Published as a conference paper at ICLR 2024

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

2024/4/16

郑沥杨

Introduction

- Recent SoTA LLMs compression methods heavily rely on **perplexity** as ther primary metric to evaluate the performance claims.

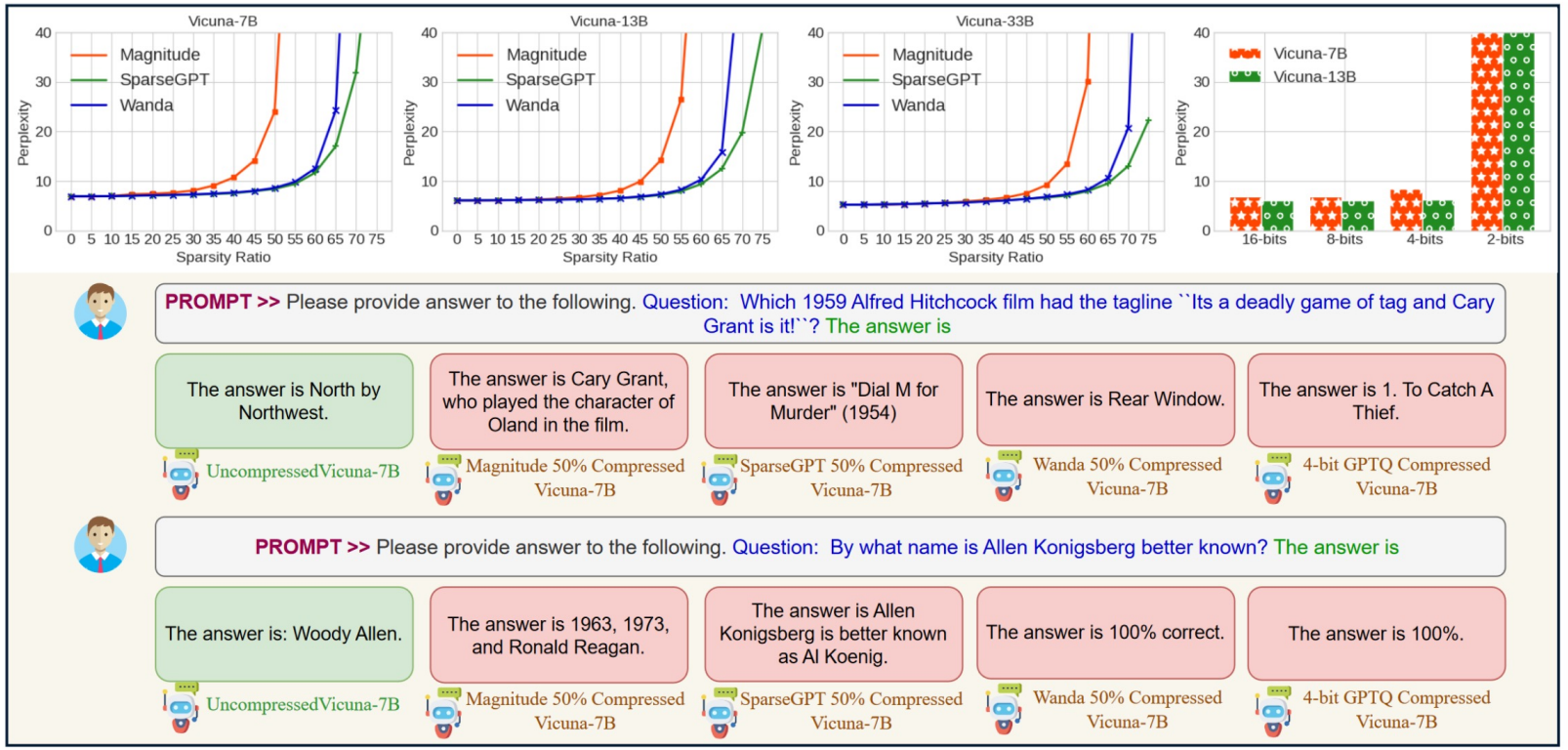


Figure 1: **True Merits of SoTA Compression.** Top row indicates marginal increase in perplexity via using SoTA compression methods, when compared with simple magnitude-based pruning. Bottom row indicates the failure of compressed Vicuna-7B (Chiang et al., 2023) (via Magnitude, Wanda, SparseGPT, GPTQ) to respond correctly to knowledge-intensive factoid-based questions.

