

When Reasoning Meets Compression: Benchmarking Compressed Large Reasoning Models on Complex Reasoning Tasks

Nan Zhang, Yusen Zhang, Prasenjit Mitra, Rui Zhang
The Pennsylvania State University
`{njz5124, rmz5227}@psu.edu`

Abstract

Recent open-source large reasoning models (LRMs) exhibit strong performance on complex reasoning tasks, but their large parameter count makes them prohibitively expensive for individuals. The compression of large language models (LLMs) offers an effective solution to reduce cost of computational resources. However, systematic studies on the performance of compressed LLMs in complex reasoning tasks, especially for LRMs, are lacking. Most works on quantization and pruning focus on preserving language modeling performance, while existing distillation works do not comprehensively benchmark student models based on reasoning difficulty or compression impact on knowledge and reasoning. In this paper, we benchmark compressed DeepSeek-R1 models on four different reasoning datasets (AIME 2024, FOLIO, Temporal Sequences of BIG-Bench Hard, and MuSiQue), ranging from mathematical to multihop reasoning, using quantization, distillation, and pruning methods. We benchmark 2.51-, 1.73-, and 1.58-bit R1 models that adopt dynamic quantization. We also benchmark distilled R1 models that are based on LLaMA or Qwen and run SparseGPT on them to obtain various sparsity levels. Studying the performance and behavior of compressed LRMs, we report their performance scores and test-time compute (number of tokens spent on each question). Notably, using MuSiQue, we find that parameter count has a much greater impact on LRMs' knowledge memorization than on their reasoning capability, which can inform the choice of compression techniques. Through our empirical analysis of test-time compute, we find that shorter model outputs generally achieve better performance than longer ones across several benchmarks for both R1 and its compressed variants, highlighting the need for more concise reasoning chains.

1 Introduction

Large reasoning models (LRMs) excel at complex reasoning tasks. DeepSeek-R1 (Guo et al., 2025) is the first open-source LRM to successfully replicate OpenAI's O1-level performance. Trained based on DeepSeek-V3 (DeepSeek-AI et al., 2025), R1 adopts large-scale reinforcement learning without supervised fine-tuning (SFT). Its open-source nature facilitates research on reasoning models (Srivastava et al., 2025; Mondillo et al., 2025). However, due to its large size (671B total parameters), deploying it locally can be costly and even infeasible for individuals, which hinders AI democratization.

Compression of large language models (LLMs) includes quantization, distillation, and pruning. Its goal is to reduce computational resources (*e.g.*, GPU memory and disk space) and provide inference speedup. As the weights of vanilla LLMs are typically in 16-bit floating values (*e.g.*, FP16 or BF16), quantization means to convert these high-precision values to lower-precision ones. Representative techniques include dynamic quantization by Unsloth (Daniel Han & team, 2023), activation-aware quantization (Lin et al., 2024), and post-training quantization (Frantar et al., 2022). Distillation aims to transfer knowledge from

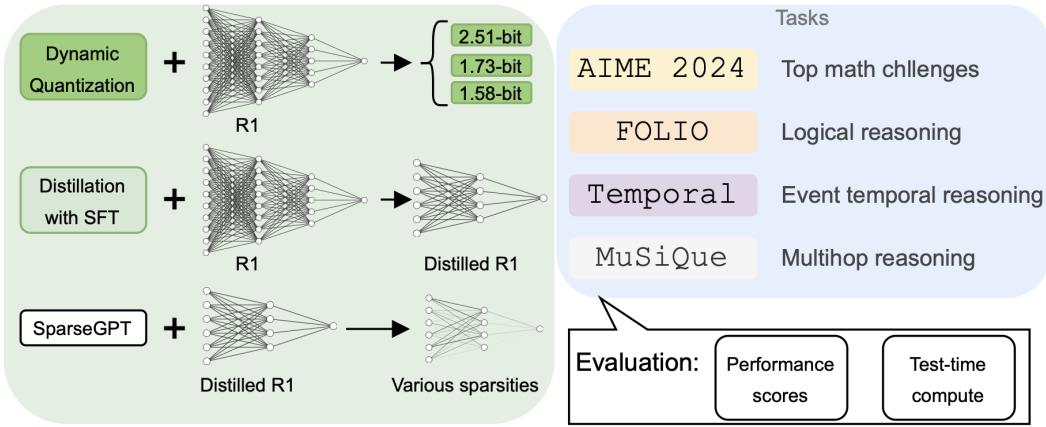


Figure 1: An overview of selected compression methods and tasks. We evaluate R1 and its compressed variants based on their performance scores and test-time compute.

a larger teacher model to a smaller student model under black-box (Li et al., 2024a) or white-box (Gu et al., 2024) settings. As for pruning, it aims to sparsify LLMs by introducing 0s to replace nonzero weights. Representative techniques include unstructured pruning (Zhang et al., 2024a; Frantar & Alistarh, 2023) and structured pruning (Xia et al., 2024; Ma et al., 2023).

