Minitron-SSM: Efficient Hybrid Language Model Compression through Group-Aware SSM Pruning

NeurIPS 2025

Ali Taghibakhshi*, Sharath Turuvekere Sreenivas*, Saurav Muralidharan*, Marcin Chochowski*, Yashasw Karnati*, Raviraj Joshi, Daniel Korzekwa, Mostofa Patwary, Mohammad Shoeybi, Jan Kautz, Bryan Catanzaro, Ashwath Aithal, Nima Tajbakhsh, Pavlo Molchanov

Introduction

Training LLM Model Families

- Model providers often train a family of LLMs, where each model targets a specific deployment scale/size

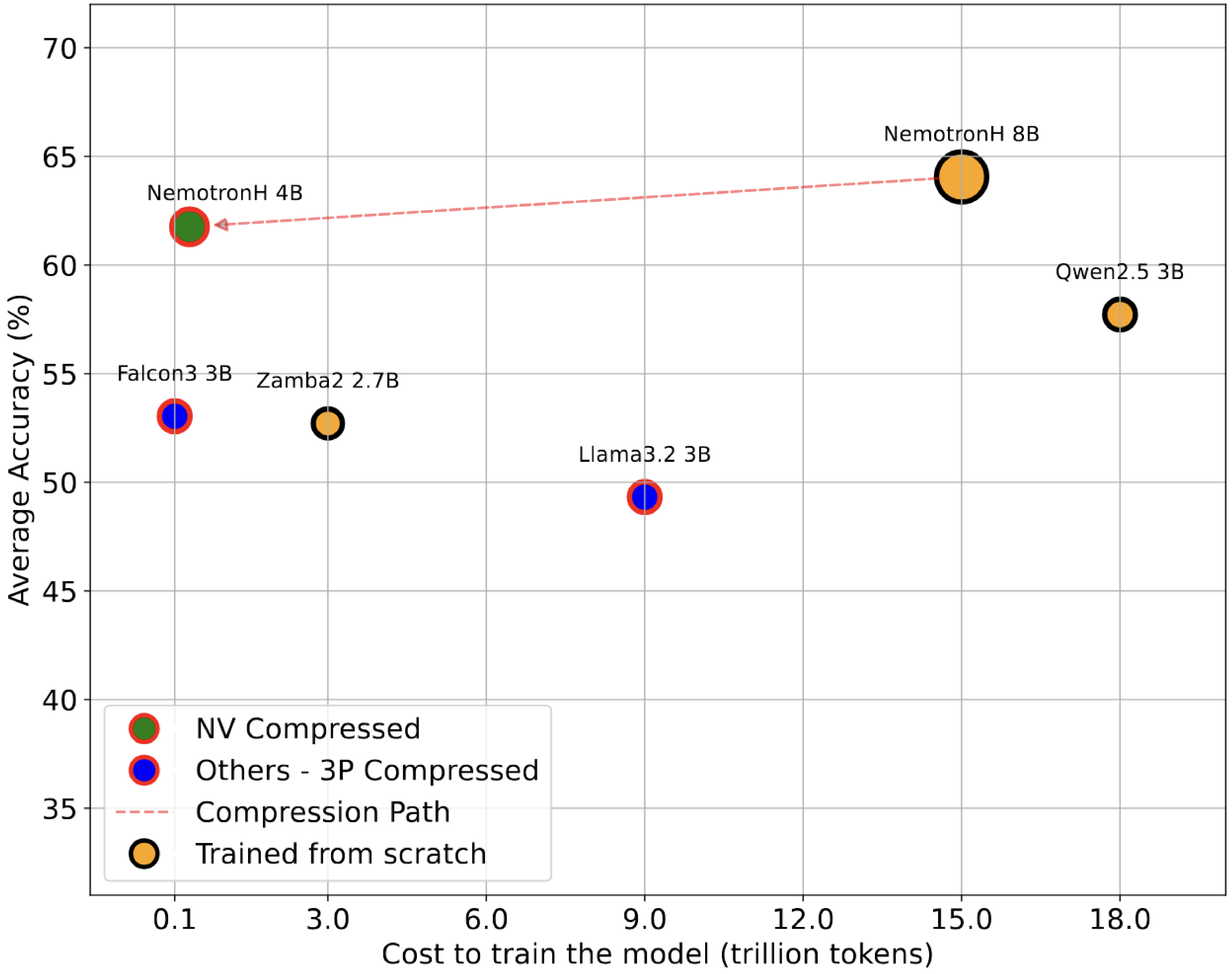
MISTRAL AI_ 7B, 8x7B, 8x22B, Small, Medium, Large

