APT: Adaptive Pruning and Tuning Pretrained Language Models for Efficient Training and Inference

Bowen Zhao

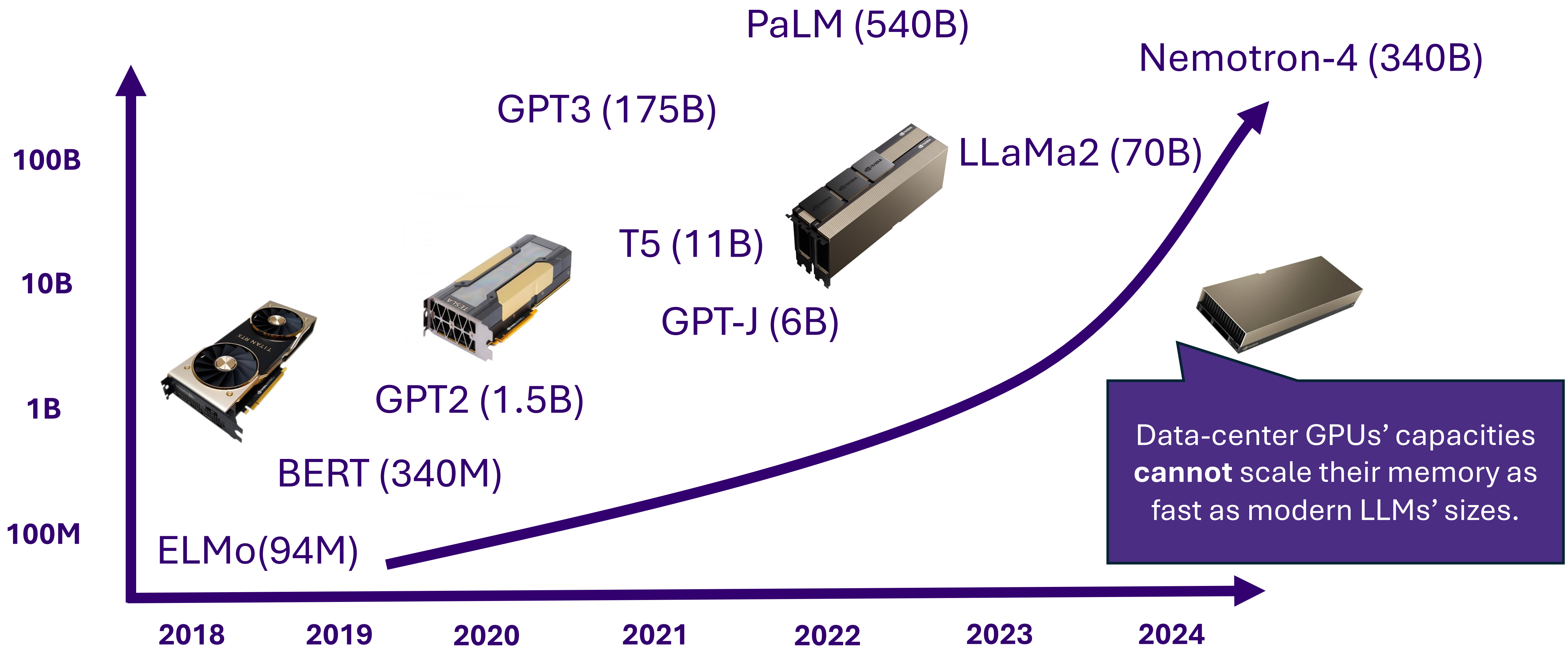
Hanna Hajishirzi

Qingqing Cao

UNIVERSITY of WASHINGTON

ICML International Conference On Machine Learning

Modern LLMs are becoming larger and computational costly



Current solutions to run LMs efficiently

Category	Method	Training Time	Training Mem.	Inference Time	Inference Mem.
PEFT	Adapter	+++	-	+	+
	LoRA	+++	-	=	=
Pruning	MvP	+++	+	-	-
	CoFi	+++	++	---	-
Combined	SPA	+++	+	---	-
	LRP	+++	-	---	-

Combined methods suffer from substantial end-task performance loss

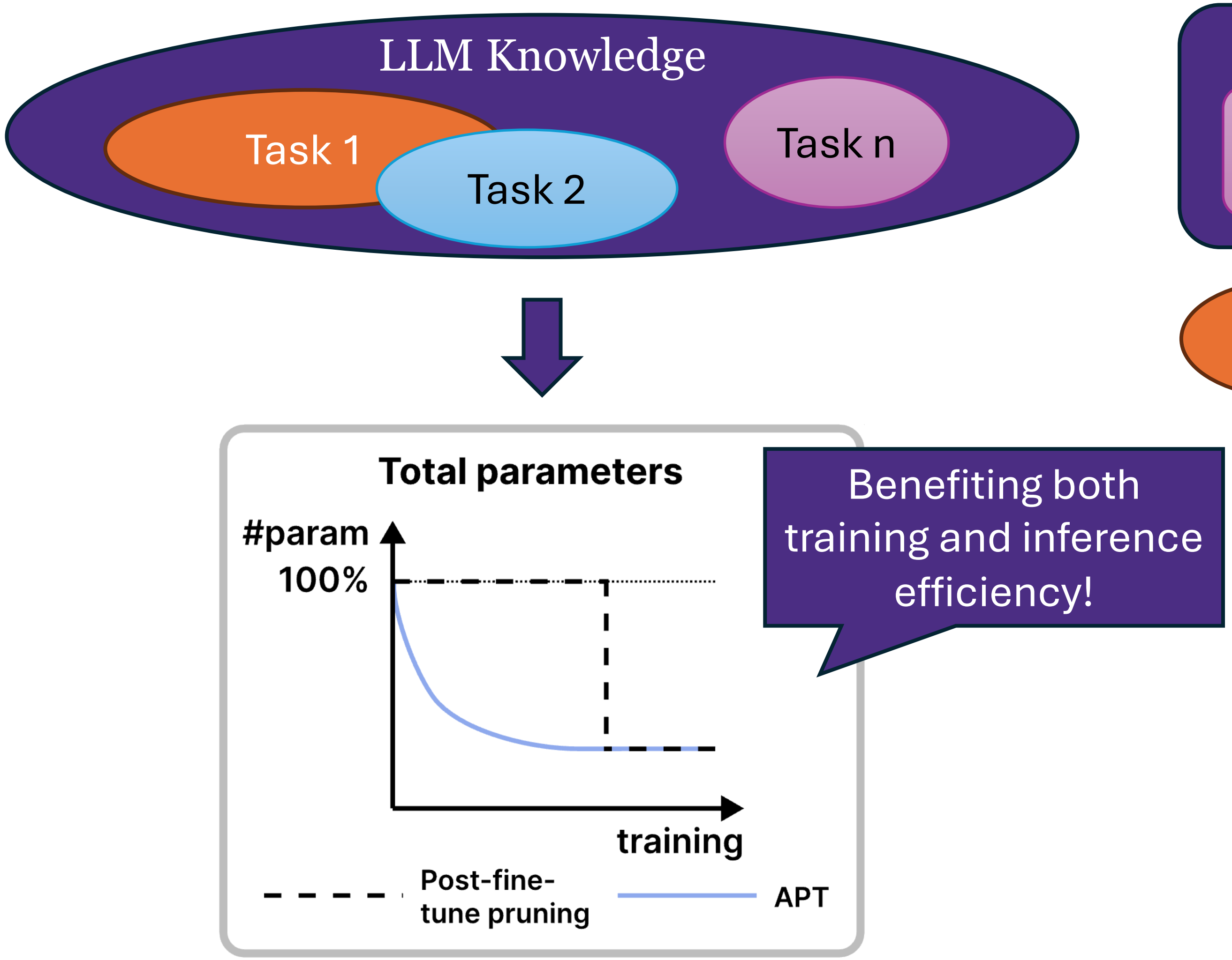
Existing efficient methods often requires longer training time to converge the LM

