Compressing LLMs: The Truth is Rarely Pure and Never Simple

Ajay Jaiswal[1], Zhe Gan[2], Xianzhi Du[2], Bowen Zhang[2], Zhangyang Wang[1], Yinfei Yang[2]
[1]University of Texas at Austin, [2]Apple

TEXAS The University of Texas at Austin

VITA

Motivation and LLM-KICK

Surprise?

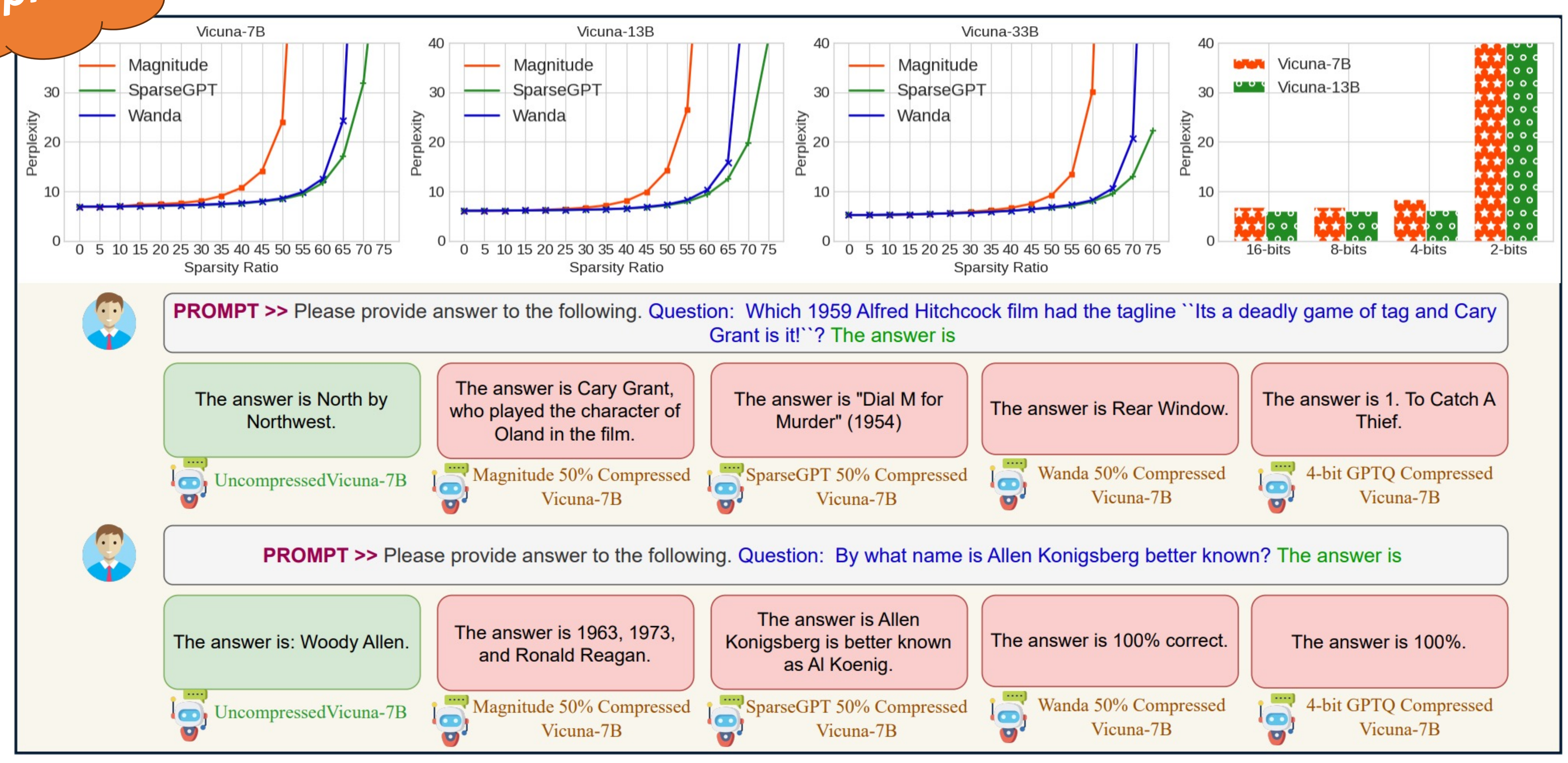
– Most (if not all) LLM compression works report perplexity
  – Perplexity measures how well a model predicts a given text but does not capture aspects such as coherence, relevance, knowledge faithfulness, or context understanding