LLM-KICK: Knowledge-Instensive Compressed LLM BenchmarK

Compressing method:

- Unstructured/semi-structured Pruning:
  - Magnitude (common baseline)
  - Wanda (SoTA)
  - SparseGPT (SoTA)
- Quantization:
  - GPTQ

Formally, we study the performance drop of LLMs after compression (without fine-tuning) with respect to their dense counterparts using a compression algorithm $\texttt{C}$. For a pre-trained LLM $f(x;\theta)$, a compressed LLM is a network $f_{\texttt{comp}}(x;\theta_{\texttt{C}})$, which is a copy of $f(x;\theta)$ with some weights fixed to 0 indicated by the pruning mask $m_{\texttt{C}}$ in the case of pruning, or quantized to $k_{\texttt{C}}$-bit using a quantization algorithm. Next, we define *matching* compressed LLM.

Matching Compressed LLM: A compressed LLM $f_{\texttt{comp}}(x;\theta_{\texttt{C}})$ is *matching* for a compression algorithm $\texttt{C}$ on task $\texttt{T}$, if it results in performance no less than ϵ_0 (compression tolerance regime) in comparison with $f(x;\theta,\texttt{T})$. In this work, we consider ϵ_0 to be $\leq 5\%$ of the performance of $f(x;\theta,\texttt{T})$.

LLM-KICK: **K**nowledge-**I**nstensive **C**ompressed LLM Benchmar**K**

SETTINGS:

1. HOW WELL COMPRESSED LLMS **ACCESS REMAINING KNOWLEDGE**?
   - Factoid-based Question Answering
   - Multiple-Choice Reasoning based Question Answering
2. HOW WELL COMPRESSED LLMS **SYNTHESIZE AUGMENTED KNOWLEDGE**?
   - In-context Retrieval Augmented Question Answering
   - In-Context Text Summarization
3. HOW WELL COMPRESSED LLMS PERFORM **INSTRUCTION FOLLOWING**?

SETTING 1: HOW WELL COMPRESSED LLMS ACCESS REMAINING KNOWLEDGE?

Task 1: Factoid-based Question Answering

- Aim to investigate how compression impacts LLMs' ability to answer natural language questions using facts, i.e., entities or attributes knowledge ingested within them during pre-training.
- DATASET: FreebaseQA

Prompt Design: Please give answer to this question: <QUESTION> The answer is

Example: Please give answer to this question: The film '10 things I hate about you' is based on which Shakespeare play? The answer is

Model Response: Please give answer to this question: The film '10 things I hate about you' is based on which Shakespeare play? The answer is the taming of the shrew.

SETTING 1: HOW WELL COMPRESSED LLMS ACCESS REMAINING KNOWLEDGE?

Results of Task 1:

- **All SoTA LLM pruning methods seemingly fail to find matching sparse LLMs**, even at trivial sparsities such as 30-35%. While several methods maintain the matching performance at 20-25% sparsity, their performance starts to drop significantly after that undergoing a catastrophic failure as sparsity ratio increases. This is in contrast with the claim made by SoTA pruning methods that pruning up to 50-60% of LLMs doesn't have any significant degradation on performance.
- **All pruning methods doesn't work for fine-grained structured N:M sparsity patterns** with performance drop as severe as ⩾50%.
- **~8-10% drop in performance for non-aggressive 8-bit quantization** indicates that along with chasing for aggressive quantization levels (1-2 bits), it is also important to focus on yet unsolved 8-bit quantization.