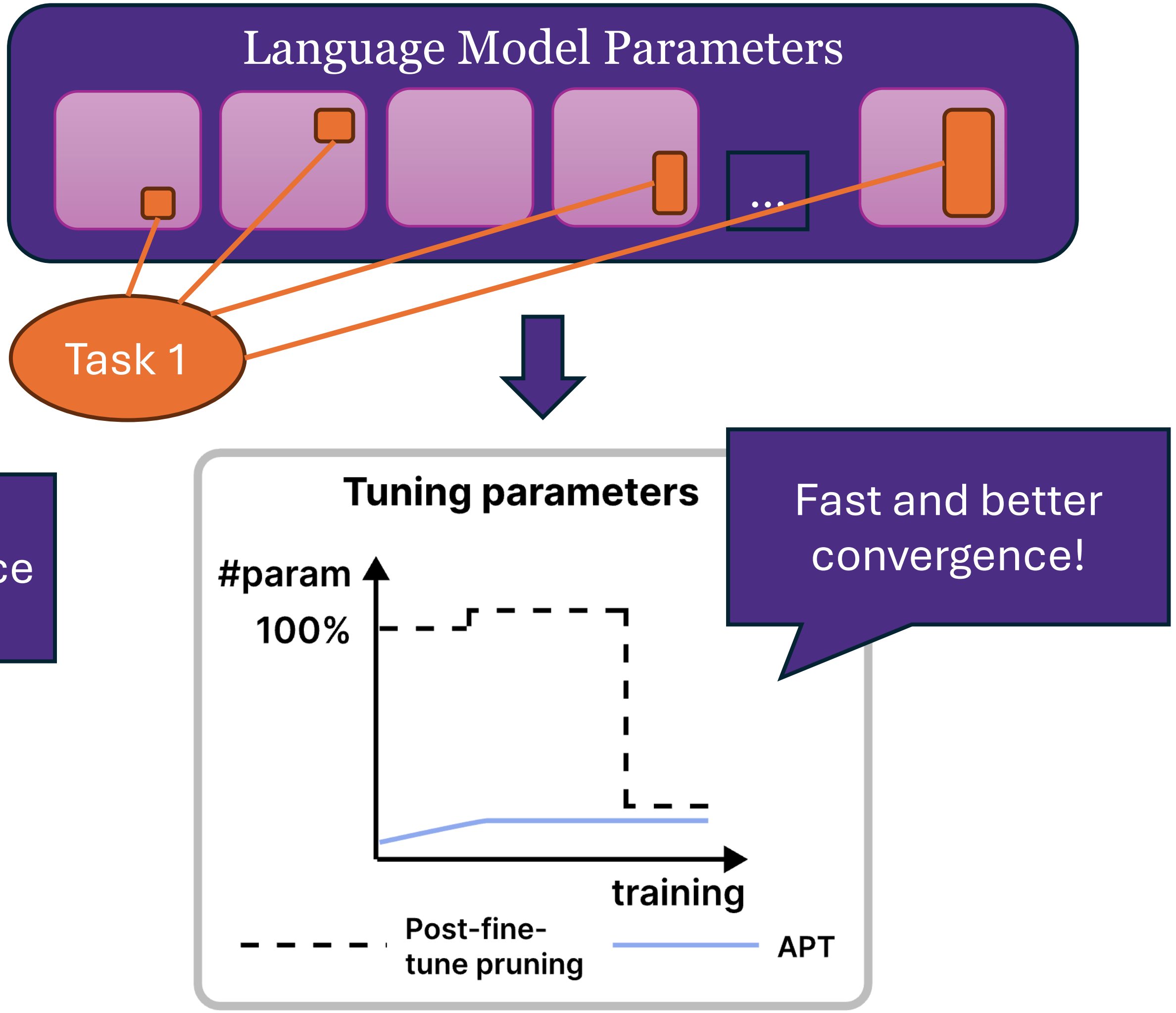
Pruning methods tend to cost extra training memory due to knowledge distillation