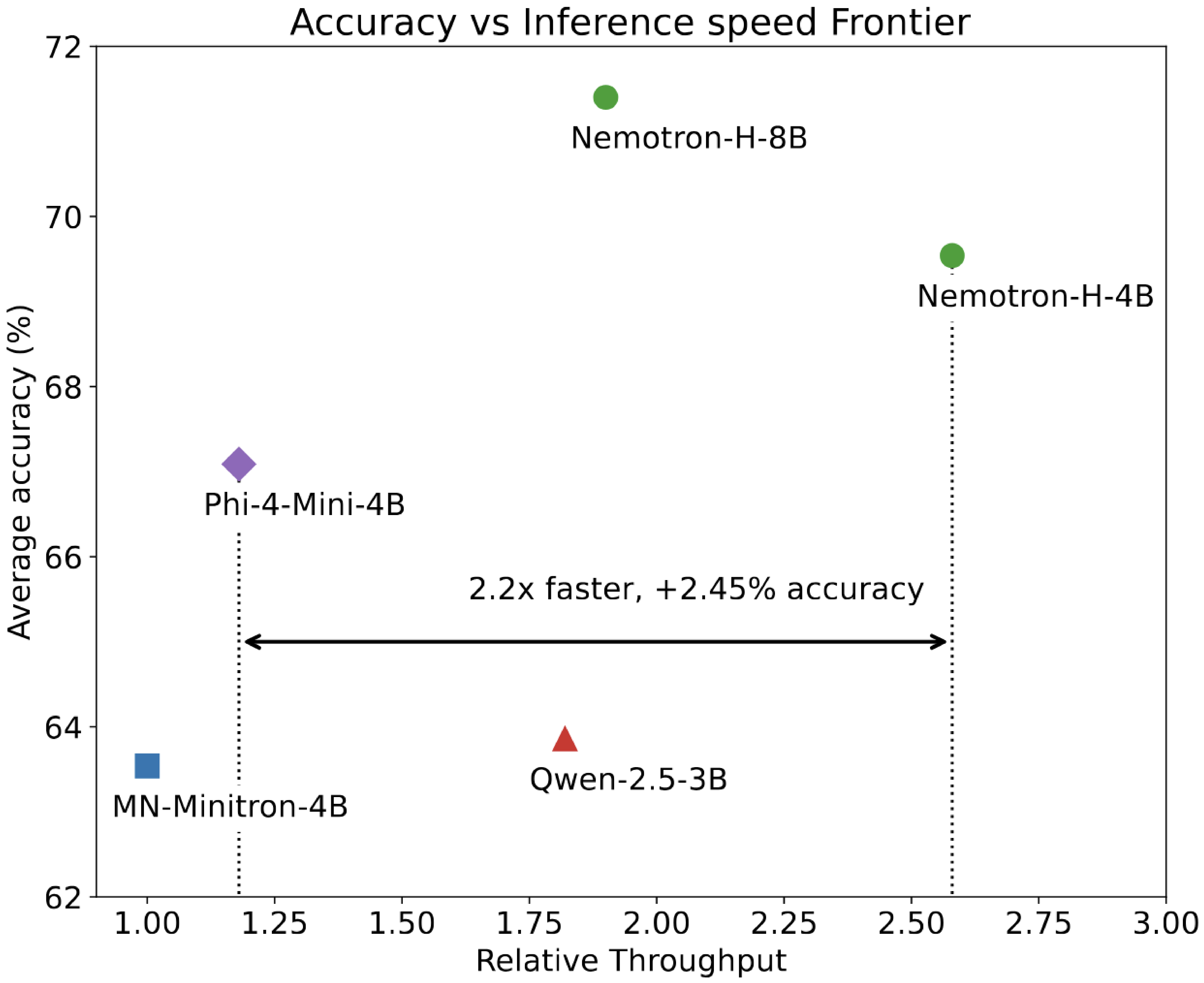
Meta LLaMa-3.1 8B, 70B, 405B

- Each model in the family is trained from scratch – expensive in compute, data, memory, etc.

“Can we train one big model, and obtain smaller, more accurate models from it through a combination of weight pruning and retraining, while only using a small fraction of the original training data?”