Although existing quantization and pruning methods claim to preserve the performance of non-reasoning LLMs after compression such as LLaMA (Touvron et al., 2023) and OPT (Zhang et al., 2022), exploring these methods on LRM s offers a timely research of democratizing the latest LRM s. Moreover, since these methods mainly adopt perplexity and simple end-task scores for evaluation, the assessment of compressed LRM s will be fundamentally different. Without focusing on basic linguistic capabilities (*e.g.*, simple language understanding and generation), evaluation of compressed reasoning models needs to align with their uncompressed counterparts. Regarding distillation, recent works either fail to comprehensively evaluate their student models on diverse reasoning benchmarks of varying difficulty or neglect to consider compression impact on knowledge and reasoning. A comprehensive evaluation that includes all popular compression methods is valuable.

Therefore, due to the lack of compression research on LRM s, we benchmark the performance and behavior of compressed DeepSeek-R1 on various reasoning tasks in this paper. Our benchmarking framework is shown in Figure 1. We test dynamic quantization (Daniel Han & team, 2023), distillation with SFT (Guo et al., 2025), and SparseGPT (Frantar & Alistarh, 2023) on R1 (or distilled R1) due to their popularity and effectiveness. Specifically, our chosen dynamic quantization reduces the MoE (Mixture of Experts) layers of R1 model to 2.51-bit, 1.73-bit, and 1.58-bit precision. Testing distillation with SFT, we evaluate the distilled R1 models that adopt LLaMA-70B (Meta AI) and Qwen-32B (Qwen et al., 2025) as students. Using SparseGPT, we prune the distilled R1 models to different sparsity levels. We select four complex reasoning benchmarks: mathematical, logical, temporal, and multihop reasoning.

In addition to assessing model performance on selected benchmarks, we study the behavior of compressed R1 models by measuring their test-time compute (number of tokens spent on each question). Specifically, we compare their performance on the shortest and longest outputs. Our major findings are:

- Parameter count has a much greater impact on LRM s’ knowledge memorization than on their reasoning capability. Therefore, both distillation and pruning are discouraged when the end task relies on models’ parametric knowledge. Otherwise, larger student models or low sparsity levels are recommended.
- In almost all cases, both R1 and its compressed variants achieve higher scores when they spend less compute during inference time. Shorter outputs tend to be more

accurate than longer ones across our benchmarks. This highlights the need to reduce verbosity to improve reasoning performance.

2 Related Work

2.1 Quantization

Quantization reduces the number of bits used to represent LLM weights, thereby lowering their precision (Srivastava et al., 2025). Recent survey (Zhu et al., 2024) categorizes quantization methodologies into quantization-aware training (QAT) and post-training quantization (PTQ). QAT requires retraining of model weights to recover performance loss during quantization while PTQ does not require retraining. Recent QAT includes LLM-QAT (Liu et al., 2024a) that adopts distillation to train a quantized LLM, BitDistiller (Du et al., 2024) that develops a self-distillation approach for the full-precision model to act as the teacher of its low-bit counterpart, BitNet (Wang et al., 2023) that proposes a 1-bit Transformer architecture for training LLMs from scratch, and OneBit (Xu et al., 2024) that quantizes LLM weight matrices to 1-bit from a knowledge transfer perspective.

PTQ is more popular in terms of the number of recent publications, because there is no retraining involved. For example, GPTQ (Frantar et al., 2022) is a one-shot weight quantization method that uses approximate second-order information, while AWQ (Lin et al., 2024) leverages activation distribution for finding the salient weight channels to skip. Other PTQ methods include weight-activation quantization (Shao et al., 2024; Yao et al., 2022; Liu et al., 2023) and KV cache quantization (Hooper et al., 2024; Liu et al., 2024b).

2.2 Distillation

Distillation involves two settings: black-box and white-box settings. For black-box setting, teacher model is typically a closed-source LLM and only the outputs of teacher are available for student model. For white-box setting, both weights and output distribution of the teacher model are available. Existing black-box distillation (Huang et al., 2024; Li et al., 2024b; Ho et al., 2023; Huang et al., 2022; Li et al., 2024a) prompts the teacher model to generate a training dataset for the student to learn. Specifically, researchers have started to distill OpenAI’s O1 model (Huang et al., 2024), which marks the beginning of LRM compression. White-box distillation allows the student model to learn from teacher’s knowledge representation. Works has been done to align the output distribution (Agarwal et al., 2024; Gu et al., 2024) or the hidden representation (Liang et al., 2023) between teacher and student models.

