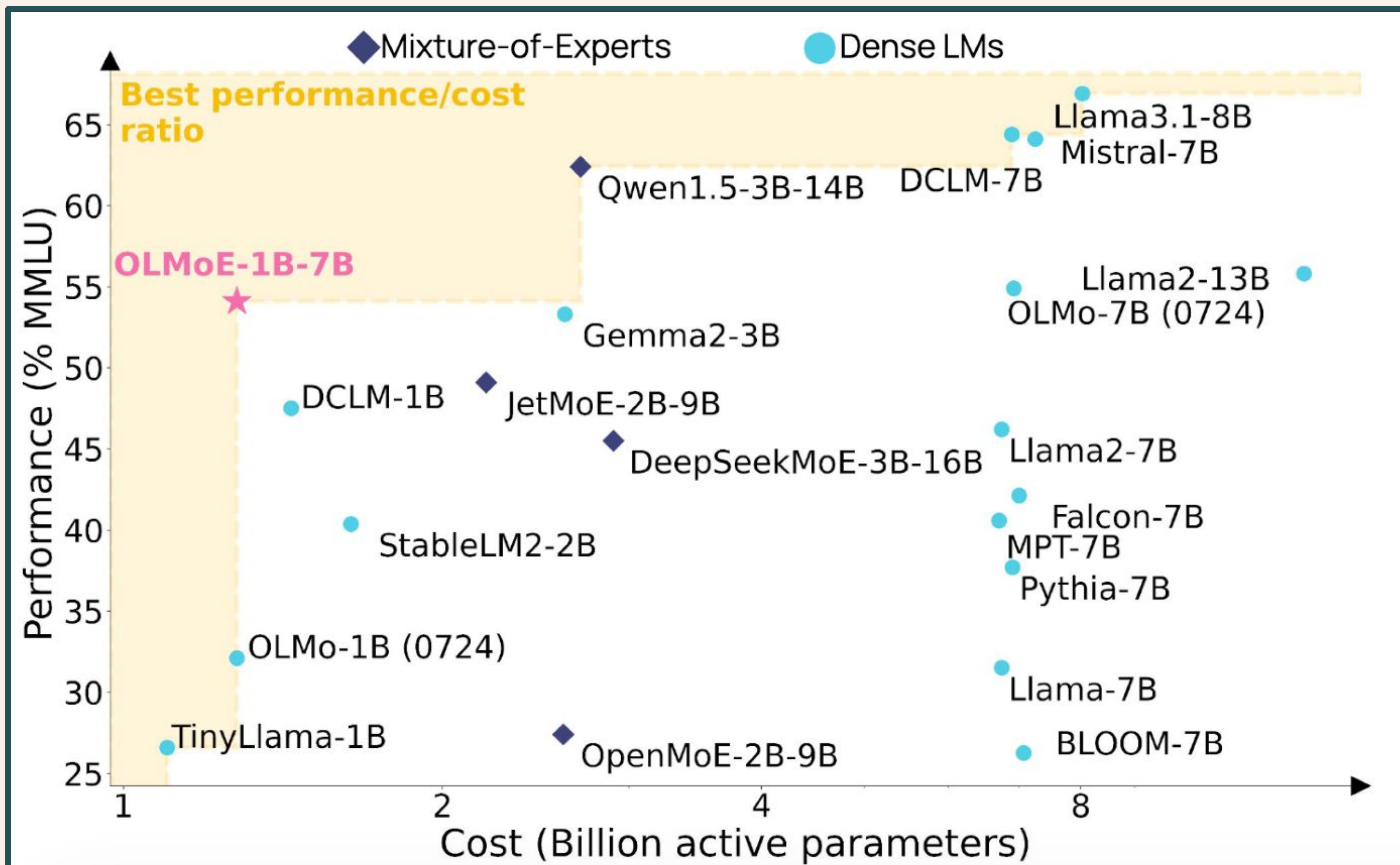


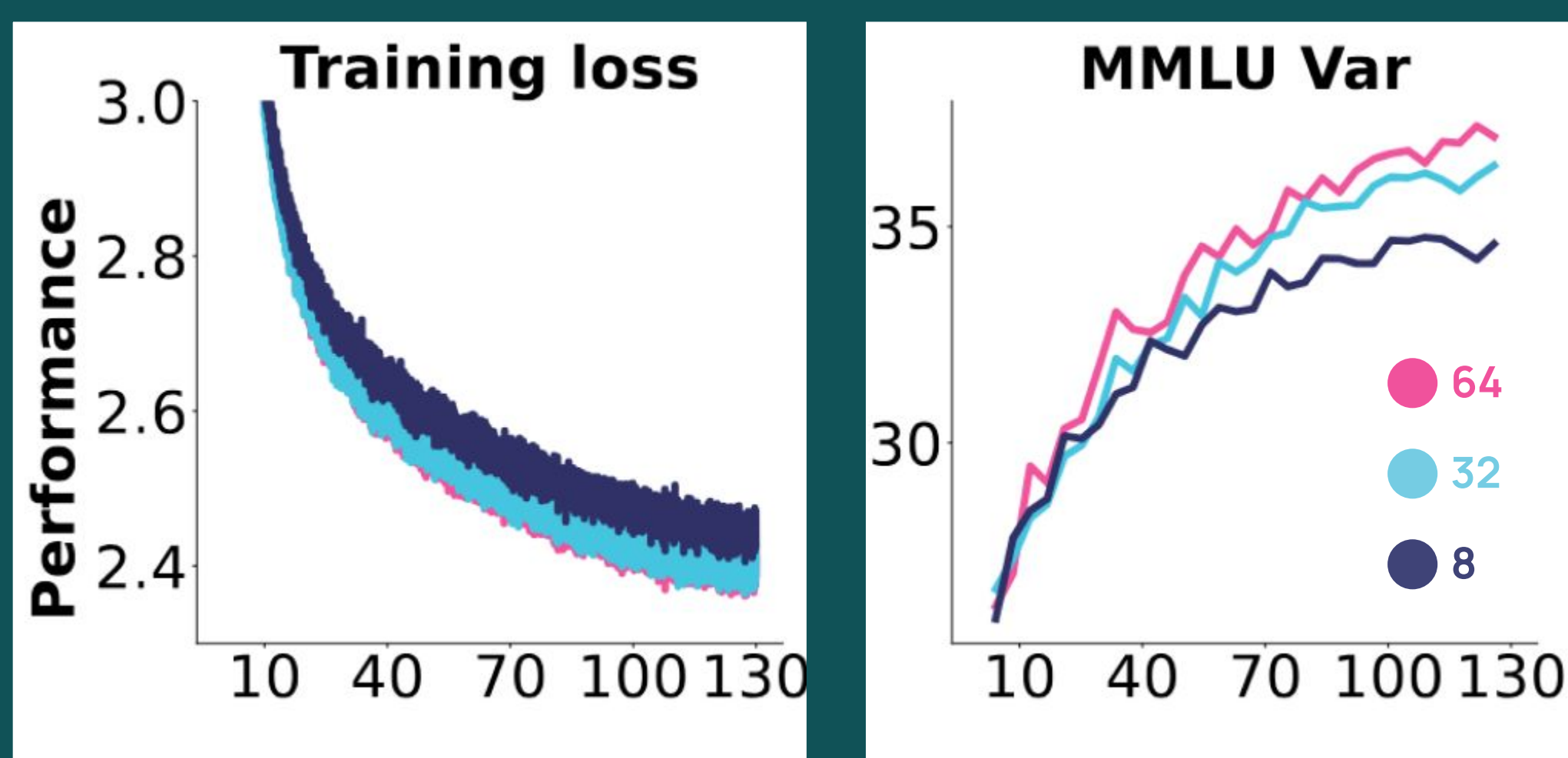


Open Mixture-of-Experts Language Models



How open are open MoEs?					
Name	Model	Data	Code	Logs	#ckpts
Grok-86B-314B	✓	✗	✗	✗	1
Mixtral-39B-141B	✓	✗	✗	✗	1
DBRX-36B-132B	!!	✗	✗	✗	1
Skywork-22B-146B	!!	✗	✗	✗	1
DeepSeekV2-21B-236B	!!	✗	✗	✗	1
Arctic-17B-480B	✓	!!	✗	✗	1
Qwen2-14B-57B	✓	✗	✗	✗	1
Mixtral-13B-47B	✓	✗	✗	✗	1
Jamba-12B-52B	✓	✗	✗	✗	1
DeepSeekMoE-3B-16B	!!	✗	✗	✗	1
Qwen1.5-3B-14B	!!	✗	✗	✗	1
OpenMoE-3B-9B	✓	✓	✓	✗	6
JetMoE-2B-9B	✓	!!	!!	✗	1
OLMoE-1B-7B	✓	✓	✓	✓	244

Number of Experts?

