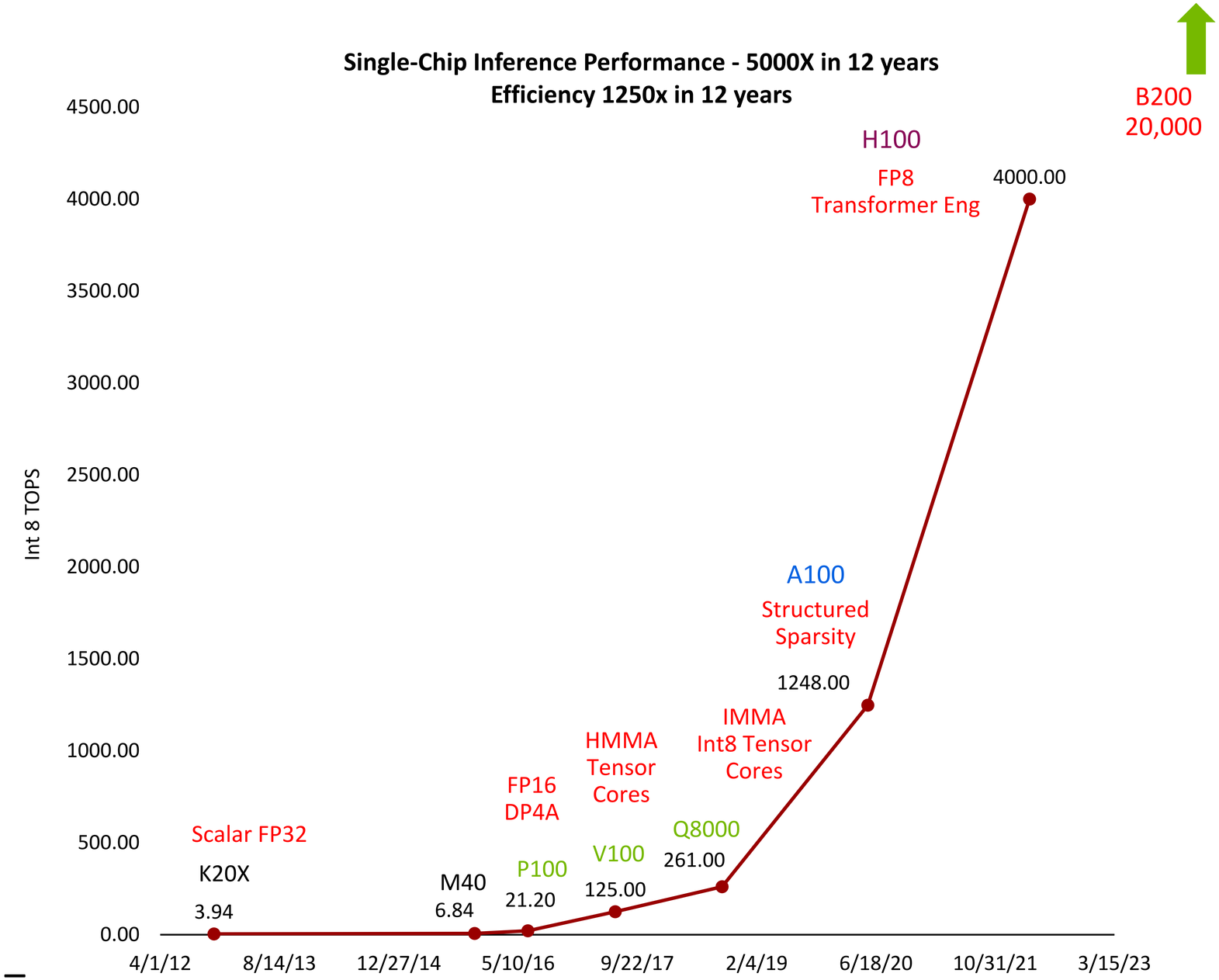
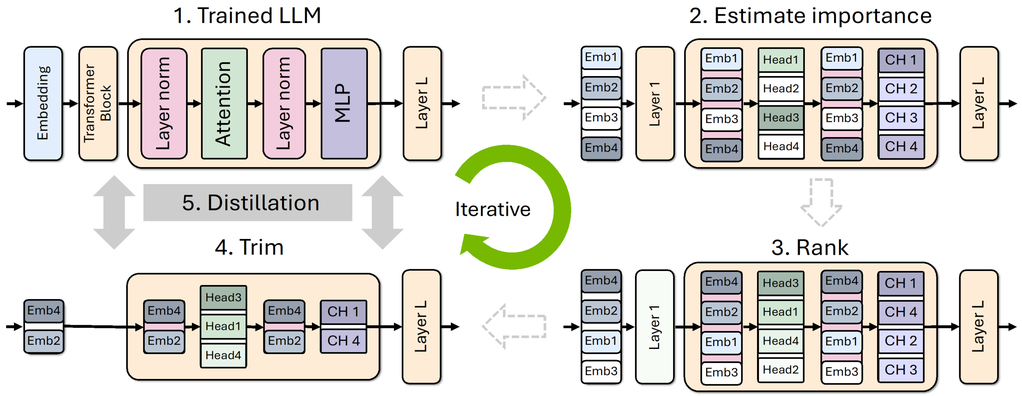
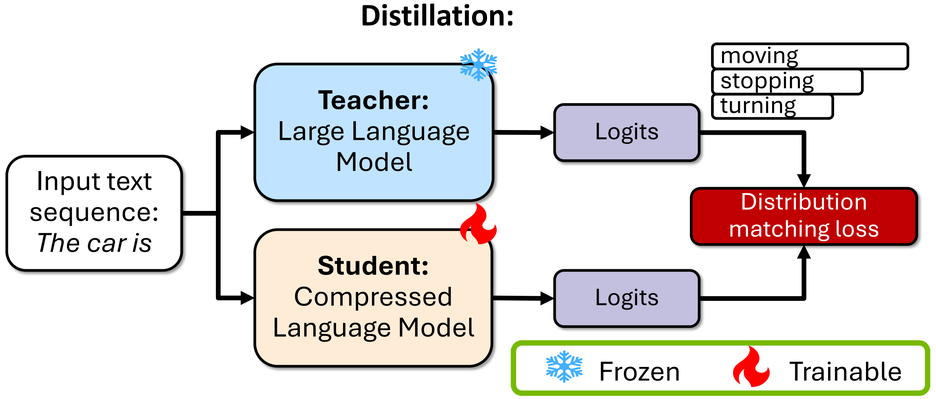
Gains from

