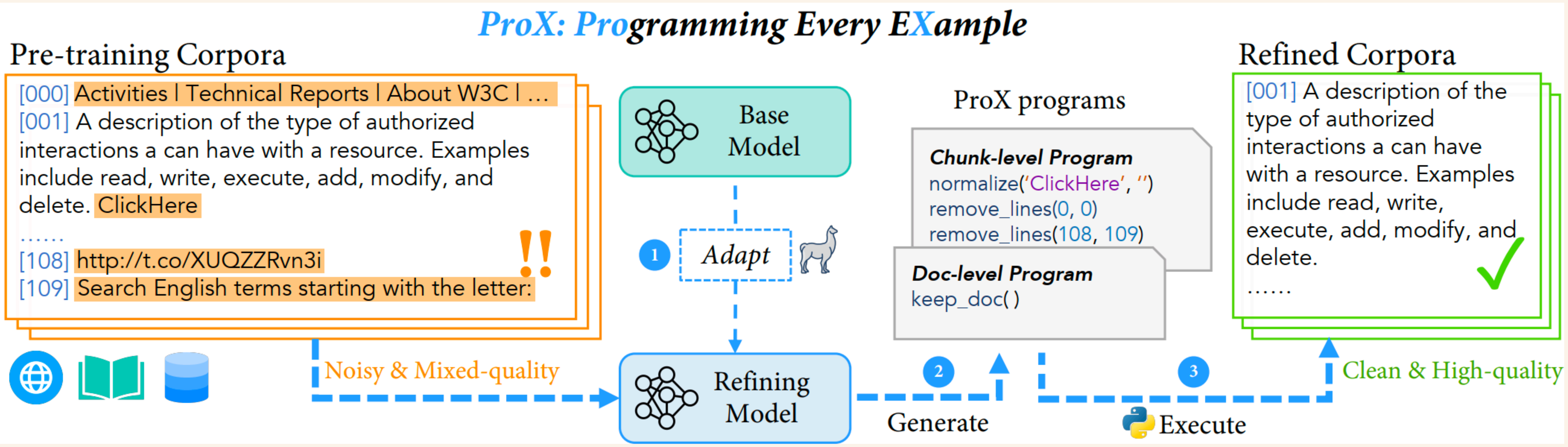
Fan Zhou [1,3*], Zengzhi Wang [1,3*], Qian Liu [2], Junlong Li [1], Pengfei Liu [1,3]

[1]Shanghai Jiao Tong University [2]Sea AI Lab [3]GAIR Lab

TL;DR Still worried about the potential noise and the low quality of your rule-cleaned pre-training corpora? **Try ProX!** In ProX, we use 0.3B LLMs to seamlessly refine your pre-training dataset, providing a clean start for your LLM training. ProX serves as a LLM-driven framework which generate **sample-wise**, **executable** cleaning programs to clean **each and every one** of your pre-training data samples.

- Our 1.7B model, trained on ProX corpus with 50B tokens training, performs on par with TinyLlama-1.1B which is trained on 3T tokens.
- Continual pre-training of CodeLlama-7B on OpenWebMath with 10B tokens refined by ProX matches Llemma-7B, also pre-trained from CodeLlama-7B, trained on 200B tokens.
- We release "ProX" series of pre-training dataset including >600B general pre-training data, and 5B math corpus.

How ProX Works? *Program Design* 👉 *LLM Annotation* 👉 *SFT on Tiny LLMs*



ProX: Programming Every EXample

Pre-training Corpora

[000] Activities | Technical Reports | About W3C | …

[001] A description of the type of authorized interactions a can have with a resource. Examples include read, write, execute, add, modify, and delete. ClickHere

……

[108] http://t.co/XUQZZRvn3i

[109] Search English terms starting with the letter:

!!

Noisy & Mixed-quality

Base Model

1 Adapt

Refining Model

2 Generate

ProX programs

Chunk-level Program
```
normalize('ClickHere', '')
remove_lines(0, 0)
remove_lines(108, 109)
```

Doc-level Program
```
keep_doc()
```

3 Execute

Refined Corpora

[001] A description of the type of authorized interactions a can have with a resource. Examples include read, write, execute, add, modify, and delete.