Improve training and inference efficiency

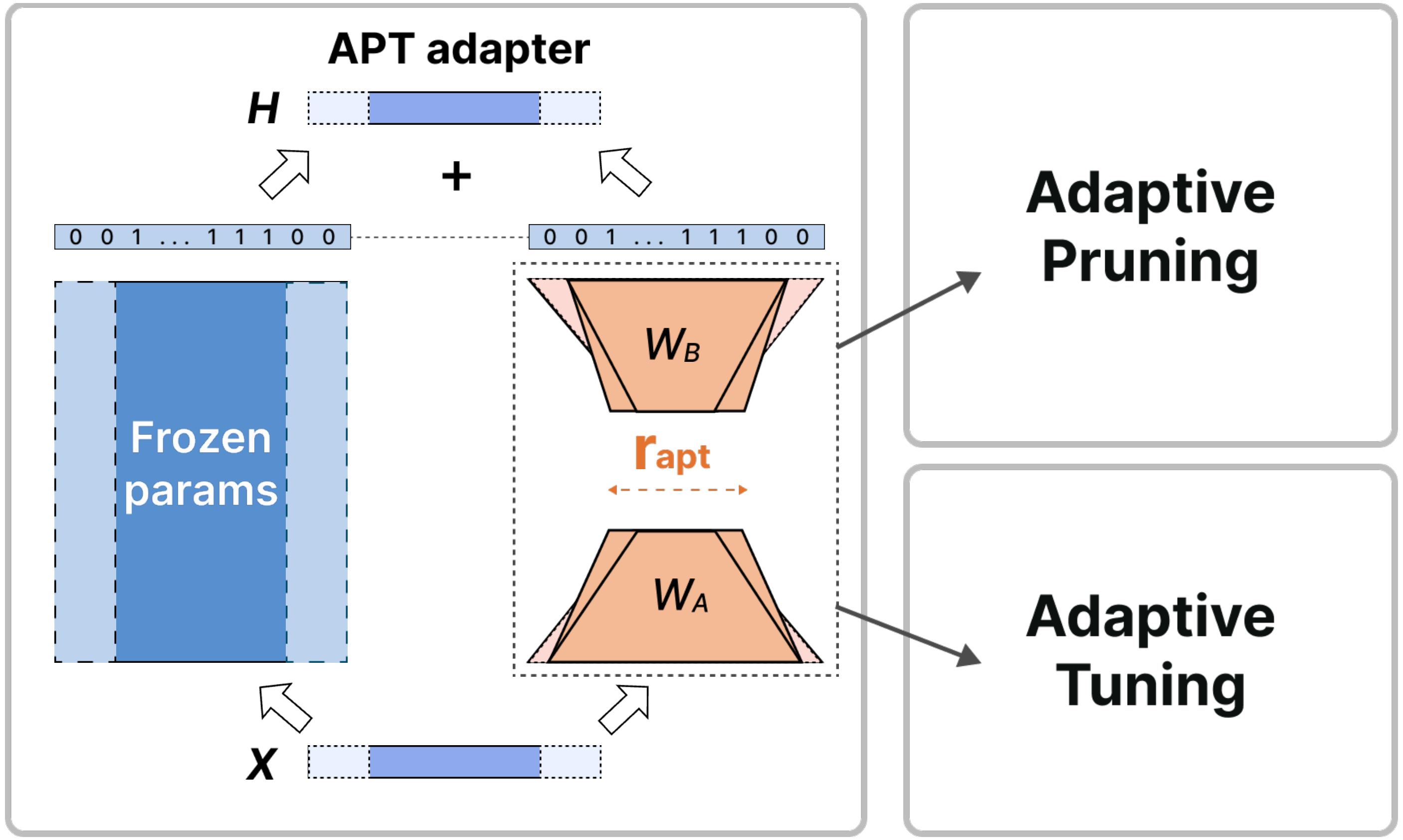
Question: can we combine the benefits of **PEFT and pruning** to improve both **training and inference efficiency** while maintaining **task performance**?