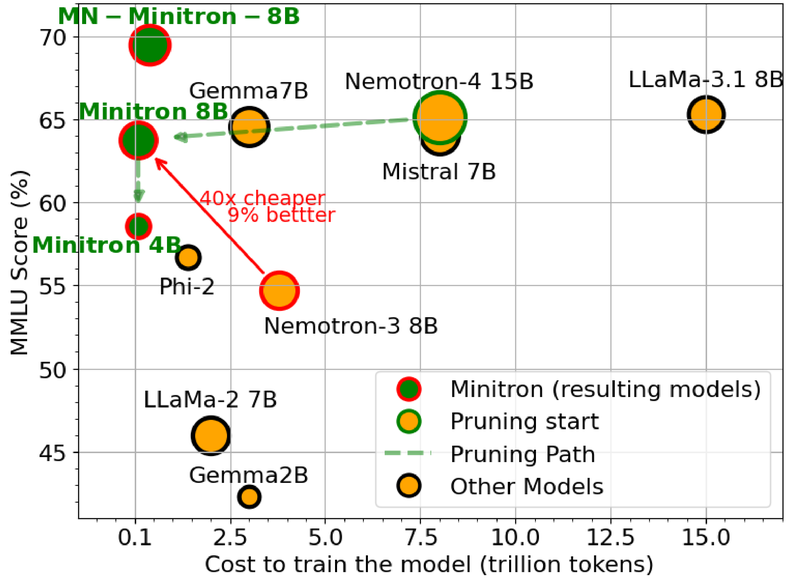
- Number Representation
  - FP32, FP16, Int8, FP4
  - (TF32, BF16)
  - ~16x, 32x
- Complex Instructions
  - DP4, HMMA, IMMA
  - ~12.5x
- Process
  - 28nm, 16nm, 7nm, 5nm, 4nm
  - ~2.5x, 3x
- Sparsity ~2x
- Die Size 2x
- Model efficiency has also improved – overall gain > 1000x