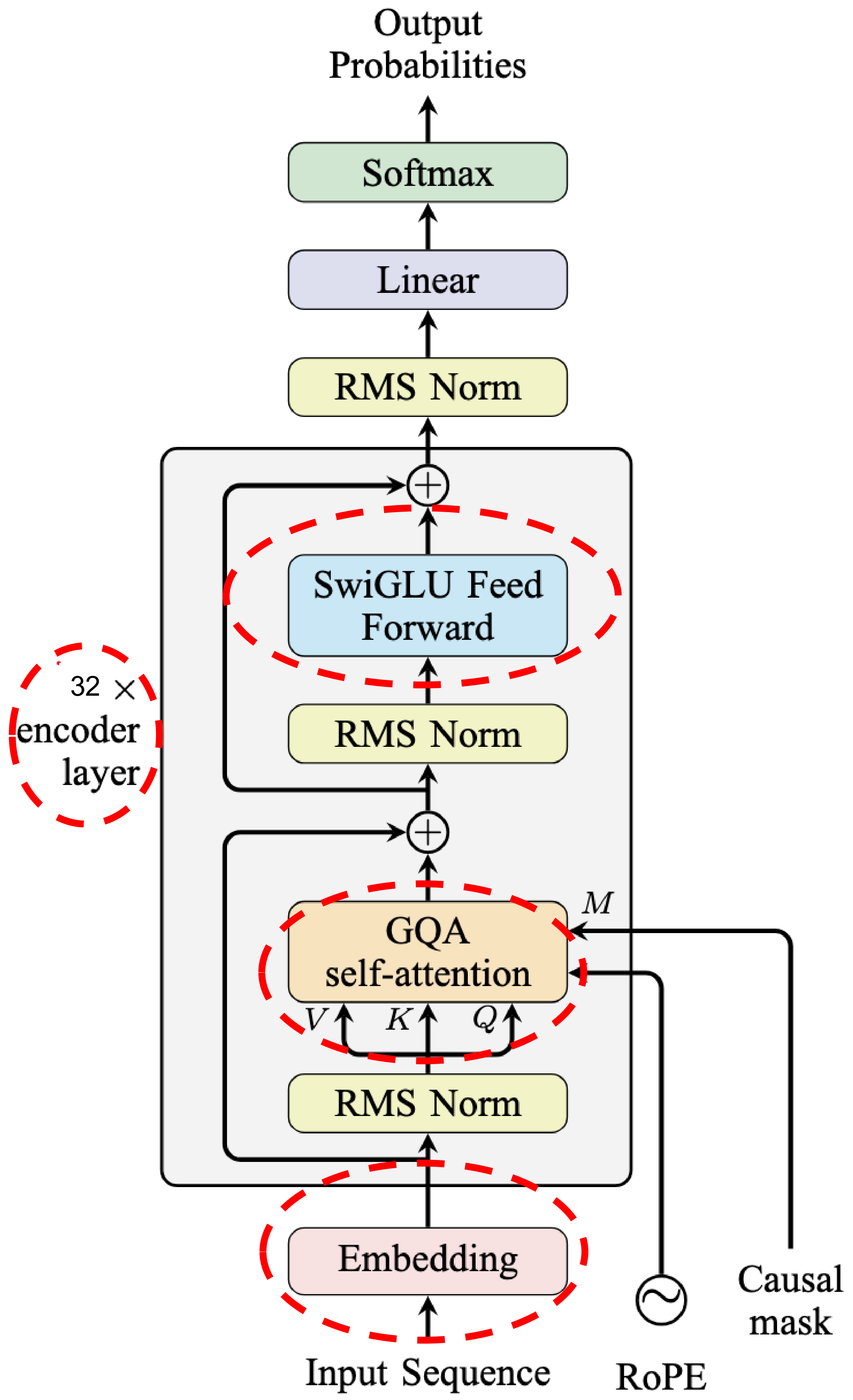
Results - 4B SOTA Accuracy and Perf





Method: What can be pruned?