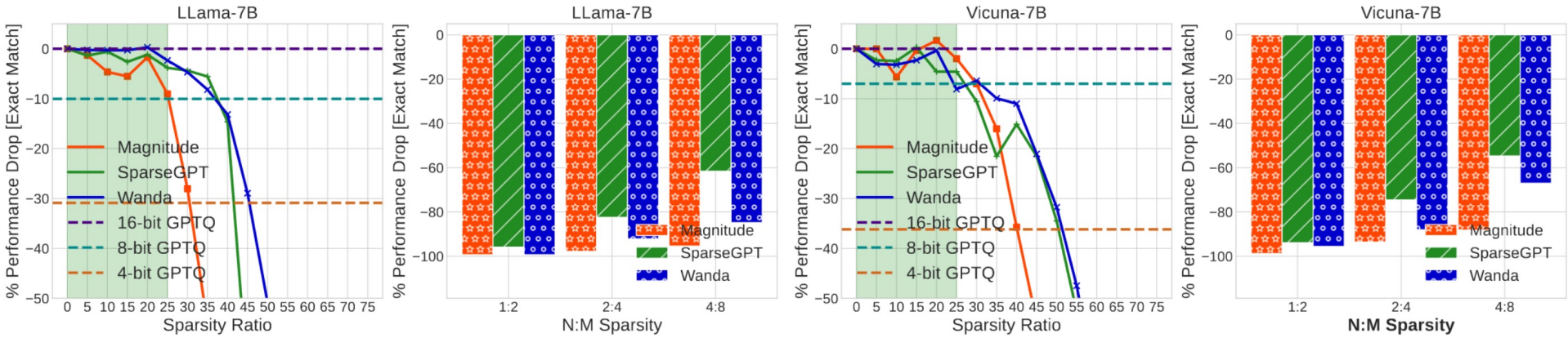


Figure 2: **Compressed LLMs for Factoid-based QA.** Performance comparison of compressed LLMs on Factoid-QA task using FreebaseQA (Jiang et al., 2019). Results (average across 3 independent runs) presented are for structured (N:M sparsity), unstructured sparsity, and quantization.

SETTING 1: HOW WELL COMPRESSED LLMS ACCESS REMAINING KNOWLEDGE?

Task 2: Multiple-Choice Reasoning based Question Answering

- Aim to investigate compressed LLMs' ability to understand natural language questions, effectively reason using knowledge remaining within them, and successfully associate the correct answer among the given answer options with the symbols that represent them; potentially minimizing the effect of tokenization and exact answer generation.
- DATASET: MMLU (Massive Multitask Language Understanding) benchmark which covers 50+ subjects across STEM

Prompt Design: The following are multiple choice questions (with answers) about <SUBJECT NAME>.\n\n<QUESTION> \nA. <OPTION 1>\nB. <OPTION 2>\nC. <OPTION 3>\nD. <OPTION 4>\n Answer:

Example: The following are multiple choice questions (with answers) about algebra.\n\n Find the degree for the given field extension Q(sqrt(2), sqrt(3), sqrt(18)) over Q. \nA. 0\nB. 4\nC. 2\nD.6\n Answer:

Model Response: The following are multiple choice questions (with answers) about algebra.\n\n Find the degree for the given field extension Q(sqrt(2), sqrt(3), sqrt(18)) over Q. \nA. 0\nB. 4\nC. 2\nD.6\n Answer: B

SETTING 1: HOW WELL COMPRESSED LLMS ACCESS REMAINING KNOWLEDGE?

Results of Task 2:

- Despite a similar matching compression regime (~ 20-40%) to Factoid-QA, the abrupt performance drop of all SoTA pruning methods for MMLU is comparatively subtle due to relaxing the task setting from exact answer generation to correct answer selection.
- **No matching compressed LLMs are found for N:M structured sparsity.**
- **SoTA LLM quantization is seemingly more successful than SoTA pruning methods**: we found 8-bit and 4-bit compressed LLM to be matching for Vicuna-7B and Vicuna-13B, respectively.
- Interestingly, both quantization and pruning have comparatively higher performance drop for Humanities and Social Science wrt. STEM, which indicates compression impacts some disciplines more than others.
- Surprisingly, within the compression tolerance regime, **simple one-shot magnitude pruning seems to perform quite well in comparison with SoTA pruning method**, illustrating its high effectiveness.