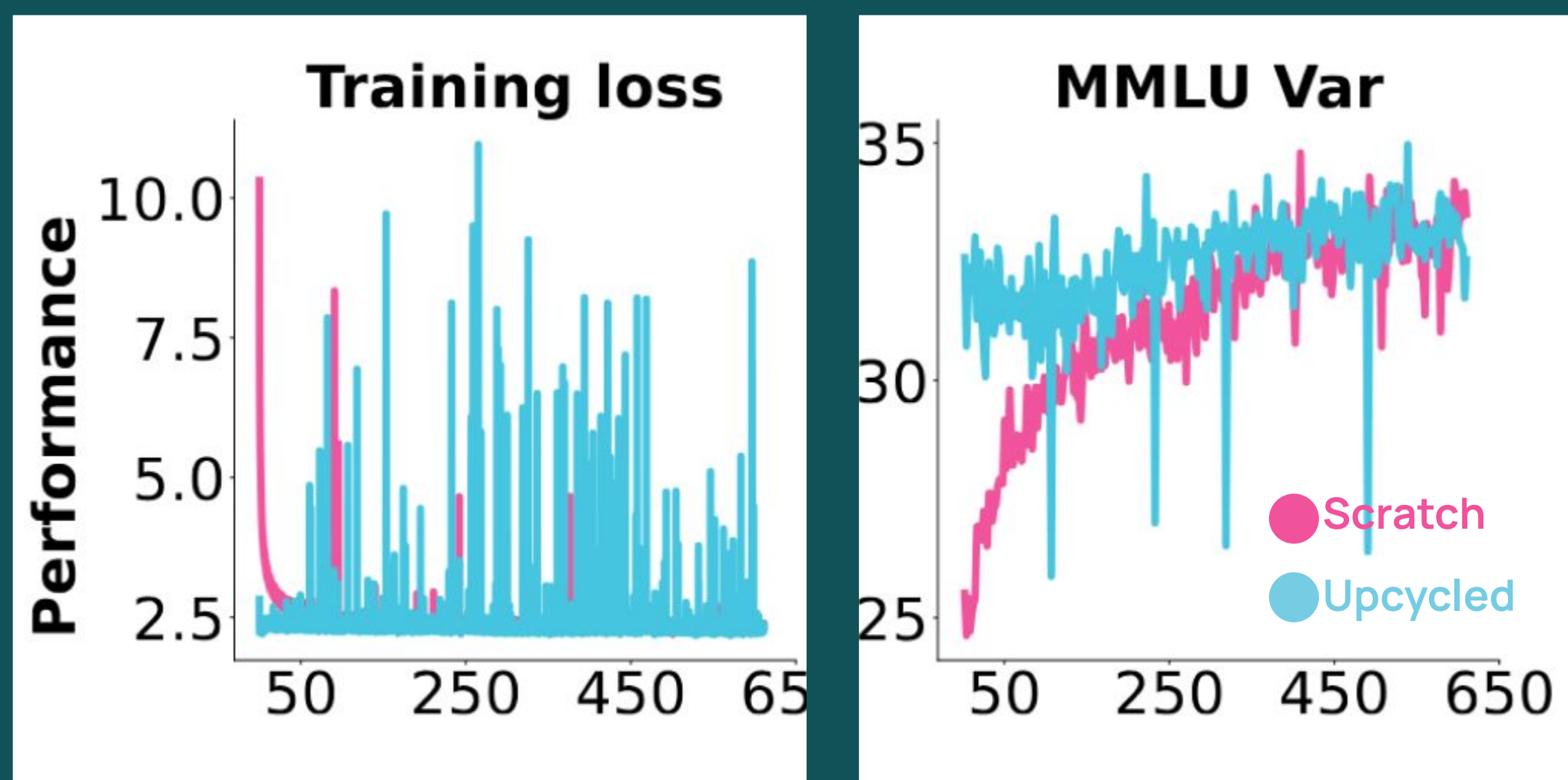


Domain Specialization?