– Specifically for compression, we observe that perplexity fails to capture subtle variations in capabilities of compressed LLMs, since they are all derived from the same dense counterparts

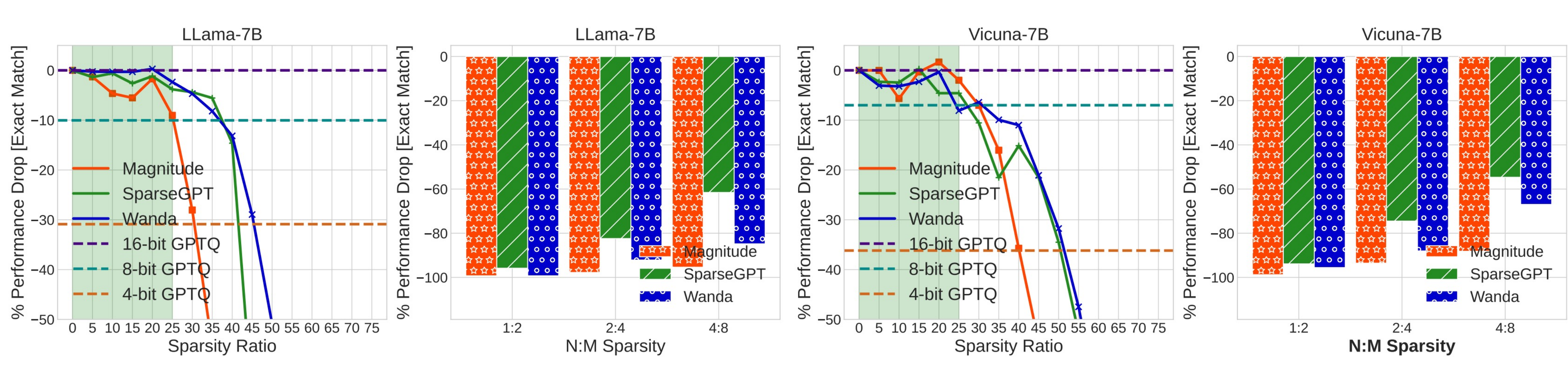
– We curate ***K**nowledge-**I**nstensive **C**ompressed LLM Benchmar**K*** (**LLM-KICK**), bringing the attention of LLM compression community towards incompetence of perplexity to reflect subtle changes in the LLM ability, and to understand what LLM compression truly promises and loses

Compression Methods:

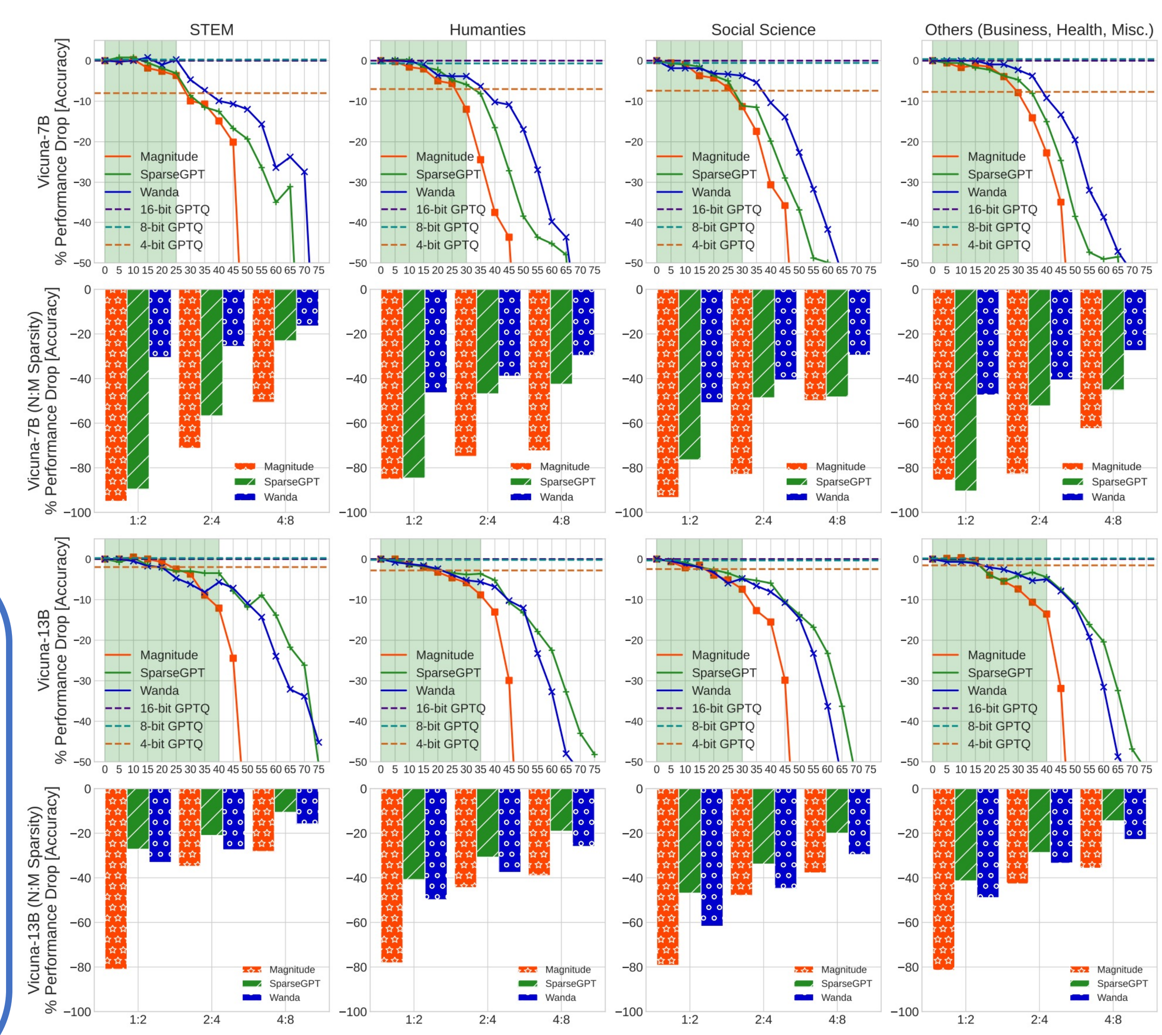
- ***Pruning:*** SparseGPT, Wanda, Magnitude-based
  - Test both unstructured sparsity (usually better accuracy) and semi-structured sparsity (hardware-friendly, 1:2, 2:4, 4:8)
- ***Quantization:*** GPTQ (4, 8, 16 bits)



How well Compressed LLMs Retain Knowledge?