2.3 Pruning

There are unstructured and structured pruning. For unstructured pruning, individual weights are targeted, which leads to irregular sparsity structure. In contrast, structured pruning involves removing entire network components such as channels or layers (Zhang et al., 2024a). Unstructured pruning usually has better compression performance than structured pruning, while it is easier to achieve inference speedup via structured methods (Zhu et al., 2024). Recent unstructured pruning includes one-shot pruning (Frantar & Alistarh, 2023; Sun et al., 2023), global pruning that makes pruning decisions based on all layers (Bai et al., 2024), and domain-specific pruning (Zhang et al., 2024a). Structured pruning includes gradient-based (Xia et al., 2024; Ma et al., 2023) and non-gradient-based (Ashkboos et al., 2024) methods.

2.4 LRM

Trained with reinforcement learning, LRM extends LLMs with advanced reasoning mechanisms (Besta et al., 2025). Popular closed-source LRM are OpenAI’s o1-mini, o1 (OpenAI et al., 2024), and o3-mini. Open-source LRM include DeepSeek-R1 and QwQ-32B-Preview (Team, 2024). Since quantization, white-box distillation, and pruning methods

	#Questions	#Pass	Answer Type	Metric	Knowledge required?
AIME 2024	30	1	Integer	Accuracy	False
FOLIO	203	1	True/False/Uncertain	Accuracy	False
Temporal	250	1	(A)/(B)/(C)/(D)	Accuracy	False
MuSiQue	100	1	A few words	(EM, F1)	True

Table 1: Statistics of selected benchmarks: AIME 2024, FOLIO, Temporal Sequences of BIG-Bench Hard, and MuSiQue. We use exact match (EM) and F1 to measure model performance on MuSiQue.

require access to model weights, they are not suitable for closed-source LRMAs. Only black-box distillation will work on closed-source models.

3 Background

Scope We benchmark quantization, distillation, and pruning methodologies, because we consider them as major LLMs compression paradigms in this paper. Due to lower popularity, other methodologies that can potentially facilitate compression such as low-rank factorization (Sharma et al., 2024; Srebro & Jaakkola, 2003) are not elaborated. When selecting models for compression, we focus exclusively on LRMAs.

Recent efforts on LRMAs compression As reviewed in Section 2, compression on LRMAs (not LLMs) is relatively underexplored. Recently, Unsloth (Daniel Han & team, 2023) introduces dynamic quantization by dynamically opting not to quantize certain LLM weights. By combining four ideas (Yu et al., 2024; Dettmers et al., 2023; Ma et al., 2024; Gerganov, 2023), Unsloth quantizes DeepSeek-R1 into four variants: 2.51-, 2.22-, 1.73-, and 1.58-bit models. Here the number of bits (*e.g.*, 2.51-bit) represents the precision of MoE weights, which make up 88% of the total R1 weights. DeepSeek-R1 (Guo et al., 2025) also comes with several distilled models via black-box distillation.

Current bottlenecks Very few quantization or pruning works have demonstrated effectiveness on LRMAs, as LRMAs have only recently emerged. Moreover, evaluation of quantized or pruned LRMAs will be different from quantized or pruned LLMs. Current works evaluate quantization and pruning performance primarily using perplexity and simple end tasks, such as the EleutherAI evaluation harness (Gao et al., 2024) and commonsense reasoning. However, compressed LRMAs should be assessed on more complex reasoning tasks with varying difficulty levels. As for distillation, although recent distillation works tend to test on more challenging reasoning tasks (compared to other compression literature) such as GSM8K (Cobbe et al., 2021), it is unclear how the compression of LRMAs affects models' parametric knowledge and reasoning capability. Some of them do not comprehensively select diverse reasoning benchmarks. Our benchmarking framework aims to address these bottlenecks.

4 Experiment Setup

4.1 Reasoning Benchmarks

To address the evaluation bottleneck discussed in Section 3, we select four different reasoning benchmarks with varying levels of difficulty: AIME 2024 (Mathematical Association of America) for mathematical reasoning, FOLIO (Han et al., 2024) for logical reasoning, Temporal Sequences of BIG-Bench Hard (Suzgun et al., 2022) for temporal reasoning, and MuSiQue (Trivedi et al., 2022) for multihop reasoning. Table 1 summarizes their statistics. For each model, we do a single pass on every benchmark.