Identify prunable elements (row/cols but keeping the model "working")



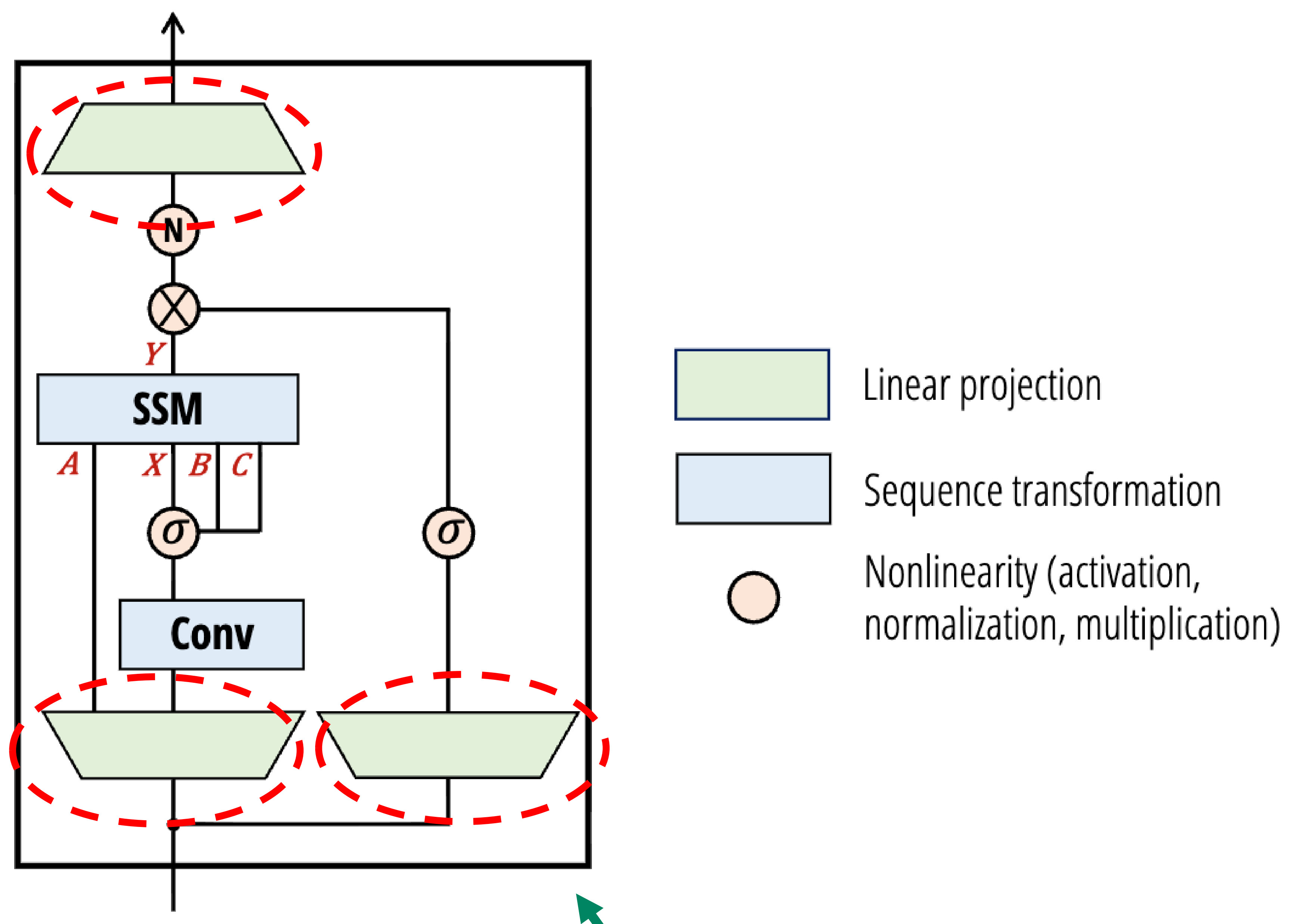
Classical GPT style transformer

- N layers
- MLP inner dimension
- Query heads
- Hidden size (embedding dimension)

Minitron

Method: What can be pruned?