Intuitions for improving LM efficiency

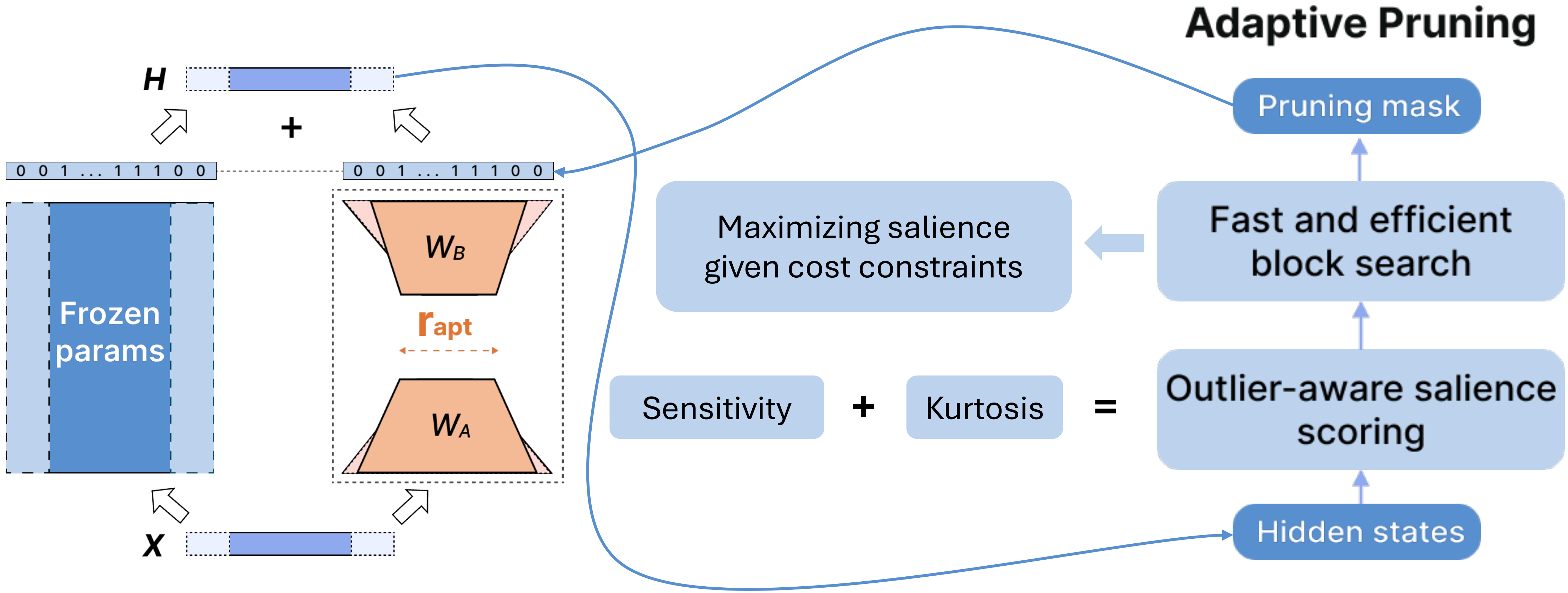




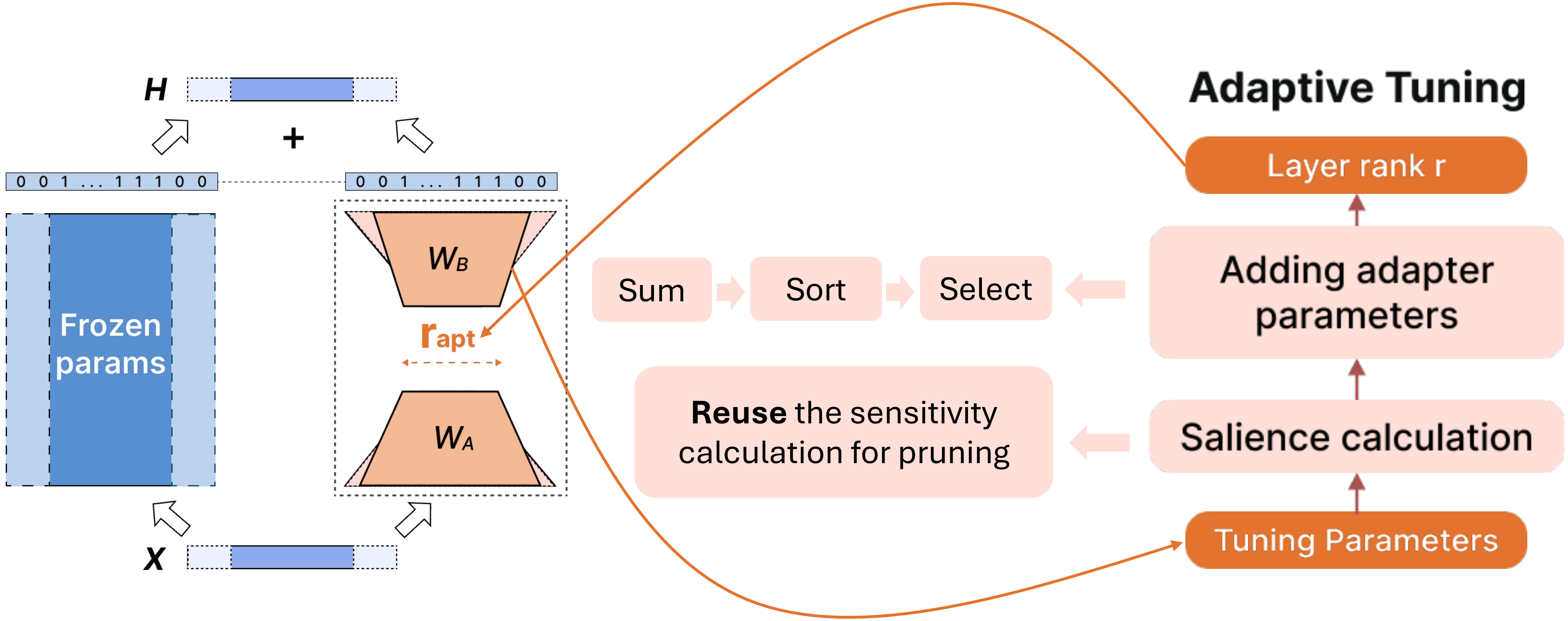
Our solution – APT: pruning & tuning **adaptively**



Low-cost adaptive pruning



Efficient adaptive tuning



Evaluation setup

- LM backbones and tasks:
  - Small-scale LMs:
    - BERT, RoBERTa: NLU tasks – GLUE, SQuAD
    - T5: NLU & NLG tasks – GLUE, CNN/DM
  - Large LMs:
    - LLaMa2 7B & 13B: standard few-shot tasks – ARC, HellaSwag, MMLU, TruthfulQA
- Metrics:
  - Task accuracy/F1/ROUGE score
  - Training efficiency: time to accuracy (seconds), training peak memory consumption (MB)
  - Inference efficiency: peak memory (MB), relative speedup

TTA: training time to a percentage of the baseline (finetuning) accuracy

Evaluation baselines

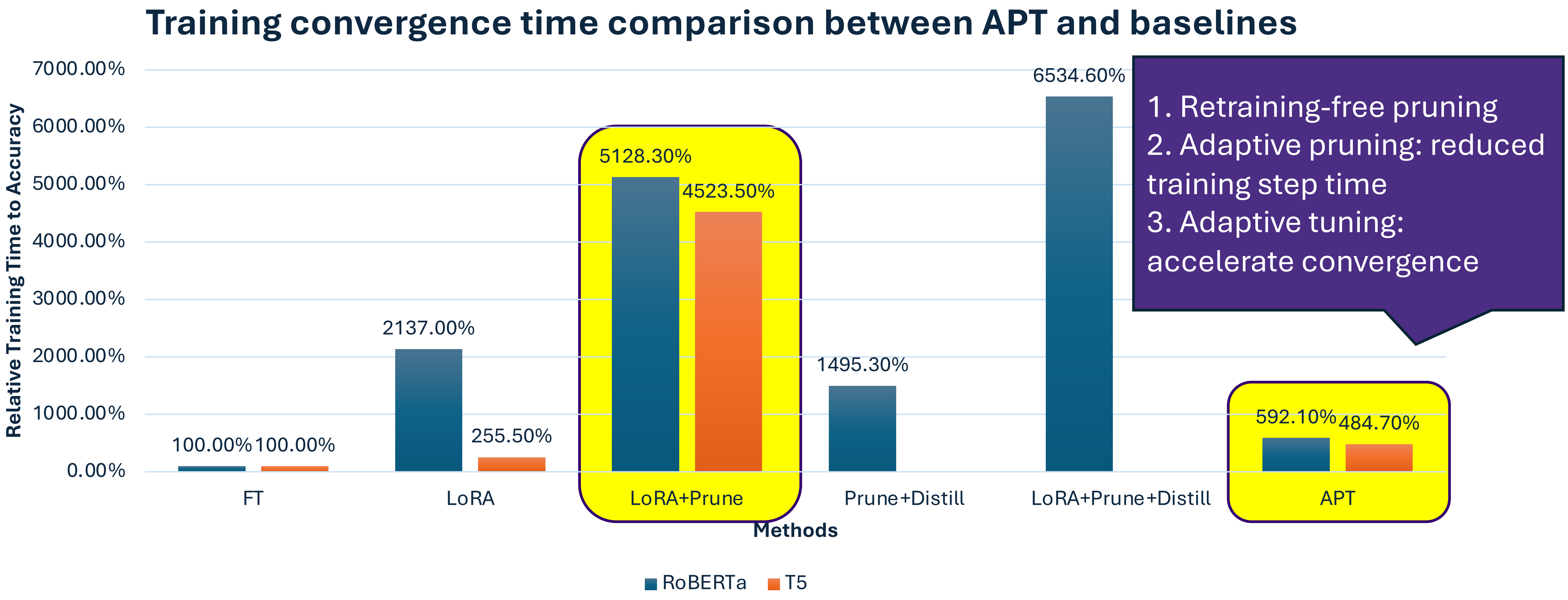
Direct baselines:

- Full-parameter finetuning
- LoRA

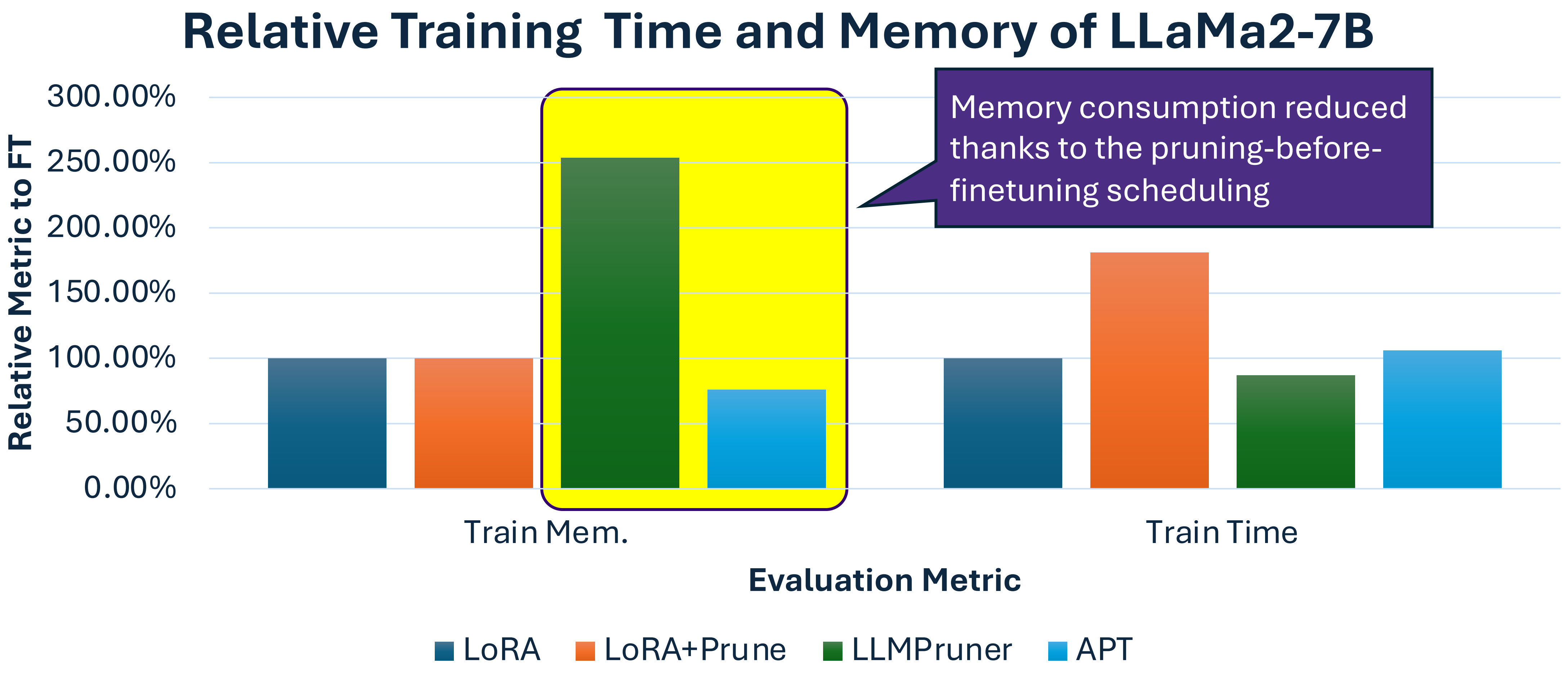
PEFT, pruning, and their combinations:

- LoRA+Prune: conducting post-training pruning (Mask-tuning; Kwon, et al., 2022) after LoRA-tuning
- Prune+Distill: structured pruning plus coarse-to-fine grained distillation (CoFi; Xia, et al., 2022)
- LoRA+Prune+Distill: using CoFi for pruning, but tuning LoRA only
- LLMPruner (Ma, et al., 2023): state-of-the-art structured pruning method on billion-level LLMs.

