



# Compressing LLMs: The Truth is Rarely Pure and Never Simple



**★** Ajay Jaiswal<sup>1</sup>, Zhe Gan<sup>2</sup>, Xianzhi Du<sup>2</sup>, Bowen Zhang<sup>2</sup>, Zhangyang Wang<sup>1</sup>, Yinfei Yang<sup>2</sup> University of Texas at Austin, <sup>2</sup>Apple

#### ➤ Motivation and LLM-KICK

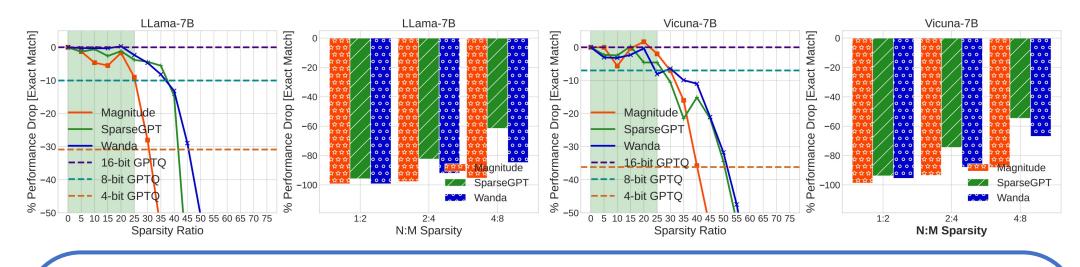
- Most (if not all) LLM compression works report perplexity
  - Perplexity measures how well a model predicts a given text but does not capture aspects such as coherence, relevance, knowledge faithfulness, or context understanding
- Specifically for compression, we observe that perplexity fails to capture subtle variations in capabilities of compressed LLMs, since they are all derived from the same dense counterparts
- We curate Knowledge-Instensive Compressed LLM Benchmark (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:**

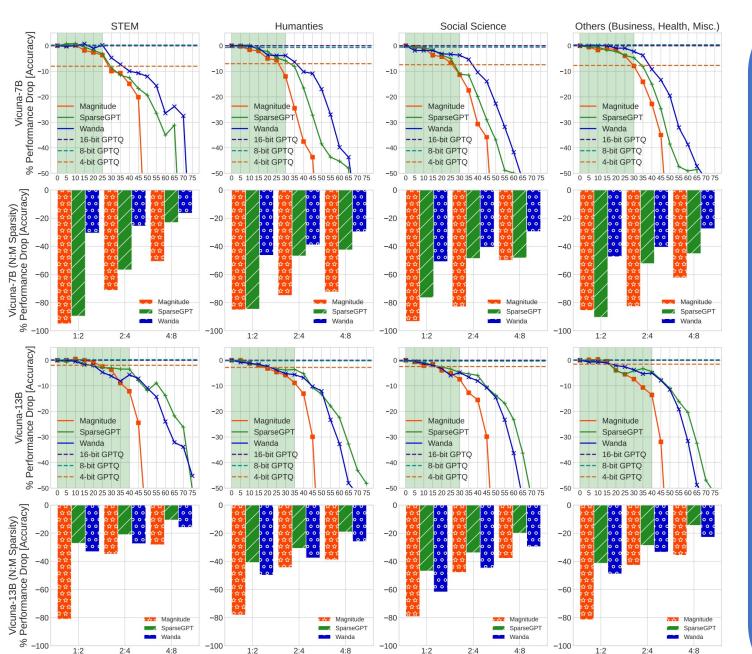
- **Pruning:** SparseGPT, Wanda, Magnitude-based
  - Test both unstructured sparsity (usually better accuracy) and semi-structured sparsity (hardware-friendly, 1:2, 2:4, 4:8)
- Quantization: GPTQ (4, 8, 16 bits)