Unlike other benchmarks where we select all questions, we randomly sample 100 out of 1000 from MuSiQue, as it is much larger. This pool of 1000 questions is the same as SiReRAG (Zhang et al., 2025) and HippoRAG (Jimenez Gutierrez et al., 2024). Since

MuSiQue requires knowledge memorization besides multihop reasoning, it is popular under retrieval-augmented generation (RAG) setup. Instead of RAG, we follow a closed-book setting (directly prompting LRM to get final answers) to test models' parametric knowledge. In this case, unlike other benchmarks that test only reasoning capability, we use MuSiQue to evaluate both reasoning and knowledge memorization. Additional details of selected benchmarks are specified in Appendix A.

4.2 Selected Compression Methods

We decide to compress DeepSeek-R1 due to its large parameter count. For quantization, we select 2.51-, 1.73-, and 1.58-bit models¹ by Unsloth. These dynamically quantized models are among the most widely used quantization methods on R1. As for distillation, we select the two largest distilled models that accompany R1: DeepSeek-R1-Distill-Llama-70B² and DeepSeek-R1-Distill-Qwen-32B³. Since we do not find a pruning method designed for models with around 700B parameters and MoE layers (e.g., SparseGPT can only prune models with around 176B parameters), we use SparseGPT to prune our selected distilled models. In order to precisely analyze the collapse point of this pruning method, we run SparseGPT on each distilled model seven times to get 10%, 30%, 40%, 50%, 60%, 70%, and 80% sparsity levels.

4.3 Evaluation Metrics

Accuracy metric is used for AIME 2024, FOLIO, and Temporal Sequences. Since MuSiQue involves question answering and its answers are in a few words, we adopt exact match (EM) and F1 to measure question answering performance of different models. Without worrying inference speedup of selected compression methods, we focus on benchmarking their performance. The reason is that these methods run on different inference platforms. As a result, different optimization strategies might be adopted when we do inference, and it is hard to control the consistency of inference optimization across various platforms.

Our analysis of test-time compute involves computing the number of tokens spent on each question. We first preprocess each output by extracting reasoning tokens and sentences of final model prediction. Instead of relying on a specific tokenizer, we then treat words within each output as tokens for simplicity. We study whether there is any relationship between length and reasoning performance.

4.4 Implementation Details

Required by our quantized models, we run their inference on `llama.cpp`⁴. While our pruned and distilled models can be deployed on various inference platforms, we use vLLM (Kwon et al., 2023) for its fast inference. In order to comprehensively analyze performance change after compression, we also evaluate R1 on our reasoning benchmarks by using DeepSeek API. Aligning with DeepSeek-R1 report (Guo et al., 2025), we keep the same parameters for all models: maximum generation length is set to 32768, temperature is set to 0.6, and top-p value is set to 0.95.

5 Results and Analysis

Our results and analysis aim to answer the following research questions:

- **RQ 1:** How does each compression methodology compare against each other (5.1)?
- **RQ 2:** Do our selected compression methodologies collapse (5.2)?
- **RQ 3:** What is the recommended methodology to compress LRMs (5.1, 5.3, and 5.5)?

¹<https://huggingface.co/unsloth/DeepSeek-R1-GGUF>

²<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B>

³<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B>

⁴<https://github.com/ggml-org/llama.cpp>

Models			Accuracy				
Model	#Param	Compression	AIME 2024	FOLIO	Temporal	Avg	MuSiQue (EM, F1)
DeepSeek-R1	671B	-	73.3	76.4	99.6	83.1	(17.0, 27.51)
DeepSeek-R1	671B	2.51-bit	76.7	77.8	100.0	84.8	(17.0, 24.43)
DeepSeek-R1	671B	1.73-bit	66.7	78.3	99.6	81.5	(15.0, 22.11)
DeepSeek-R1	671B	1.58-bit	66.7	75.4	94.0	78.7	(14.0, 22.34)
R1-Distill-Llama	70B	Distillation	63.3	78.8	100.0	80.7	(13.0, 21.80)
R1-Distill-Llama	70B	Distillation & 50% sparse	26.7	70.9	97.2	64.9	(6.0, 12.75)
R1-Distill-Qwen	32B	Distillation	66.7	82.3	100.0	83.0	(1.0, 9.38)
R1-Distill-Qwen	32B	Distillation & 50% sparse	30.0	75.4	96.0	67.1	(3.0, 9.29)

Table 2: Overall results of R1 and its compressed variants. The avg column represents the average accuracy scores of AIME 2024, FOLIO, and Temporal. EM and F1 of MuSiQue are shown in tuples. We segment this table based on model family (e.g., R1-Distill-Llama).