Hybrid models introduce Mamba2 layer

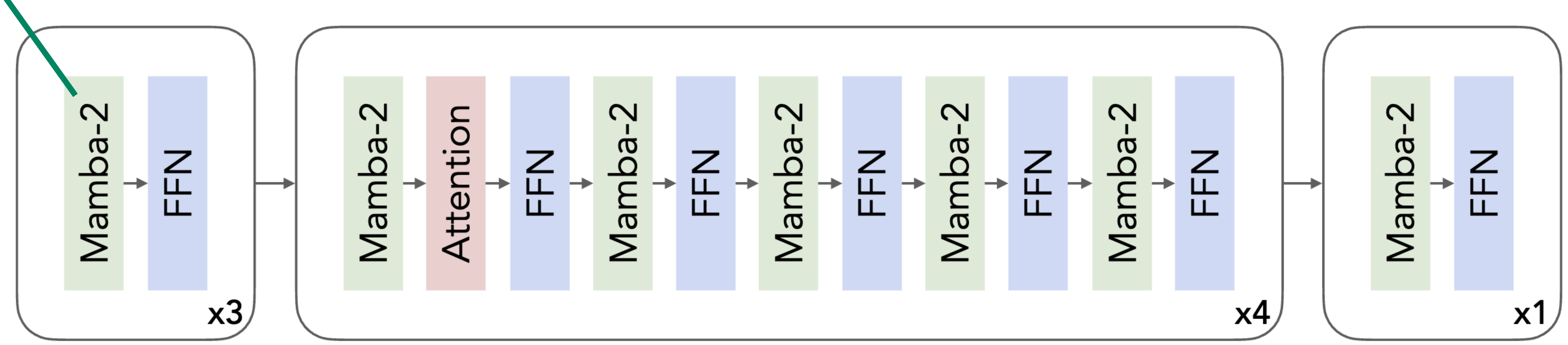


Hybrid model (transformer mixed with Mamba-2)

- N layers
- MLP inner dimension
- Query heads
- Hidden size (embedding dimension)
- Mamba heads / heads channels
  - Additional group-aware head permutation constraints !