#### Surprise? Vicuna-13B Vicuna-33B Magnitude Magnitude Magnitude SparseGPT SparseGPT Sparsity Ratio Sparsity Ratio Sparsity Ratio PROMPT >> Please provide answer to the following. Question: Which 1959 Alfred Hitchcock film had the tagline "Its a deadly game of tag and Cary Grant is it!"? The answer is The answer is Cary Grant, The answer is North by The answer is "Dial M for The answer is 1. To Catch A The answer is Rear Window who played the character of Murder" (1954) Northwest. Oland in the film. 4-bit GPTQ Compressed Magnitude 50% Compressed SparseGPT 50% Compressed Wanda 50% Compressed UncompressedVicuna-7B Vicuna-7B Vicuna-7B Vicuna-7B Vicuna-7B **PROMPT >>** Please provide answer to the following. Question: By what name is Allen Konigsberg better known? The answer is The answer is Allen The answer is 1963, 1973. The answer is: Woody Allen. The answer is 100% correct Konigsberg is better known The answer is 100% and Ronald Reagan. as Al Koenig. Wanda 50% Compressed Magnitude 50% Compressed 4-bit GPTQ Compressed SparseGPT 50% Compressed UncompressedVicuna-7B Vicuna-7B Vicuna-7B Vicuna-7B

## How well Compressed LLMs Retain Knowledge?

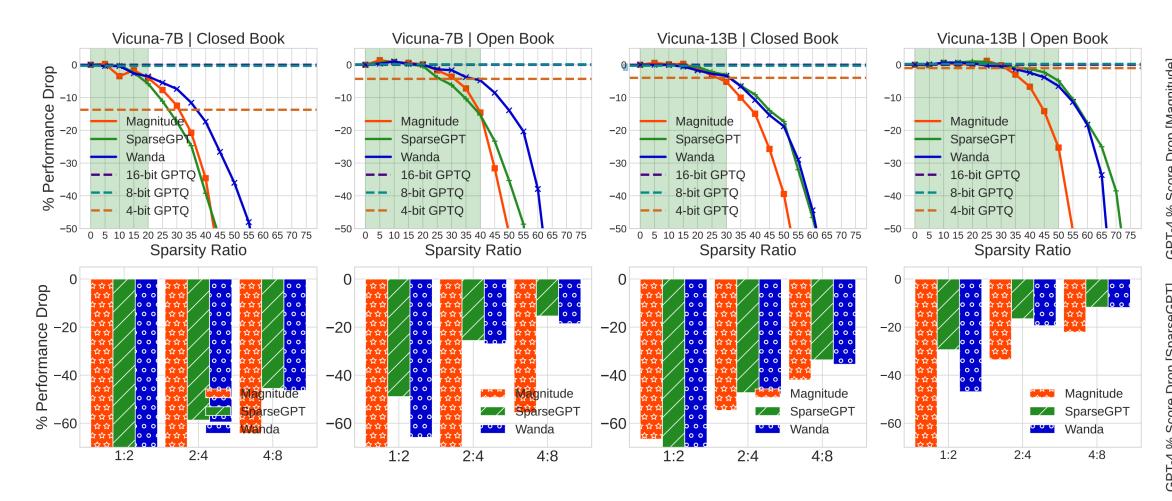


- All SoTA LLM unstructured pruning seemingly fail, even at "trivial" sparsities such as 30-35%
- No pruning method yet work for fine-grained structured N:M sparsity patterns, with performance drop as severe as ≥50%.
- Quantization seems better, but still is not a solved problem:  $\sim$ 8-10% drop in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