- **RQ 4:** Is it a good sign when models spend extra amounts of test-time compute on certain questions (5.4)?

5.1 Overall Performance

The overall results of R1 and its compressed variants are presented in Table 2 and 3. We only show performance of pruned R1-Distill-Llama and pruned R1-Distill-Qwen under 50% sparsity in Table 2, as it is the default sparsity level of SparseGPT and many other works (Zhang et al., 2024a; Sun et al., 2023). Table 3 displays performance of our two distilled models under various sparsity levels.

Comparing Performance of Compression Methodologies In Table 2, 2.51-bit R1 achieves the highest average accuracy scores, surpassing even the original R1. We suspect that certain security mechanisms in DeepSeek may have lowered R1’s performance, as we often receive empty responses when using its API, requiring us to prompt R1 iteratively until valid responses are obtained. Both R1-Distill-Llama and R1-Distill-Qwen have competitive average accuracy scores. According to Table 3, we see that pruning also maintains good average accuracy scores when sparsity levels are low (e.g., 30% sparsity of R1-Distill-Llama). On MuSiQue dataset, the 2.51-bit model has the same EM as R1, with a slight decrease in F1. According to scores of the closed-book setting on GPT-4o (Zhang et al., 2025), all quantized models achieve relatively strong performance on MuSiQue, with higher EM and F1 than distilled and pruned models. Therefore, this 2.51-bit model has the best overall performance than other compression methods.

Comparing Benchmark Difficulties We can make judgment about the difficulties of our accuracy-based benchmarks through Table 2. By comparing the scores of AIME 2024, FOLIO, and Temporal, we see all models struggle more on AIME 2024. This indicates that AIME 2024 is more difficult than the other two benchmarks. MuSiQue is also a difficult benchmark in terms of knowledge requirement, because its scores in Table 2 are much lower than RAG setup (Zhang et al., 2025; Jimenez Gutierrez et al., 2024). This suggests that existing LRMs lack sufficient knowledge for knowledge-intensive tasks, making RAG a more suitable approach. We demonstrate a diverse selection of reasoning benchmarks with varying difficulty levels, which is important for comprehensive evaluation.

Comparing Distilled Models based on LLaMA and Qwen When focusing on accuracy-based benchmarks, we see that R1-Distill-Qwen delivers an average 2.3% improvement over R1-Distill-Llama. This demonstrates that Qwen-32B has stronger reasoning capability than LLaMA-70B, which also aligns with the evaluation results of DeepSeek-R1 report (Guo et al., 2025). However, R1-Distill-Qwen scores significantly lower on MuSiQue, highlighting its inability to memorize detailed knowledge.

5.2 Collapse Point of Each Compression Method

We investigate whether LRMs degrade as they undergo increasing levels of compression. In Table 2, the performance of dynamically quantized models steadily declines as we

Models			Accuracy				
Model	#Param	Sparsity	AIME 2024	FOLIO	Temporal	Avg	MuSiQue (EM, F1)
R1-Distill-Llama	70B	0%	63.3	78.8	100.0	80.7	(13.0, 21.80)
R1-Distill-Llama	70B	10%	60.0	81.3	99.6	80.3	(12.0, 21.69)
R1-Distill-Llama	70B	30%	63.3	79.3	99.6	80.7	(14.0, 21.40)
R1-Distill-Llama	70B	40%	56.7	73.9	98.8	76.8	(6.0, 13.79)
R1-Distill-Llama	70B	50%	26.7	70.9	97.2	64.9	(6.0, 12.75)
R1-Distill-Llama	70B	60%	0.0	65.0	95.6	53.5	(0.0, 6.42)
R1-Distill-Llama	70B	70%	0.0	49.8	15.6	21.8	(0.0, 2.23)
R1-Distill-Llama	70B	80%	0.0	11.8	12.4	8.1	(0.0, 0.94)
R1-Distill-Qwen	32B	0%	66.7	82.3	100.0	83.0	(1.0, 9.38)
R1-Distill-Qwen	32B	10%	70.0	81.3	100.0	83.8	(5.0, 13.19)
R1-Distill-Qwen	32B	30%	56.7	81.3	100.0	79.3	(1.0, 10.47)
R1-Distill-Qwen	32B	40%	53.3	78.3	100.0	77.2	(2.0, 10.16)
R1-Distill-Qwen	32B	50%	30.0	75.4	96.0	67.1	(3.0, 9.29)
R1-Distill-Qwen	32B	60%	0.0	65.0	87.2	50.7	(0.0, 4.13)
R1-Distill-Qwen	32B	70%	0.0	32.5	19.6	17.4	(0.0, 1.72)
R1-Distill-Qwen	32B	80%	0.0	8.7	2.0	3.6	(0.0, 1.29)