Minitron-SSM

Nemotron-H-8B

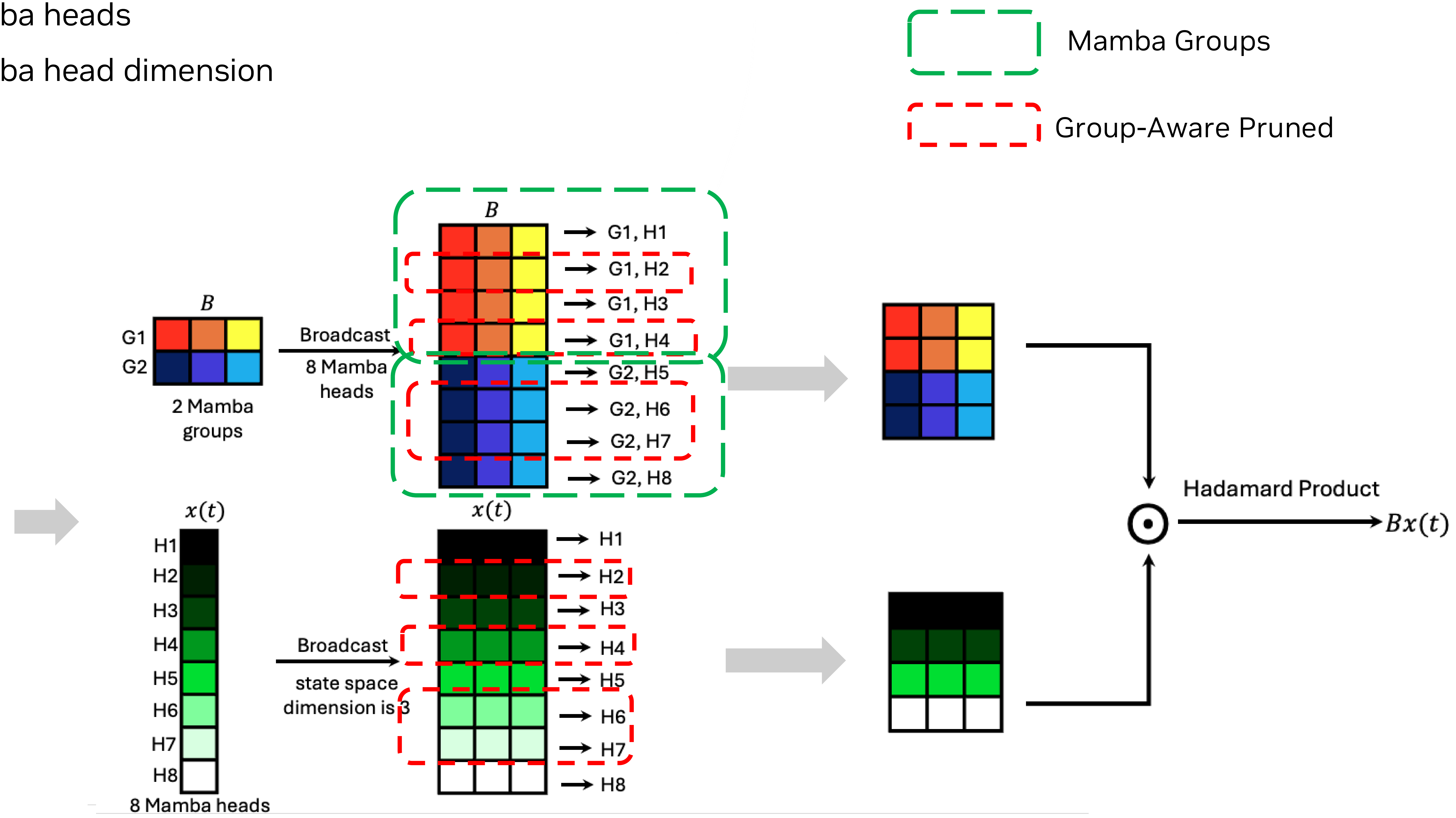


Method: Mamba2 Pruning

Pruning Mamba has constraints

- Group-aware SSM pruning
  - Arises from SSM implementation
    - Pruning Mamba heads
    - Pruning Mamba head dimension

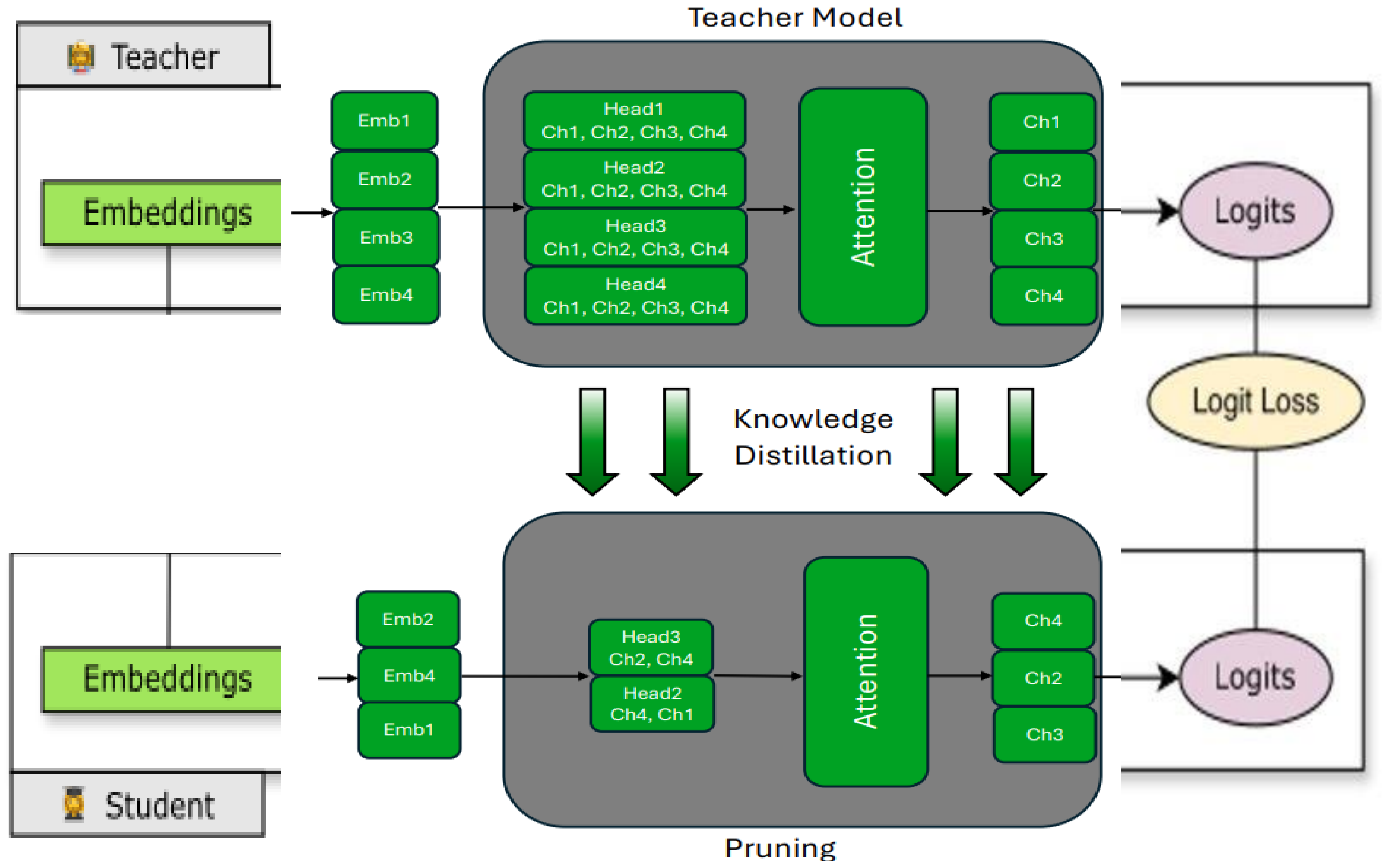
$$h_t = A_t h_{t-1} + B_t x_t$$
$$y_t = C_t^\top h_t + D_t x_t$$