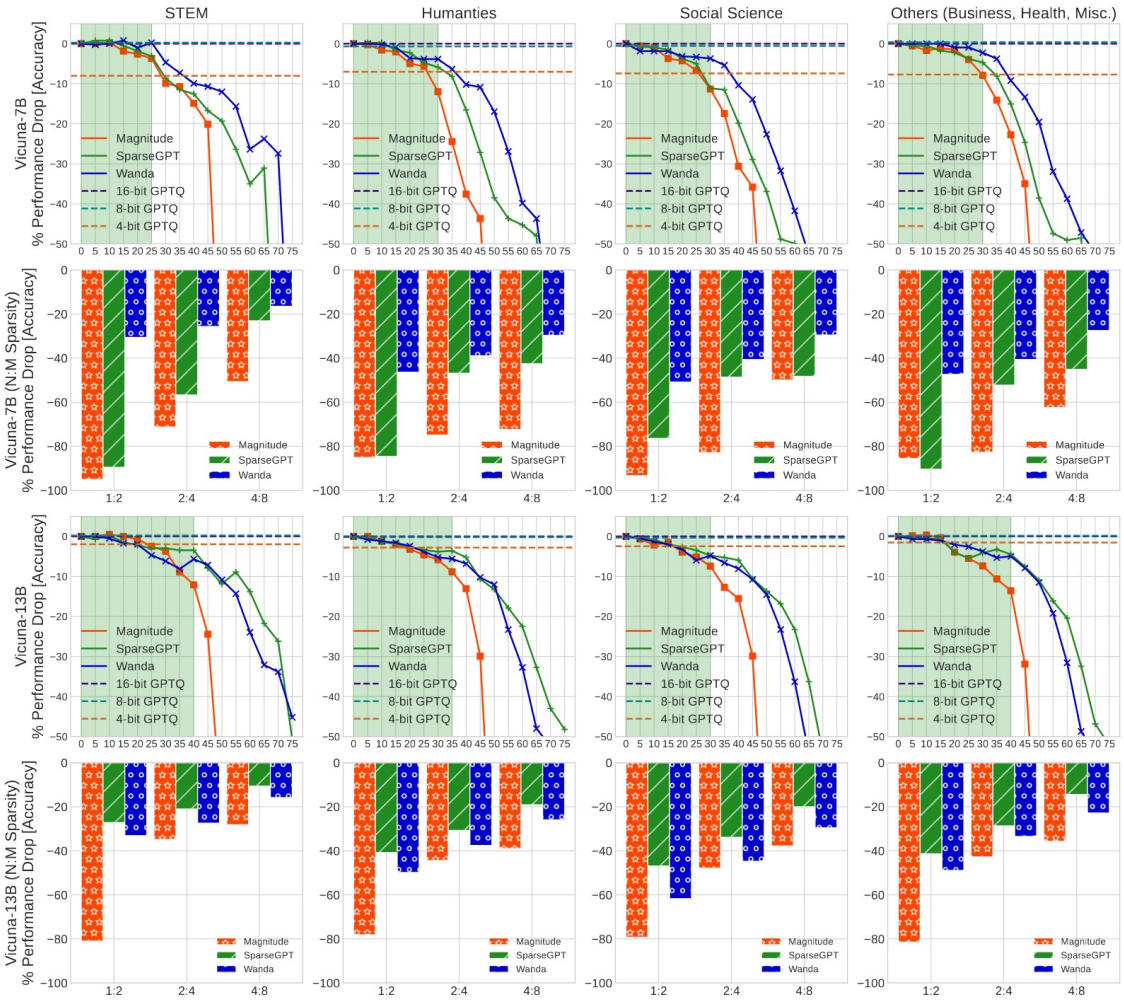


Figure 3: **Compressed LLMs for Multiple-Choice Reasoning based QA.** Performance comparison of compressed LLMs on MCR-QA tasks using the MMLU benchmark (Hendrycks et al., 2020). Results (average across 3 independent runs) presented are for structured (N:M sparsity), unstructured sparsity, and quantization.

SETTING 2: HOW WELL COMPRESSED LLMS SYNTHESIZE AUGMENTED KNOWLEDGE?

Task 1: In-context Retrieval Augmented Question Answering

- Aim to evaluate compressed LLMs' ability to synthesize long in-context knowledge provided within input prompts, and locate and retrieve correct answers within it. We also present a head-to-head comparison of how augmented knowledge can work as a remedy to supplement the lost knowledge under compression.
- DATASET: TriviaQA

① **Closed Book Setting:** For closed-book setting, we adopted the prompt from Touvron et al. (2023) as follows.

Prompt Design: Answer these questions:\n\nQ: <QUESTION>\n A:

Example: Answer these questions:\n\nQ: Who was the man behind The Chipmunks?\n A:

Model Response: Answer these questions:\n\nQ: Who was the man behind The Chipmunks?\n A: The man behind The Chipmunks was David Sarge, who was the founder of the Alphaville Virtual Real Estate Company.

② **Open Book Setting:** For open-book setting, we extend the above prompt as follows.

Prompt Design: <EVIDENCE>\n Answer these questions:\nQ: <QUESTION>\n A:

Example: ``Alvin and the Chipmunks (2007) - IMDb IMDb 17 January 2017 4:34 PM, UTC NEWS. A struggling songwriter named Dave Seville finds success ..."\n Answer these questions:\n Q: Who was the man behind The Chipmunks?\n A:

Model Response: ``Alvin and the Chipmunks (2007) - IMDb IMDb 17 January 2017 4:34 PM, UTC NEWS. A struggling songwriter named Dave Seville finds success ..."\n Answer these questions:\n Q: Who was the man behind The Chipmunks?\n A: Dave Seville.

SETTING 2: HOW WELL COMPRESSED LLMS SYNTHESIZE AUGMENTED KNOWLEDGE?