……

✓

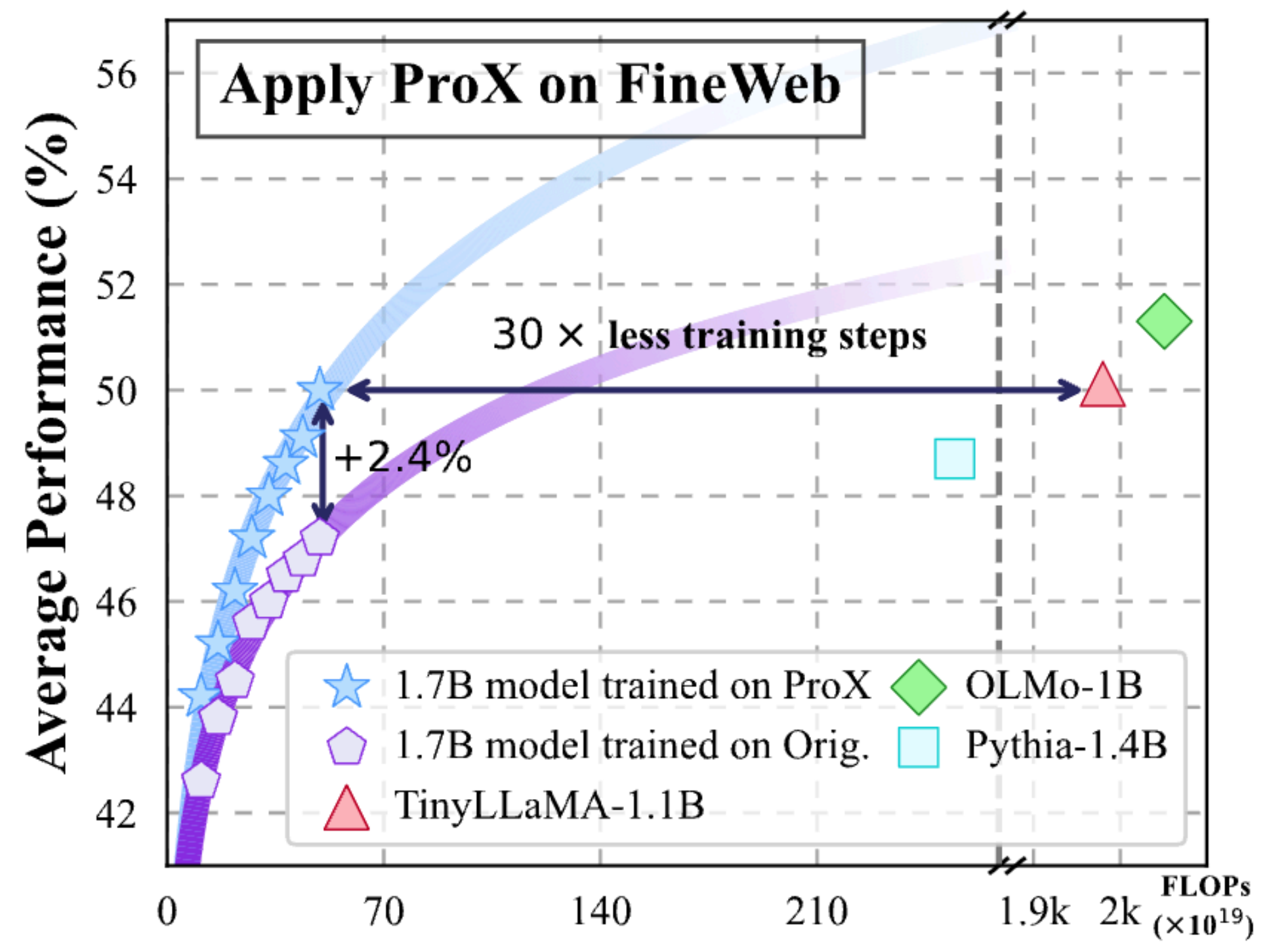
Clean & High-quality

😰 Due to visa restrictions, the authors are not able to attend in person. If you have any questions, please also feel free to reach out via the QR codes!