APT speeds up small LMs pruning 8x faster compared to LoRA+Prune baseline

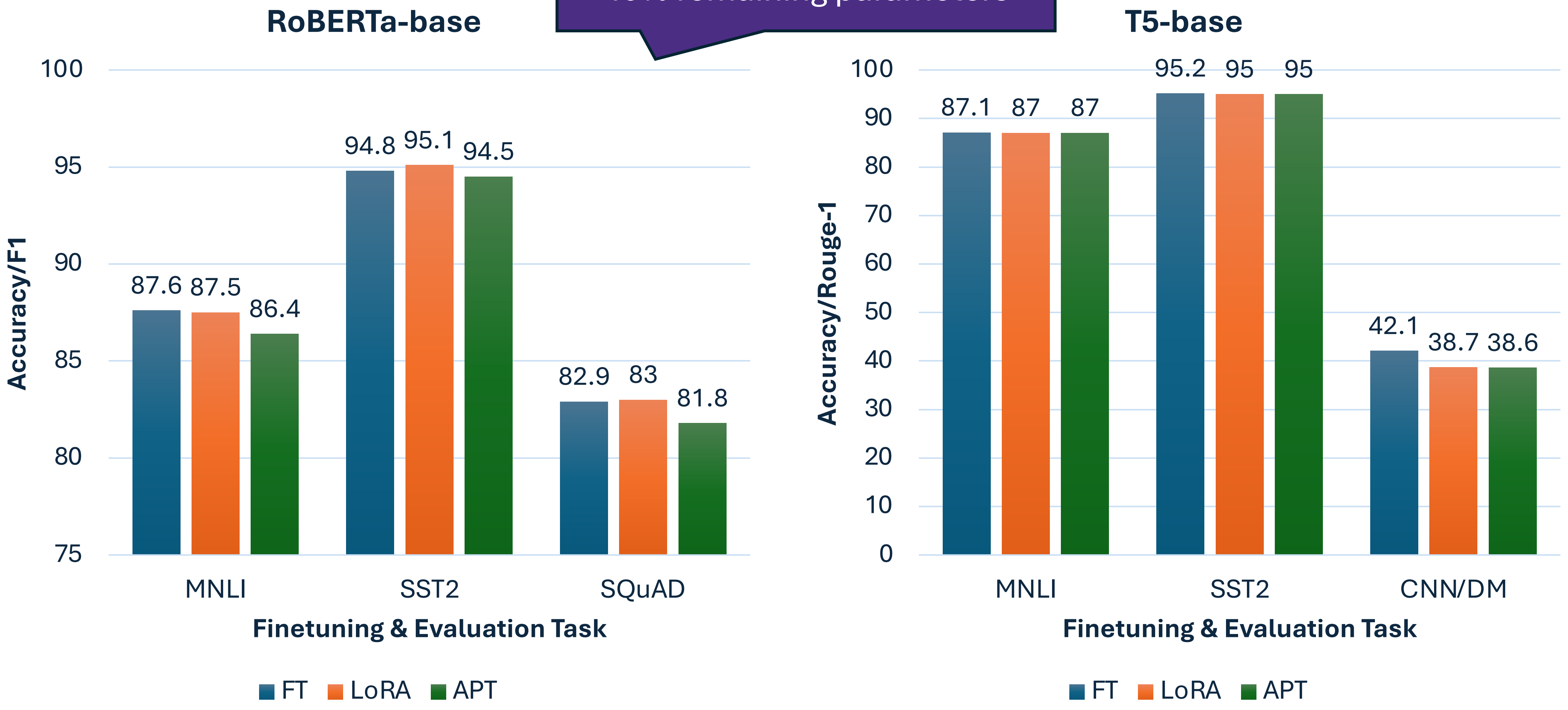


APT prunes LLMs with only 30% training memory consumption compared to LLMPruner

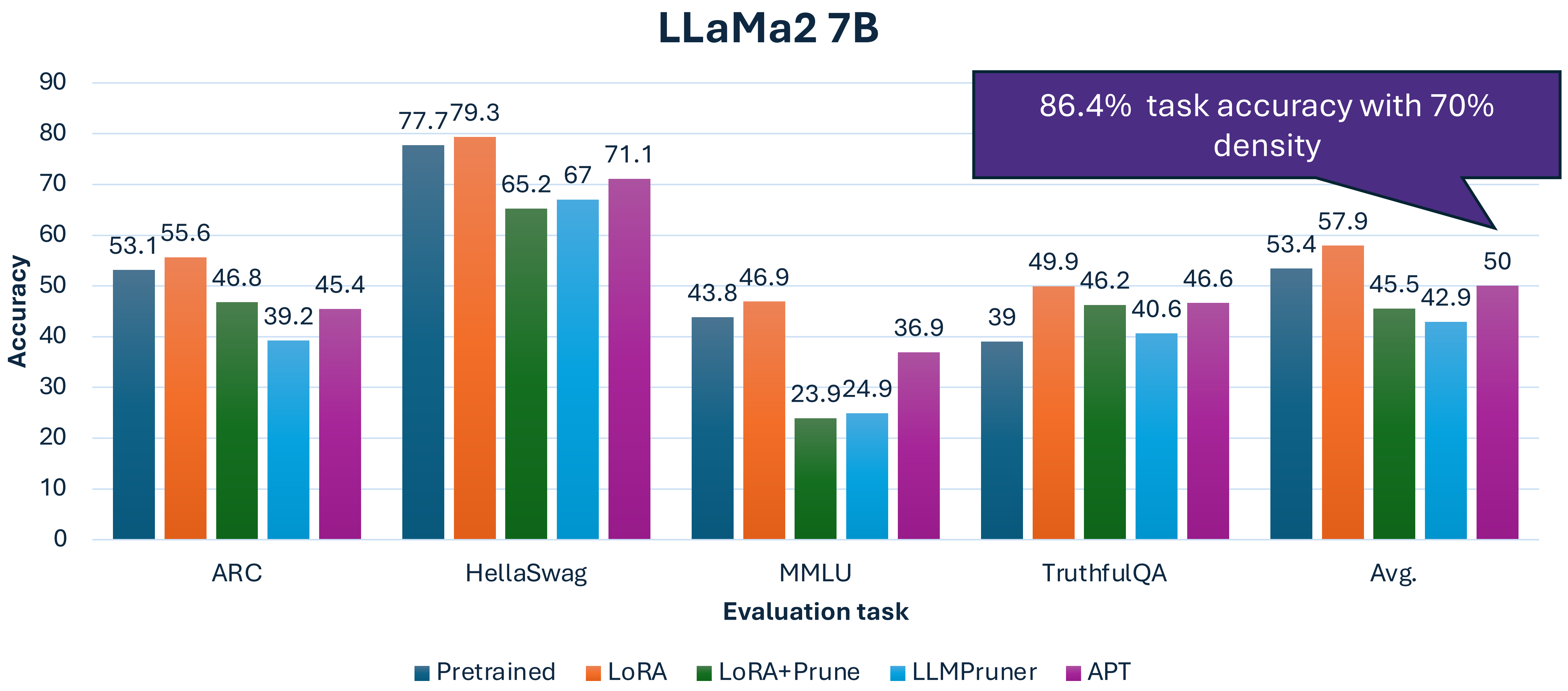


APT recovers task accuracy for small and large LMs

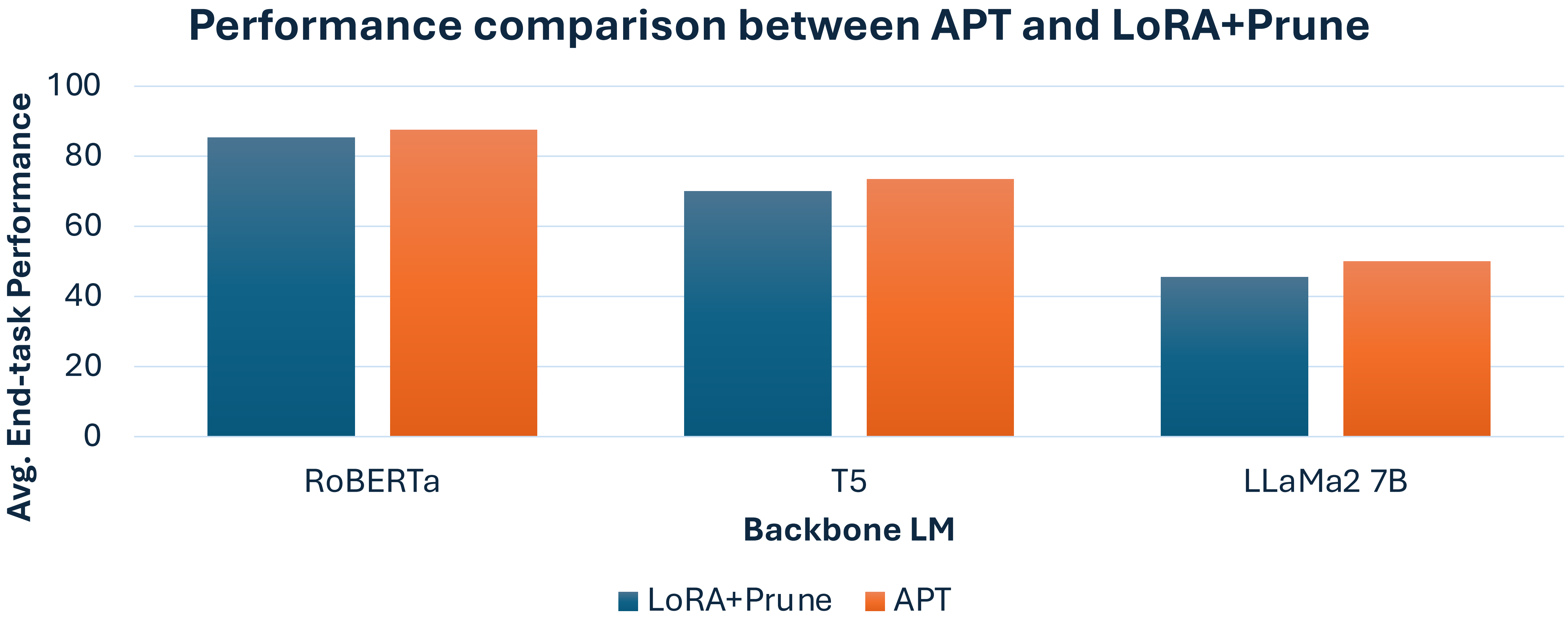
Up to 98% performance with 40% remaining parameters