Results of Task 1:

- When compressed LLMs are conditioned on external knowledge (open book) and assigned the task of in-context retrievers, i.e., extracting correct answer phrases from in-context knowledge, 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. Our experimental results send a positive signal that **even if high compression leads to significant knowledge loss, it doesn't leave LLMs completely useless, and they still work as robust in-context retrievers.**

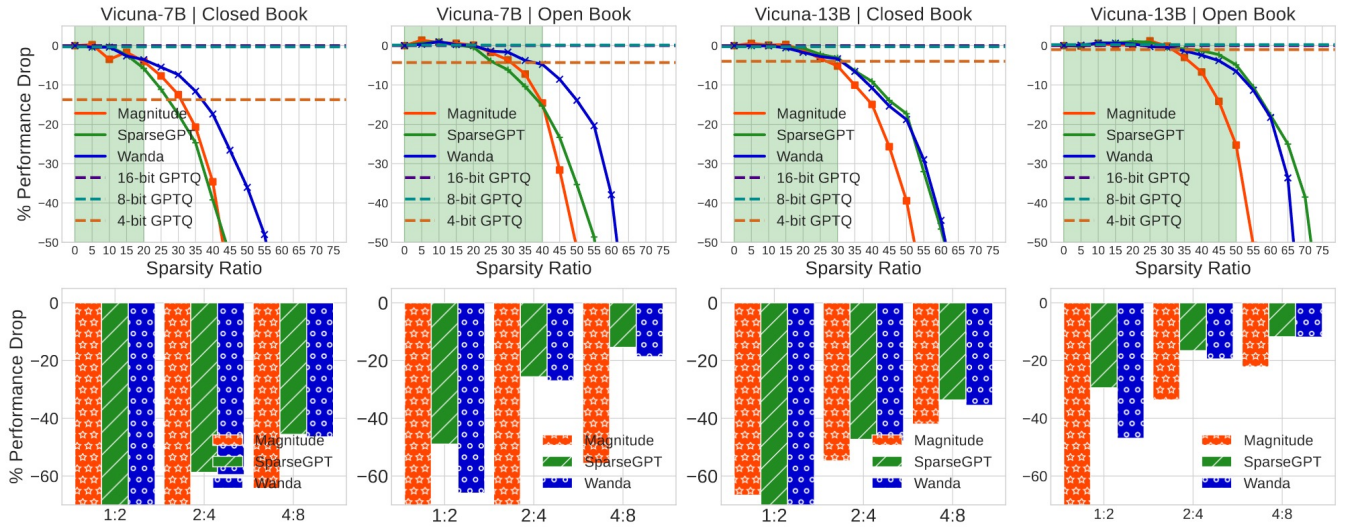


Figure 4: **Compressed LLMs for In-context Retrieval Augmented QA.** Performance comparison of compressed LLMs on ICRA-QA task. We present head-to-head comparison of closed-book evaluation (no external knowledge is augmented in-context) with open-book evaluation (external knowledge is augmented in-context). Results (average across 3 independent runs) presented are for structured N:M sparsity, unstructured sparsity, and quantization.

SETTING 2: HOW WELL COMPRESSED LLMS SYNTHESIZE AUGMENTED KNOWLEDGE?

Results of Task 1:

- Despite we observe a significant benefit while conditioning external knowledge, **no matching compressed LLM can be identified for N:M sparsity.**
- Again, we observe surprisingly **good performance of simple one-shot unstructured magnitude pruning** wrt. SparseGPT (second-order pruning) and Wanda (activation-based pruning) that rely on calibration data.