Table 3: Performance of our two distilled models under various sparsity levels. The avg column represents the average accuracy scores of AIME 2024, FOLIO, and Temporal.

move from 2.51-bit to 1.58-bit. However, we do not observe a clear collapse point, as performance declines only slightly when transitioning to a lower bit-size tier. As for distillation, comparing R1 with R1-Distill-Llama, we also do not observe a clear collapse point, although the performance drop on MuSiQue is significant. However, we see that R1-Distill-Qwen collapses on MuSiQue. With EM close to zero, it almost loses its ability on MuSiQue during distillation.

Comparing distilled models with their sparsified variants in Table 3, we find that pruned models collapse on all benchmarks at certain sparsity levels. Interestingly, their collapse point correlates to the difficulty of the benchmark. For example, on AIME 2024, R1-Distill-Llama collapses between 40% and 50% sparsity, since its performance drops by more than half. However, its collapse points on FOLIO and Temporal are roughly between 60% and 70% sparsity, which occur much later than AIME 2024. The correlation between collapse point and benchmark difficulty can also be seen on sparsified R1-Distill-Qwen. The early occurrence of collapse on AIME 2024 demonstrates that it is one of the most challenging datasets, as existing pruning works typically observe model collapse after 50% sparsity (Zhang et al., 2024a; Bai et al., 2024). Moreover, since R1-Distill-Qwen already collapses on MuSiQue, pruning it becomes meaningless for this dataset, as we observe very low scores for all sparsified R1-Distill-Qwen models on MuSiQue.

5.3 Compression Impact on Knowledge and Reasoning

As discussed above, although Qwen-32B shows stronger reasoning capability than LLaMA-70B, it has significantly lower EM and F1 scores on MuSiQue. Because MuSiQue requires knowledge memorization under the closed-book setting, the smaller parameter count of Qwen-32B puts itself at a disadvantaged position. In other words, models' parameter count affects knowledge more than reasoning. When a compression method aggressively removes the weights of an LRM, it is expected that the model's knowledge will be more seriously affected. This phenomenon can also be seen on our quantized and pruned models. Since quantization preserves parameter count and our analysis above shows that the quantized models still retain competitive reasoning capability, it is not surprising that even the 1.58-bit model outperforms other distilled and pruned LRMs on MuSiQue. In addition, we notice that pruned R1-Distill-Llama collapses between 30% and 40% sparsity on MuSiQue, which is even earlier than on AIME 2024. Since MuSiQue is difficult due to its knowledge requirement, we see that pruning hurts LRMs' knowledge memorization more than quantization.

Models		AIME 2024			FOLIO			MuSiQue		
Model	Compression	Short	Long	Ratio	Short	Long	Ratio	Short	Long	Ratio
DeepSeek-R1	-	88.9	33.3	5.3	83.3	63.3	8.0	(30.0, 42.8)	(3.3, 10.0)	6.4
DeepSeek-R1	2.51-bit	100.0	33.3	4.9	85.0	71.7	6.5	(33.3, 41.9)	(0.0, 6.9)	7.7
DeepSeek-R1	1.73-bit	88.9	22.2	4.4	86.7	73.3	4.9	(30.0, 41.6)	(10.0, 23.5)	6.9
DeepSeek-R1	1.58-bit	77.8	44.4	3.8	80.0	65.0	5.0	(30.0, 41.5)	(10.0, 16.6)	5.9
R1-Distill-Llama	Distillation	88.9	11.1	5.4	80.0	80.0	4.5	(26.7, 38.0)	(6.7, 18.4)	4.0
R1-Distill-Llama	Distillation & 10% sparse	100.0	0.0	6.6	85.0	78.3	4.9	(20.0, 29.6)	(6.7, 13.7)	7.4
R1-Distill-Llama	Distillation & 30% sparse	88.9	11.1	5.3	85.0	76.7	5.0	(26.7, 36.9)	(3.3, 12.5)	8.8
R1-Distill-Qwen	Distillation	100.0	11.1	7.1	86.7	75.0	6.9	(16.7, 24.1)	(16.7, 24.7)	7.1
R1-Distill-Qwen	Distillation & 10% sparse	88.9	33.3	4.8	83.3	75.0	5.5	(23.3, 36.7)	(3.3, 12.8)	8.1
R1-Distill-Qwen	Distillation & 30% sparse	88.9	11.1	6.8	86.7	78.3	7.8	(26.7, 40.7)	(3.3, 9.1)	8.8