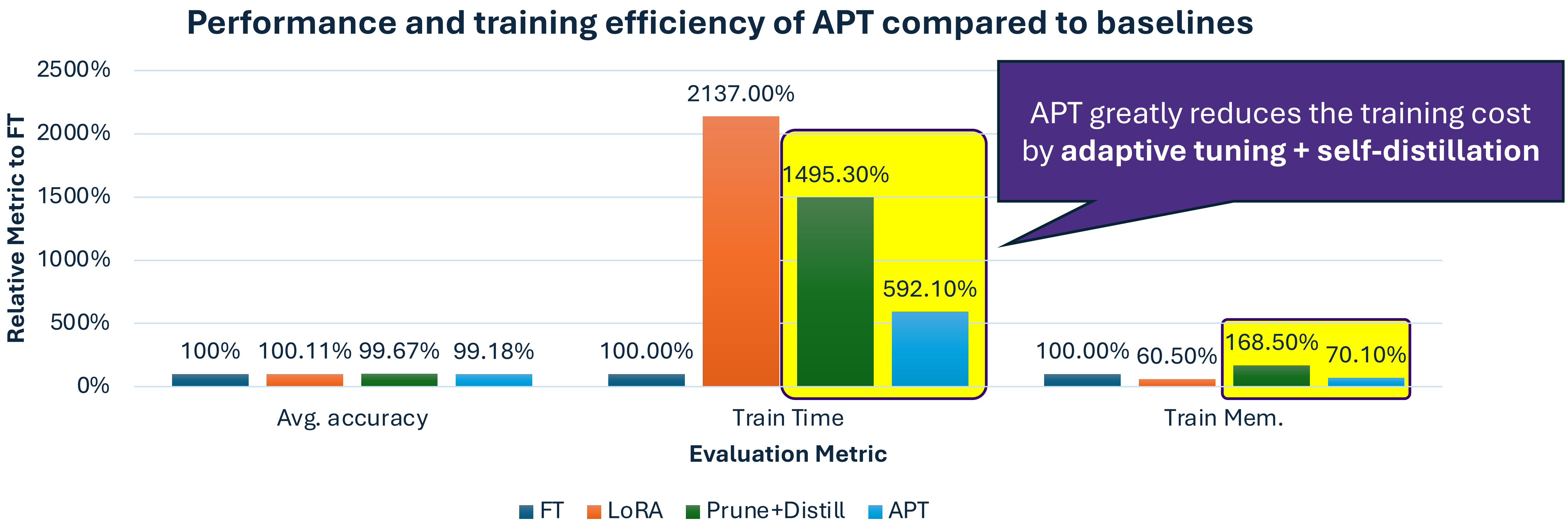
APT recovers task accuracy for small and large LMs



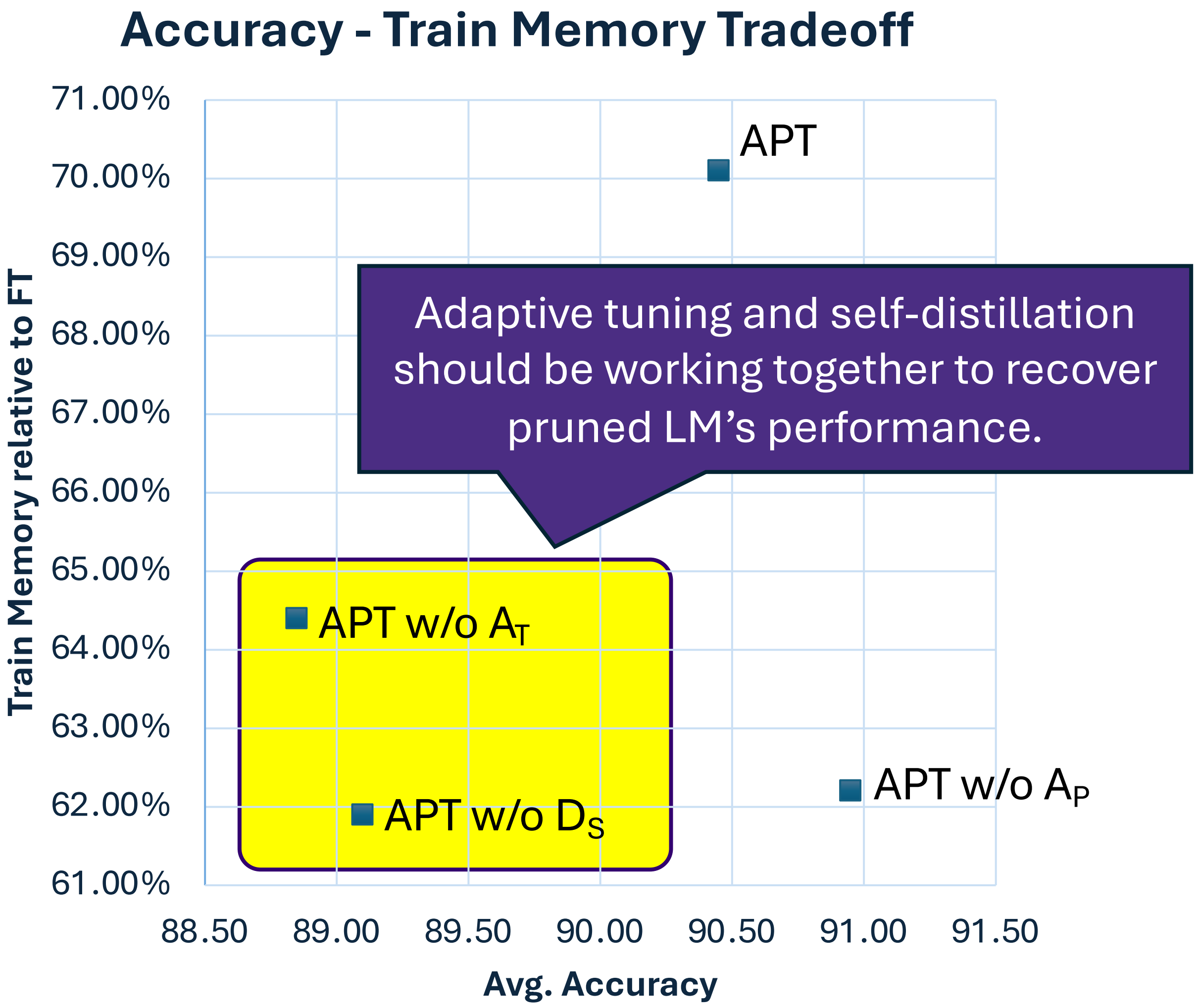
APT achieves 2.5%-9.9% higher task performance than the LoRA+Prune baseline