Minitron: Compact Models via Pruning and Distillation



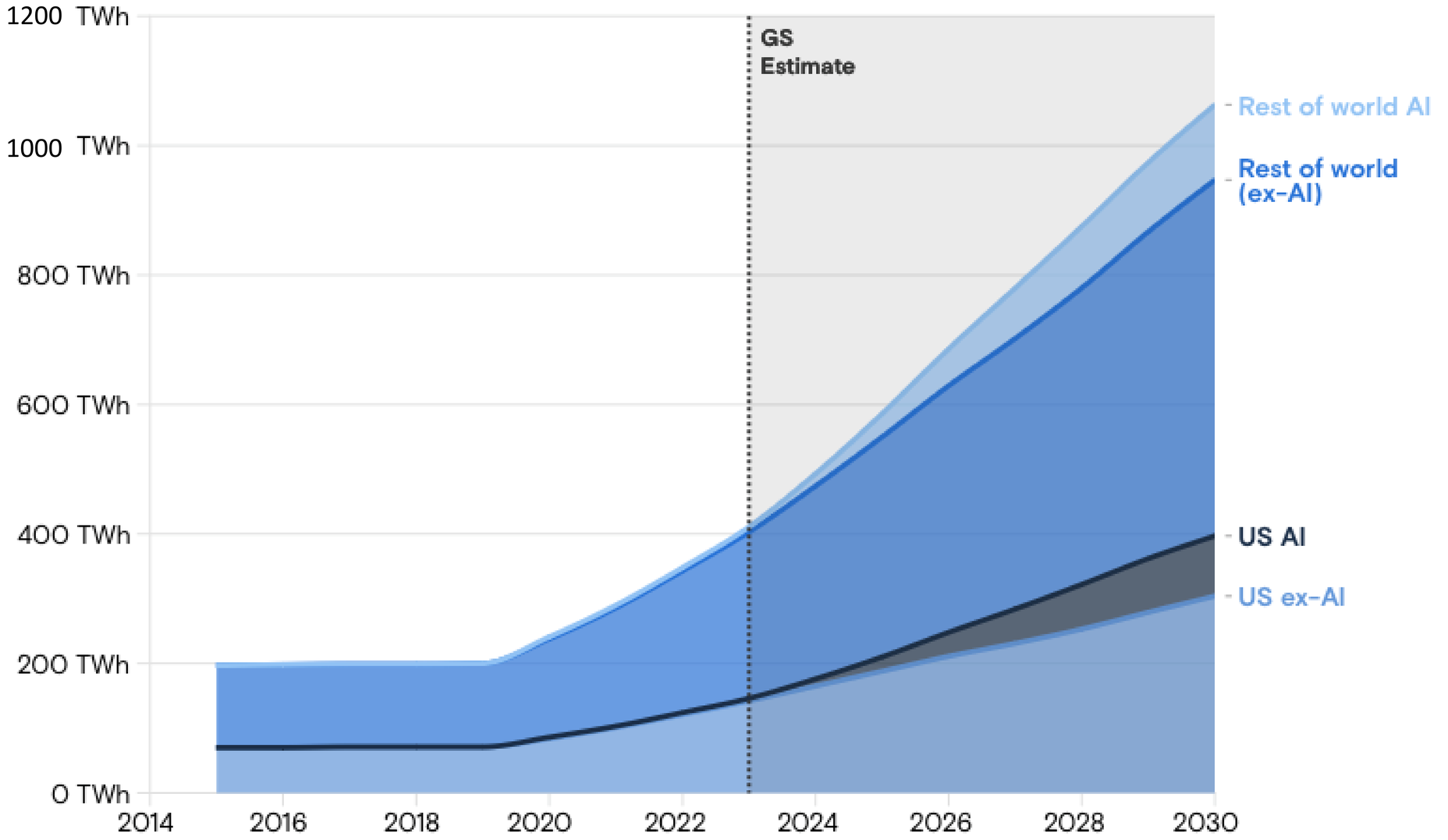




Benchmark	LLaMa-3	Mistral 7B	Gemma 7B	**MN-Minitron 8B**
Training tokens	15T	8T	6T	**0.4T**
General knowledge	65	64.1	64	**69.5**
Summarization	31	4.8	17	**32**
Reasoning	78	78.5	78	**80.4**
Coding	28	28.7	32	**36.2**

Details: *"LLM Pruning and Distillation in Practice: The Minitron Approach",* NVIDIA, arxiv: https://arxiv.org/abs/2408.11796, 2024

Data center power demand



Source: Masanet et al. (2020), Cisco, IEA, Goldman Sachs Research

Goldman Sachs