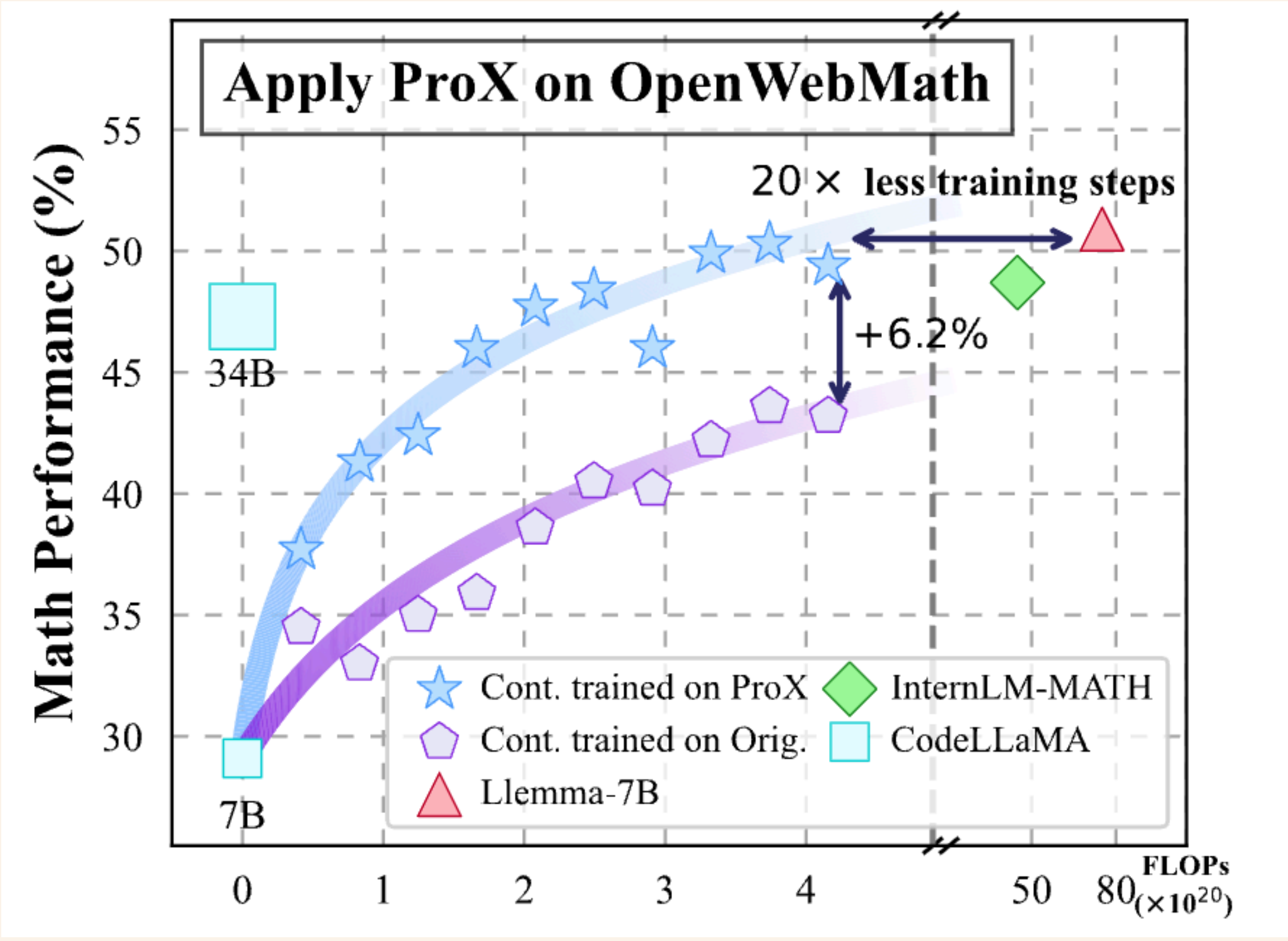
👉 Github Huggingface X.COM

Experimental Results

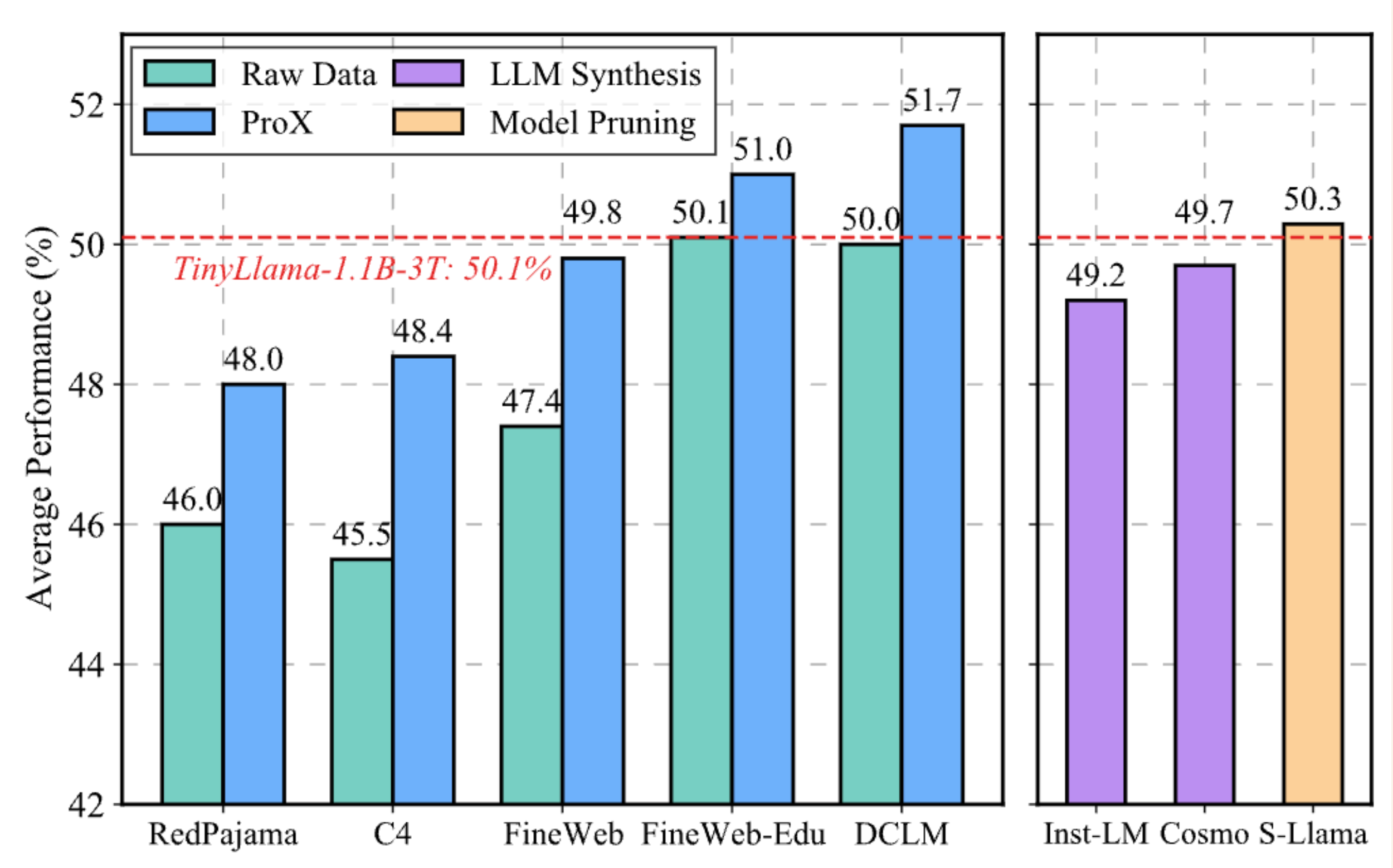
It works on general domain.



It works on specific domain.



And also better than …….



With less overall compute!