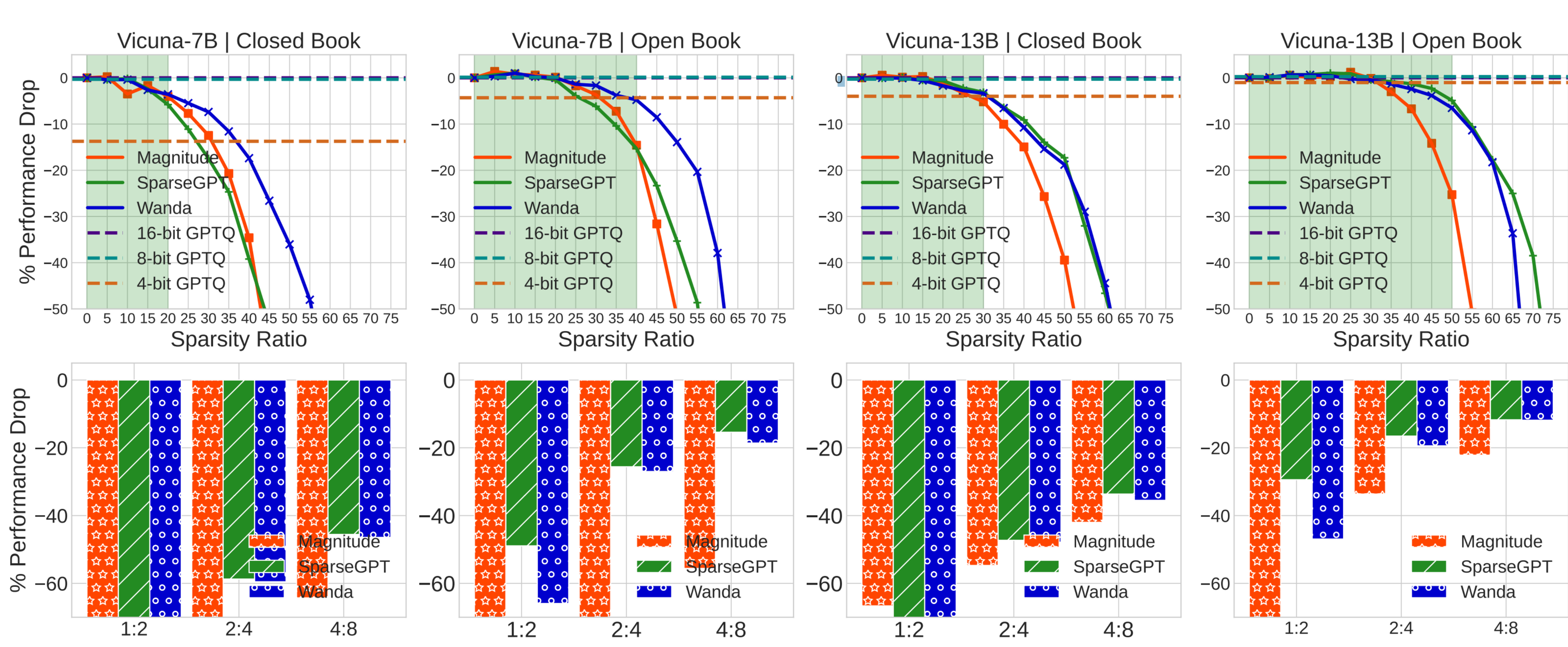


- All SoTA LLM unstructured pruning seemingly fail, even at "trivial" sparsities such as 30-35%
- No pruning method yet work for fine-grained structured N:M sparsity patterns, with performance drop as severe as ≥50%.
- Quantization seems better, but still is not a solved problem: ~8-10% drop in performance even for "non-aggressive" 8-bit quantization



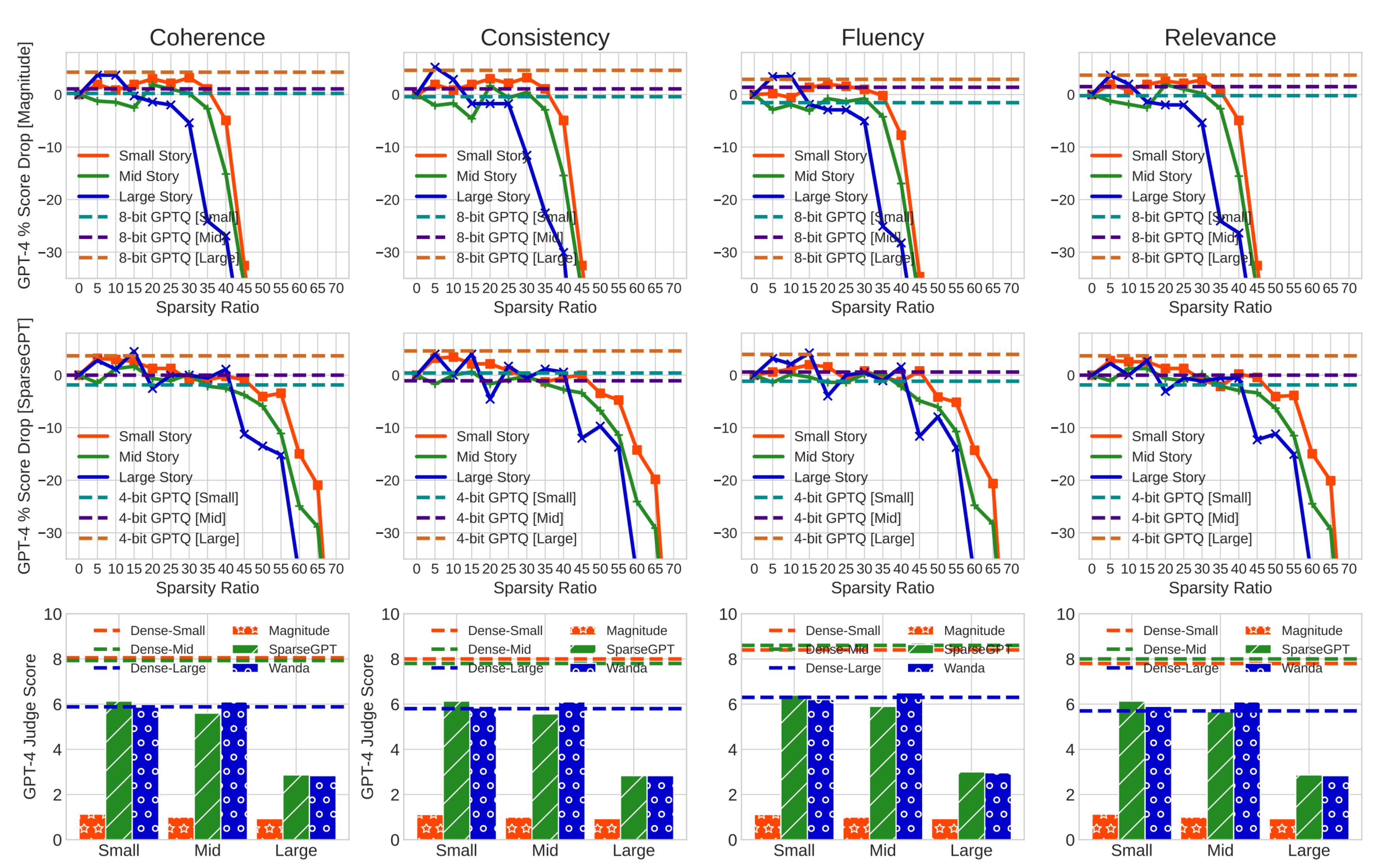
- On "simpler" MMLU, the performance drop of all pruning methods become smaller (even the naïve magnitude)
- **Yet still no success for N:M pruning!**
- Quantization becomes more successful too: 8-bit now match performance for Vicuna-7B and -13B
- Compression impacts some disciplines (Humanities, Social) more than others, uncovering data bias?