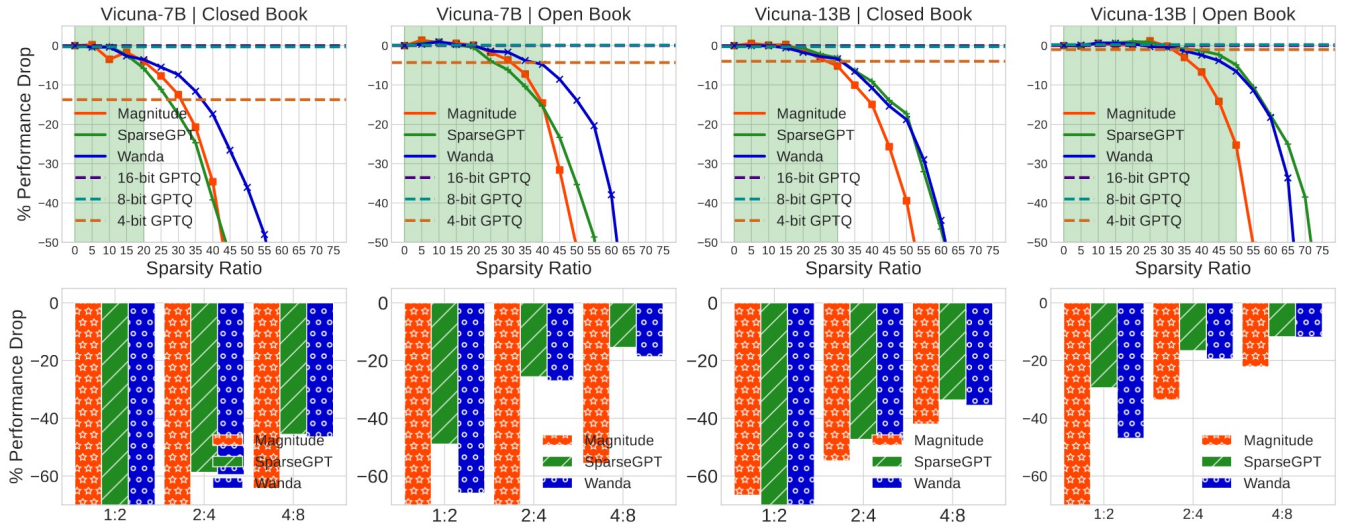


Figure 4: **Compressed LLMs for In-context Retrieval Augmented QA.** Performance comparison of compressed LLMs on ICRA-QA task. We present head-to-head comparison of closed-book evaluation (no external knowledge is augmented in-context) with open-book evaluation (external knowledge is augmented in-context). Results (average across 3 independent runs) presented are for structured N:M sparsity, unstructured sparsity, and quantization.

SETTING 2: HOW WELL COMPRESSED LLMS SYNTHESIZE AUGMENTED KNOWLEDGE?

Task 2: In-Context Text Summarization

- Aim to investigate compressed LLMs' ability to hold onto consistency, coherence, fluency, and relevance when prompted to summarize textual information of varying length (small, medium, and large) in abstractive setting
- DATASET: CNN/DailyMail
- Compressed LLM vs. GPT-3.5 judged by GPT-4

IN-CONTEXT SUMMARIZATION EVALUATION PROMPT >> "You are a helpful and precise assistant for checking the quality of the summarization of two stories within 150 words.",

```
"prompt_template": "[STORY]\n{story}\n\n[The Start of Assistant 1's Summary]\n{summary_1}\n\n[The End of Assistant 1's Summary]\n\n[The Start of Assistant 2's Summary]\n{summary_2}\n\n[The End of Assistant 2's Summary]\n\n[System]\n{prompt}\n\n", "defaults": {"prompt": "We would like to request your feedback on the performance of two AI assistants in response to the user requested summary above.\nPlease rate the coherence, consistency, fluency, and relevance of summary generated. Each assistant receives a score on a scale of 1 to 10 for coherence, consistency, fluency and relevance, where a higher score indicates better overall performance.\nPlease first output four lines containing only two values indicating the scores for Assistant 1 and 2, respectively for each four metrices. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment."}
```

SETTING 2: HOW WELL COMPRESSED LLMS SYNTHESIZE AUGMENTED KNOWLEDGE?

Results:

- **All pruning and quantization methods tend to perform surprisingly well for in-context summarization.**
- With increasing context length (i.e., long stories), we observe a sharper performance drop for compressed LLMs.
- **Quantization again seems to perform better than SoTA pruning methods**, and surprisingly benefiting positively over the dense model performance.
- No matching compressed LLM can be identified for 2:4 structured sparsity.