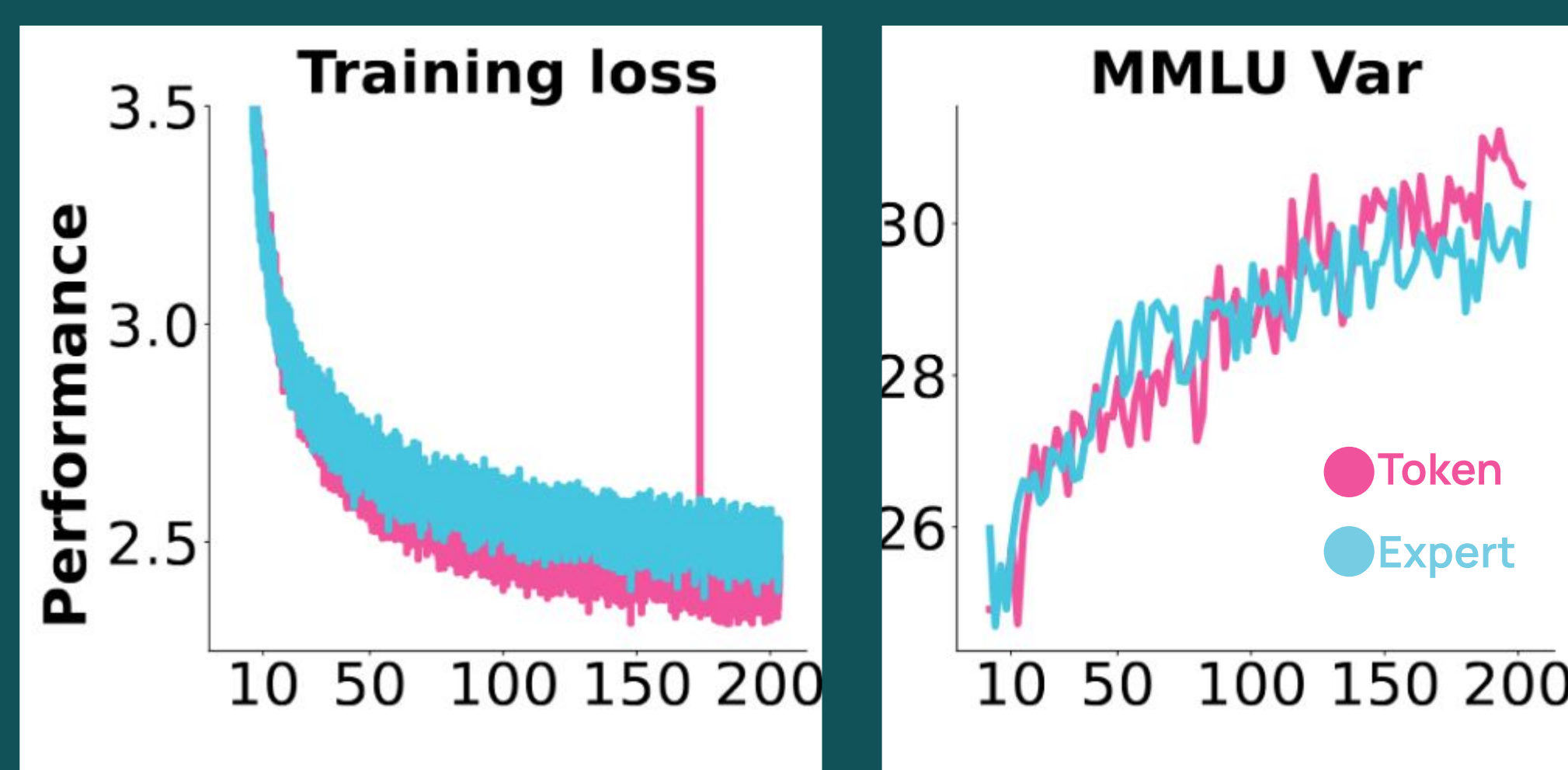


Runs
locally

Sparse Upcycling?



Token Choice vs Expert Choice?



110 tokens

per sec

on your

iPhone

Niklas Muennighoff,
Luca Soldaini, Dirk Groeneveld, Kyle Lo,
Jacob Morrison, Sewon Min, Weijia Shi,
Pete Walsh, Oyvind Tafjord, Nathan Lambert,
Yuling Gu, Shane Arora, Akshita Bhagia,
Dustin Schwenk, David Wadden, Alexander Wettig,
Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi,
Noah A. Smith, Pang Wei Koh, Amanpreet Singh,
Hannaneh Hajishirzi