How well Compressed LLMs Retrieve Knowledge?



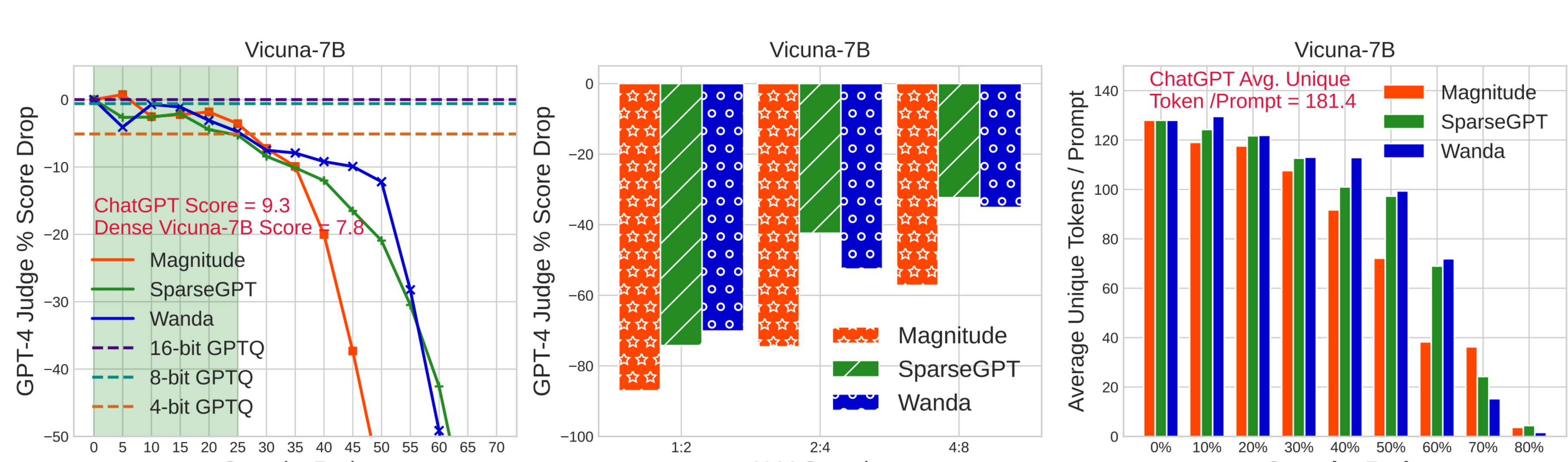
- When compressed LLMs are conditioned on external knowledge (open book QA) and assigned the task of in-context retrievers, they perform significantly well even in extremely high compression regime!
- Vicuna7B can remain matching till ~40% sparsity and 8-bit quantization, while Vicuna-13B can remain matching up to ~50% sparsity and 4-bit quantization.
- Yet, **Yet still no success for N:M pruning!**

How well Compressed LLMs Summarize Knowledge?



- All compression methods perform surprisingly well for in-context summarization
- Quantization again perform better than SoTA pruning
- While increasing context length (small -> mid -> large), the ability to digest longer context is affected more severely than smaller context
- **Yet still no success for N:M pruning!**

How well Compressed LLMs Follow Instructions?



- Pruning fails again at trivial sparsities (25-30%) while quantization remains okay. No N:M pruning works
- Interestingly, all compressed LLMs lose the ability to generate distinct unique content. Instead, they are much more prone tom producing more repetitive texts.

Follow @VITAGroupUT