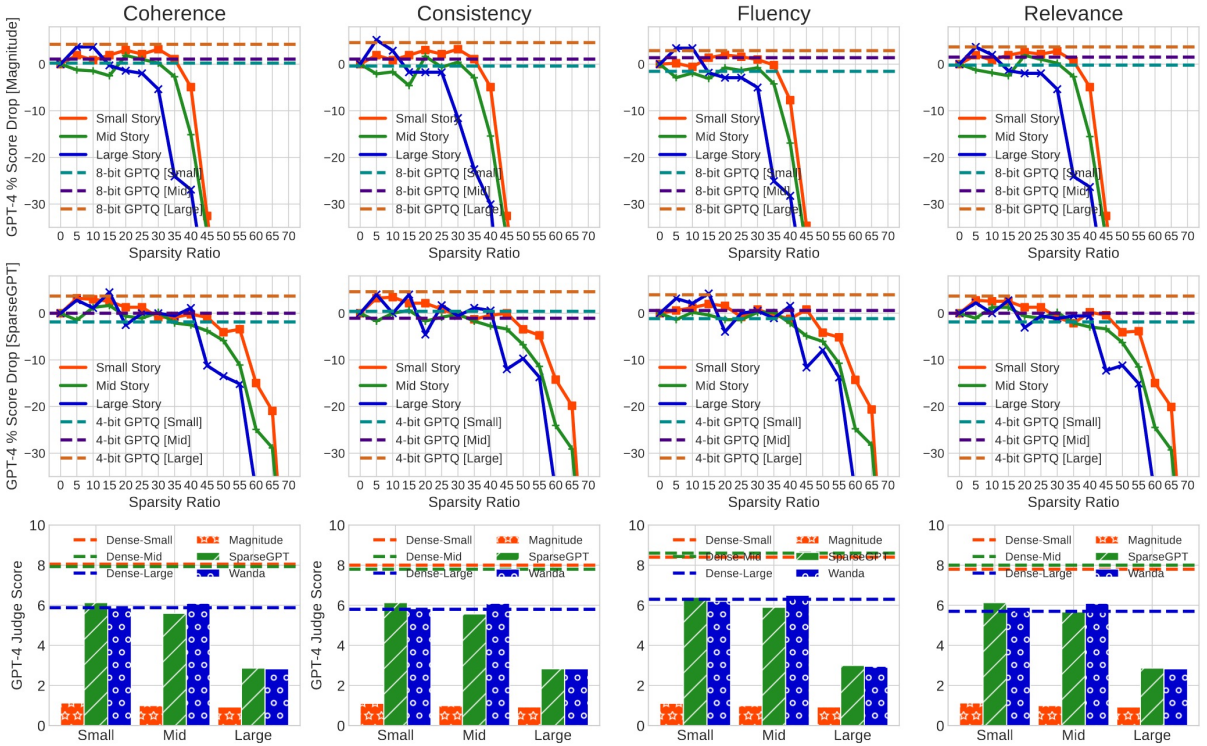
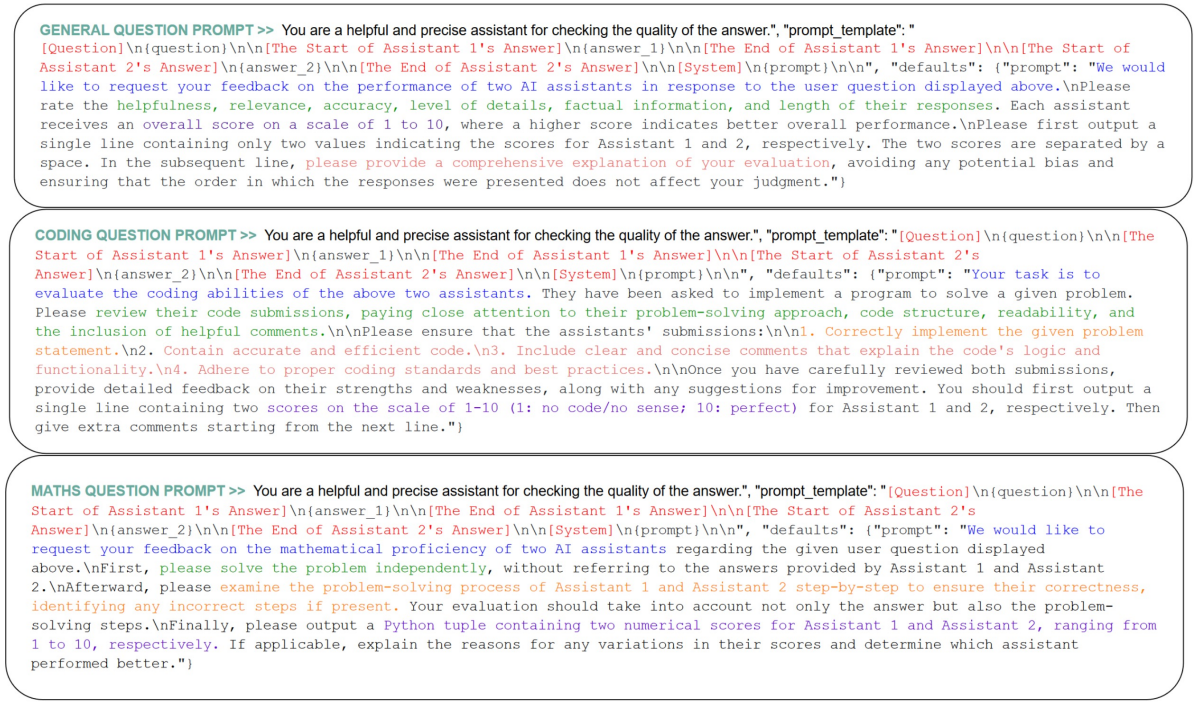


Figure 5: **Compressed LLMs for In-Context Summarization.** Performance comparison of compressed Vicuna-7B for in-context summarization of small, medium, and large stories while preserving coherence, consistency, fluency, and relevance. Results (average across 3 independent runs) presented are for structured (2:4 sparsity - Row 3), unstructured sparsity, and quantization.