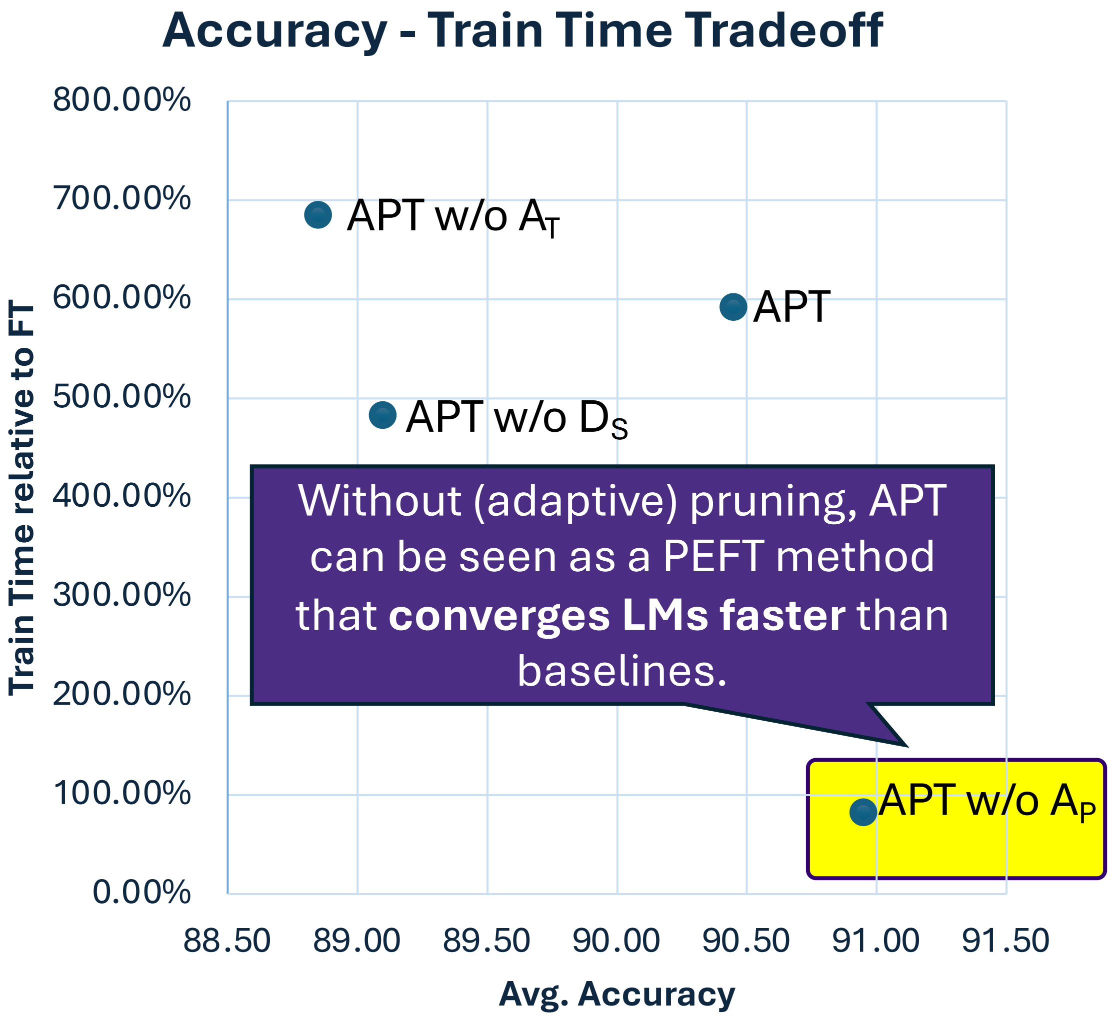


APT reaches on-par performance with the Prune+Distill baseline but trains **2.5× faster** and costs **only 41.6% memory**.

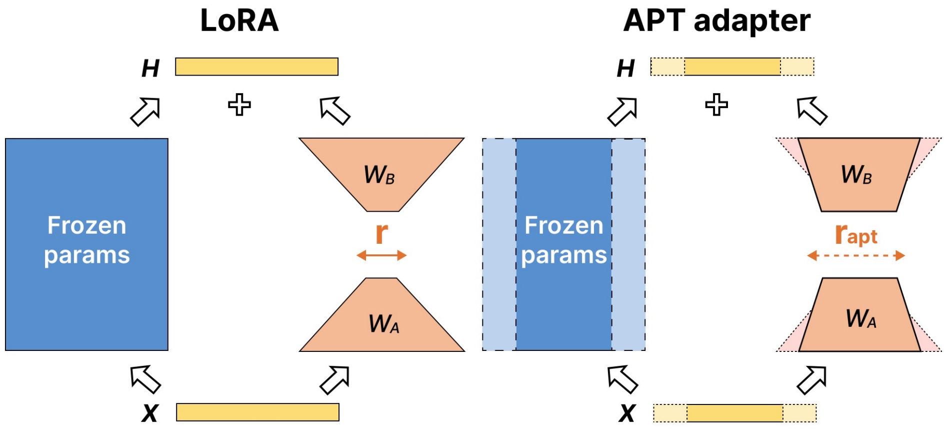


Each component in APT is effective





APT key takeaways and impact



- We propose APT, a new adaptive paradigm to prune and tune LMs effectively, targeting both training and inference efficiency via APT adapters.
- APT dynamically adjusts (adds/reduces) APT adapter input/output dimensions and the rank (r_{apt}), thus accelerating LM training convergence and also reducing inference costs.
- APT preserves LM task performance while speeding up small-scaled LMs' fine-tuning by up to 8× and reducing large LMs' training memory footprint by up to 70%.

Future work

- Even though APT proposes an efficient way to prune and tune LMs, it is definitely not always the optimal method for all LMs
- We hope that future work will focus on:
  - Adopting APT to a wider variety of PEFT backbones, e.g., prefix-tuning, prompt-tuning, parallel-adapter, VeRA, DoRA, etc.
  - Aiming at accurate, efficient, retraining-free pruning and distillation methods of large, billion-level LMs
  - Adapting APT with other efficient methods together for further inference efficiency gains, such as quantization, MoEfication, etc.

APT: Adaptive Pruning and Tuning Pretrained Language Models for Efficient Training and Inference

Bowen Zhao

Hanna Hajishirzi

Qingqing Cao

WeChat

Email: roimgame@gmail.com
X (Twitter): @BowenROIM
LinkedIn: bowen-zhao-760b83225

UNIVERSITY of WASHINGTON

ICML
International Conference On Machine Learning