Table 4: Analysis of test-time compute when selecting the shortest and longest 30% of responses output by each model on each benchmark. “Short” column contains performance scores of the shortest 30% of outputs from a model, while “long” column contains scores of the longest 30% of outputs. We compare the scores between “Short” and “Long” for every model and benchmark, and mark the best scores in **bold**. “Ratio” column represents the ratio of the average length (in number of tokens) of the longest 30% to that of the shortest 30%.

Models		AIME 2024			FOLIO			MuSiQue		
Model	Compression	Short	Long	Ratio	Short	Long	Ratio	Short	Long	Ratio
DeepSeek-R1	-	100.0	16.7	7.1	82.5	57.5	10.9	(30.0, 42.4)	(5.0, 12.5)	8.5
DeepSeek-R1	2.51-bit	100.0	33.3	6.6	87.5	62.5	8.6	(40.0, 50.3)	(0.0, 4.5)	10.7
DeepSeek-R1	1.73-bit	100.0	33.3	5.9	87.5	67.5	6.3	(30.0, 43.7)	(5.0, 15.7)	9.3
DeepSeek-R1	1.58-bit	100.0	33.3	5.6	80.0	72.5	6.2	(35.0, 46.3)	(15.0, 20.0)	7.6
R1-Distill-Llama	Distillation	83.3	16.7	7.0	82.5	75.0	5.7	(35.0, 45.2)	(5.0, 9.2)	5.3
R1-Distill-Llama	Distillation & 10% sparse	100.0	0.0	9.2	90.0	75.0	6.3	(20.0, 28.5)	(5.0, 11.7)	10.7
R1-Distill-Llama	Distillation & 30% sparse	100.0	16.7	7.1	90.0	75.0	6.4	(20.0, 31.0)	(0.0, 7.3)	13.0
R1-Distill-Qwen	Distillation	100.0	0.0	9.8	87.5	75.0	9.0	(20.0, 28.7)	(20.0, 24.2)	10.4
R1-Distill-Qwen	Distillation & 10% sparse	83.3	33.3	6.1	85.0	75.0	7.2	(30.0, 36.2)	(0.0, 8.4)	12.0
R1-Distill-Qwen	Distillation & 30% sparse	100.0	0.0	8.9	87.5	75.0	11.0	(30.0, 36.6)	(5.0, 11.2)	13.4

Table 5: Analysis of test-time compute when selecting the shortest and longest 20% of responses output by each model on each benchmark. Refer to Table 4 for the meaning of each column, except that we select shortest and longest 20% instead.

Based on the finding that parameter count has a much greater impact on knowledge than reasoning, we recommend quantization as the LRM’s compression method when some levels of models’ parametric knowledge are desired. Since both distillation and pruning involve removing weights, we recommend to choose a large student model or a low sparsity when the end task requires knowledge memorization. On the other hand, a model with strong reasoning capability does not need to be extremely large. As Qwen-32B is a stronger reasoner than LLaMA-70B, smaller models also have the potential to outperform larger ones on various reasoning tasks.

5.4 Test-time Compute

We study the behavior of R1 and its compressed variants by measuring their test-time compute. Table 4 shows the analysis of test-time compute when we select the shortest and longest 30% of responses output by each model on each benchmark. We observe that shorter model outputs consistently yield better performance across three reasoning benchmarks, with only an exception on MuSiQue. Regardless of whether an LRM is compressed, if it generates significantly more tokens for a question than other questions in the same dataset, the answer is likely to be incorrect. We exclude Temporal here, because many of the compressed R1 models achieve close to 100% accuracy. The length ratios between the longest and the shortest 30% are typically greater than 4, which indicates nontrivial length differences among model outputs.