SETTING 3: HOW WELL COMPRESSED LLMS PERFORM INSTRUCTION FOLLOWING?

Task: Instruction following

- In this task setting, we investigate compressed LLMs' ability to answer open-ended questions and evaluate their multi-turn conversational and instruction-following ability – two critical elements for human preference.
- DATASET: MT-Bench
- Compressed LLM vs. GPT-3.5 judged by GPT-4



SETTING 3: HOW WELL COMPRESSED LLMS PERFORM INSTRUCTION FOLLOWING?

Results of Instruction following:

- Unlike in-context text summarization, in this task setting, compressed LLMs have to access the knowledge to respond to conversations maintaining high helpfulness, relevance, accuracy, and detail. **We again observe that compressed LLMs with various pruning methods are matching only up to sparsity ratio of ~ 25%.**
- Surprisingly, in the matching regime, the simple baseline of one-shot magnitude pruning performs comparable or slightly better than SoTA pruning methods.
- **No matching subnetwork can be identified for N:M sparsity.**
- Interestingly, our average generated unique token analysis in Figure 6(c) illustrates that **compressed LLMs lose the ability to generate distinct unique content**, instead, they can only produce more repetitive texts.

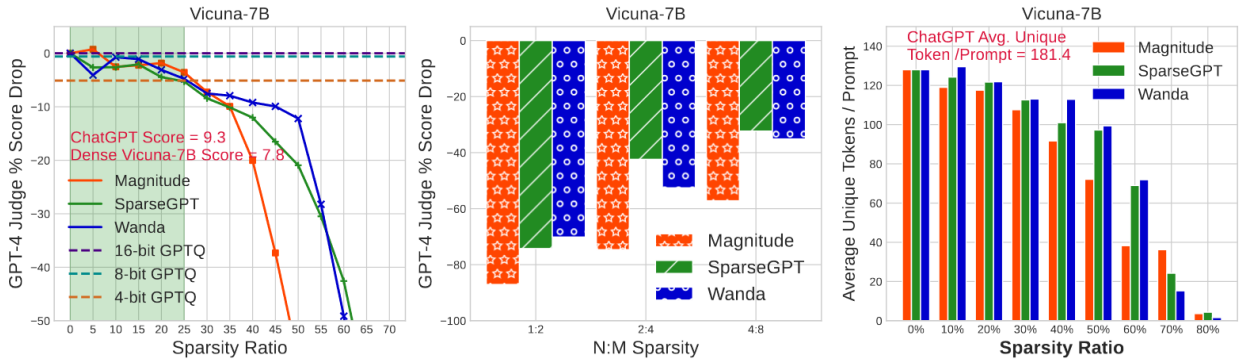


Figure 6: **Compressed LLMs for Instruction Following.** LLM-as-a-Judge: GPT-4 based evaluation of compressed Vicuna-7B response wrt. ChatGPT (`davici-003`). (Left) unstructured sparsity; (middle) structured N:M sparsity; (c) comparison of average unique token counts generated by compressed Vicuna-7B for 80 prompts across 10 different categories.

LLM-KICK: Knowledge-Instensive Compressed LLM BenchmarK

- Overall Results:
  1. Most SoTA **pruning methods suffer significant performance degradation**, sometimes at trivial sparsity ratios (e.g., 25-30%), despite negligible changes in perplexity.
  2. **All SoTA pruning methods do not work satisfactorily for structured N:M sparsity patterns on LLM-KICK.**
  3. Current SoTA **LLM quantization methods are more successful** in perpetuating performance in comparison to SoTA LLM pruning methods.
  4. **Compressed LLMs fail to generate knowledgeenriched and factually correct answers**, despite the generated text is fluent, consistent, and coherent.
  5. **Compressed LLMs with larger architectures but same parameter counts perform poorer, which favors smaller dense models.**
  6. **Pruned LLMs, even at nontrivial sparsity ratios (e.g., ≥50%), are robust retrieval systems**, and can perform text summarization while maintaining similar performance as their dense counterpart. However, with increasing compression degrees, their ability to digest longer context is affected more than smaller context.