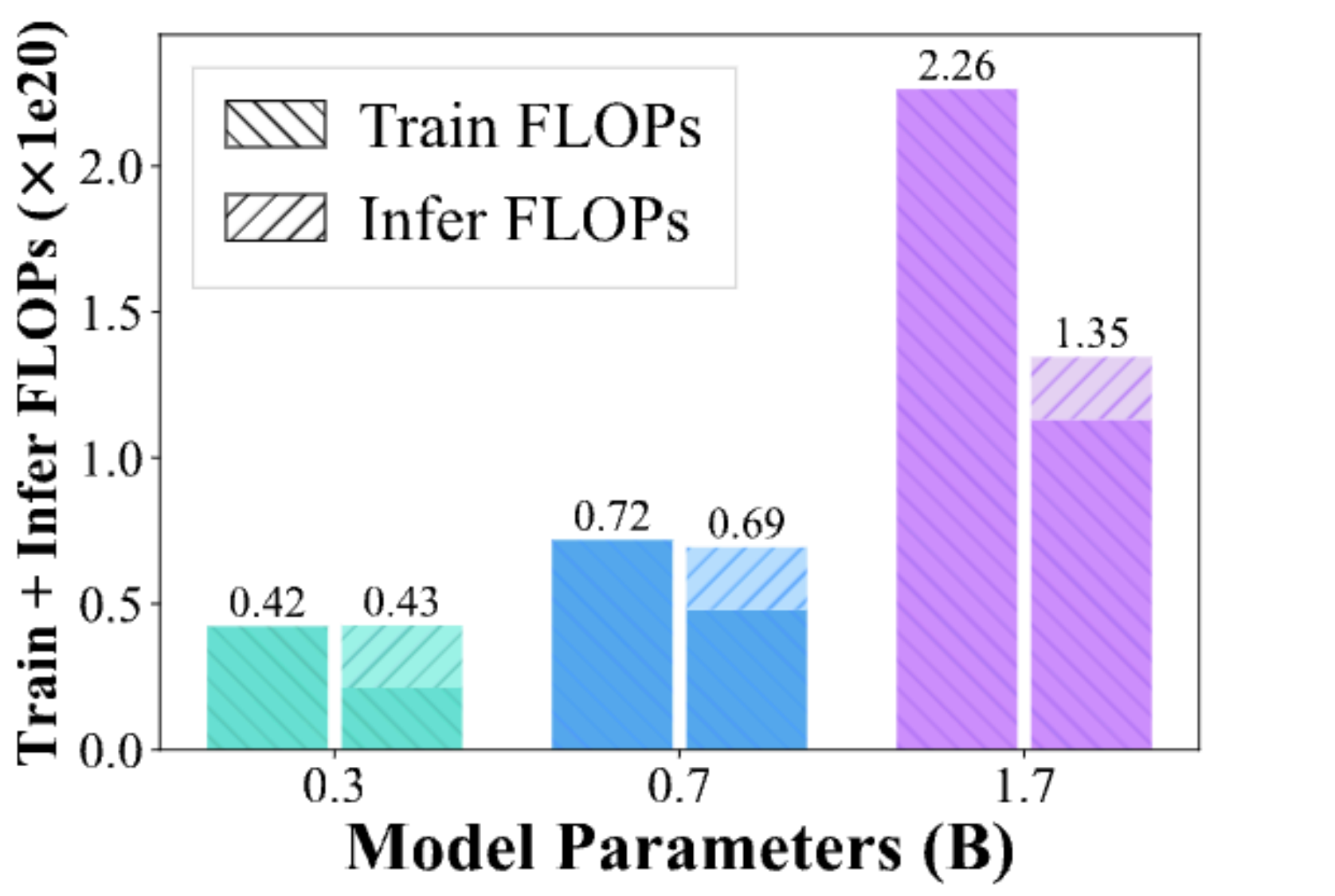


Figure 8: FLOPs comparison for comparable downstream performance with/without PROX refining: 0.3B(Avg.Perf = 40.5), 0.7B (41.6), and 1.7B (42.9).[2]