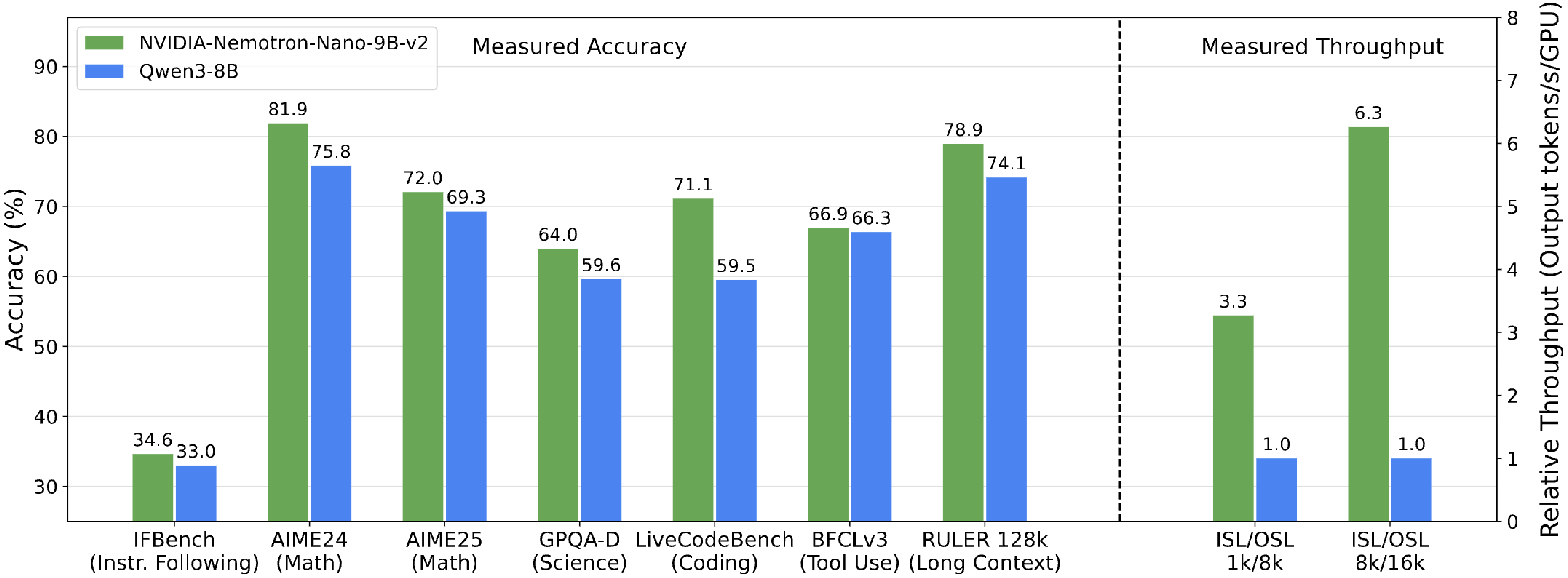
Method: How to recover?

Accuracy Recovery with Knowledge Distillation (KD)

- Knowledge distillation outperforms Cross Entropy fine-tuning
  - distilling knowledge from the original model to the pruned model
  - various loss during (logits loss only performs best)



Minitron-SSM was used for Nemotron-NanoV2



Minitron-SSM Resources

Poster Session: Wed 3 Dec 4:30 p.m. — 7:30 p.m. PST

NeurIPS Poster Page
Minitron Website
HuggingFace Base and Instruct Models

Thanks