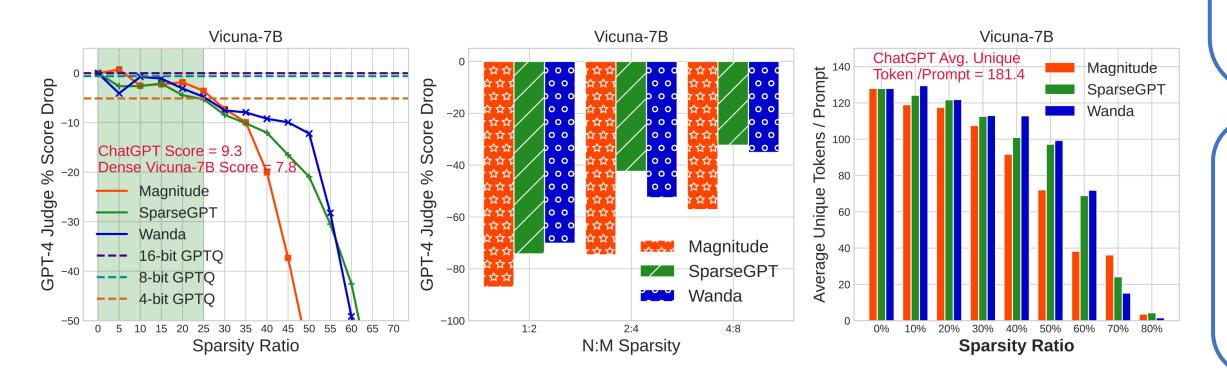
# How well Compressed LLMs Retrieve Knowledge?



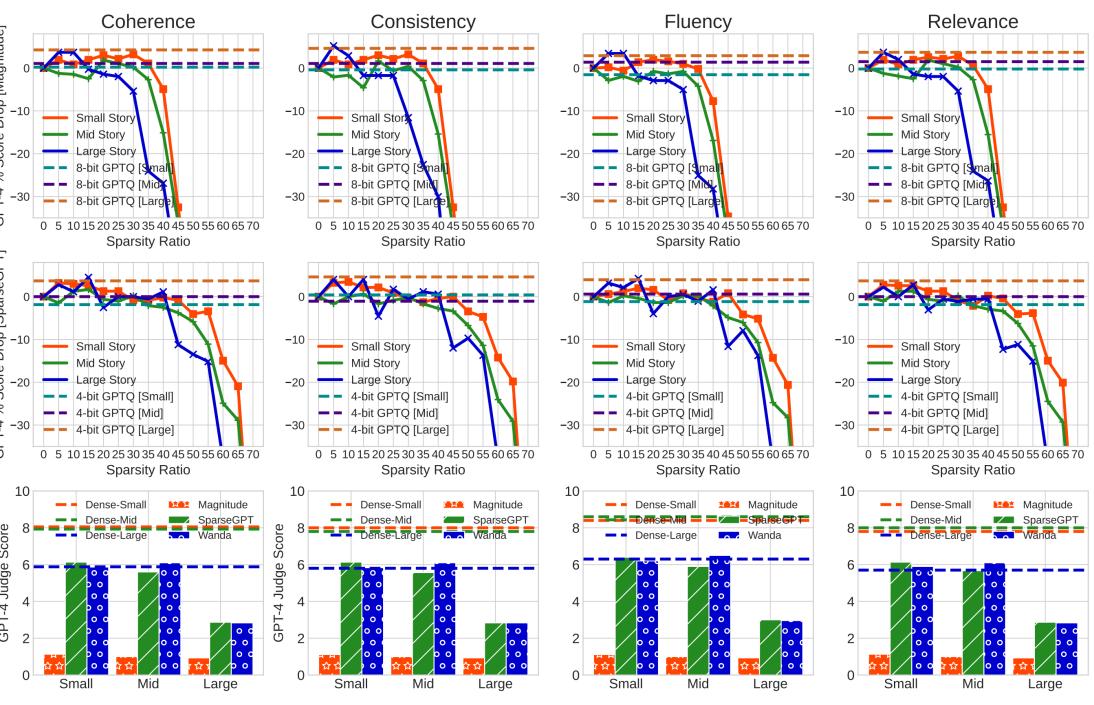
### • When compressed LLMs are conditioned on external knowledge (open book QA) and assigned the task of in-context retrievers, they perform significantly well even in extremely high compression regime!

- Vicuna7B can remain matching till  $\sim$ 40% sparsity and 8-bit quantization, while Vicuna-13B can remain matching up to  $\sim$ 50% sparsity and 4-bit quantization.
- Yet, Yet still no success for N:M pruning!

#### How well Compressed LLMs Follow Instructions?



### 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!
- 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.