Similarly, Table 5 shows the selection of the shortest and the longest 20% of model responses. Compared to Table 4, as we move toward the extreme, the performance gap between the shortest and longest responses becomes even larger. Both R1 and its compressed variants

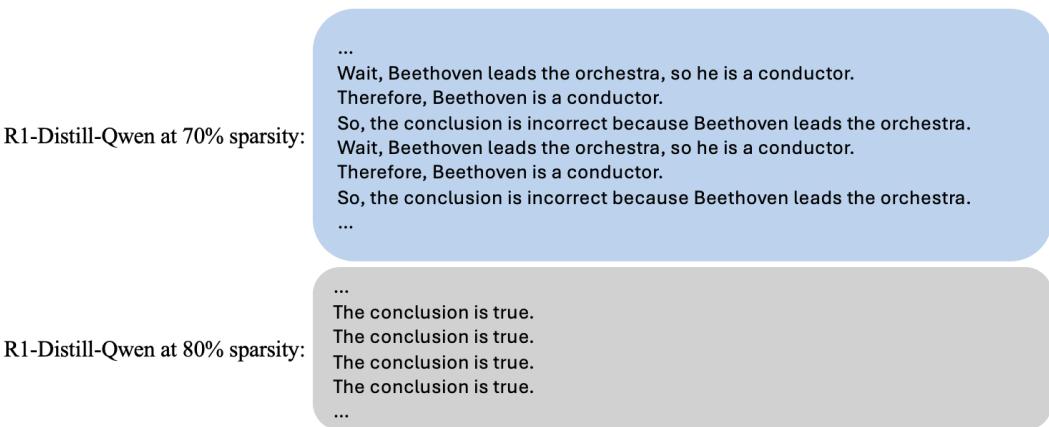


Figure 2: Two examples of the case when a model collapses and keeps repeating itself. These are two outputs (for a FOLIO question) from R1-Distill-Qwen at either 70% or 80% sparsity levels.

achieve higher scores when they spend less compute during test time. After manually checking some long responses, we notice that longer outputs tend to be more verbose and involve more backtracking in reasoning. This finding demonstrates the need to reduce verbosity to improve reasoning performance, which aligns with recent research (Zhang et al., 2024b).

5.5 Case Study

As discussed in Section 5.2, pruned models collapse on all benchmarks at certain sparsity levels. We identify a common phenomenon when a model collapses: it repeatedly generates a sentence or a chunk until reaching the maximum generation length. We show two examples of this phenomenon in Figure 2. For brevity, we omit the beginning and the end of outputs.

In both examples, the pruned models are repeating themselves without pushing their reasoning chains forward, which is a signal of model collapse. When R1-Distill-Qwen is pruned to 70% sparsity, we see that it can still organize a few sentences (e.g., a chunk) to repeat. But when it is pruned to 80% sparsity, it only repeats a simple sentence. This decline of linguistic capability is common when models are pruned to high sparsities. Therefore, aggressively pruning LRM s requires careful consideration.

6 Conclusion and Future Directions

In this paper, we benchmark compressed LRMs on four complex reasoning tasks: AIME 2024, FOLIO, Temporal Sequences of BIG-Bench Hard, and MuSiQue. We evaluate models that are quantized, distilled, and pruned. Through MuSiQue, we find that parameter count has a much greater impact on LRMs' knowledge than reasoning capability, which can potentially guide compression decision. Based on our study of test-time compute, we also find that shorter outputs of R1 and compressed LRMs generally achieve better performance than longer outputs. This finding highlights the need to reduce verbosity in reasoning chains.

Since compression on LRMs is relatively underexplored, future directions involve designing novel methodologies to specifically compress LRMs. The evaluation of compressed LRMs should also place greater emphasis on diverse reasoning tasks. Since DeepSeek-R1 has already been compressed through quantization and distillation, the question remains how to prune such a large model.

References

- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=3zKtaqxLhW>.
- Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefer, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=vXxardq6db>.
- Guangji Bai, Yijiang Li, Chen Ling, Kibaek Kim, and Liang Zhao. SparseLLM: Towards global pruning of pre-trained language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=oXHyYHp4Zb>.
- Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houlston, Tomasz Sternal, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, Łukasz Flis, Hannes Eberhard, Hubert Niewiadomski, and Torsten Hoefer. Reasoning language models: A blueprint, 2025. URL <https://arxiv.org/abs/2501.11223>.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.
- Michael Han Daniel Han and Unsloth team. Unsloth, 2023. URL <http://github.com/unslotha/unsloth>.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL <https://arxiv.org/abs/2412.19437>.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

DaYou Du, Yijia Zhang, Shijie Cao, Jiaqi Guo, Ting Cao, Xiaowen Chu, and Ningyi Xu. Bitdistiller: Unleashing the potential of sub-4-bit llms via self-distillation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 102–116, 2024.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pp. 10323–10337. PMLR, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. GPTQ: Accurate post-training compression for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL <https://zenodo.org/records/12608602>.

Georgi Gerganov. Github-ggerganov/llama.cpp: Llm inference in c/c++—github. com, 2023.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models, 2024. URL <https://arxiv.org/abs/2306.08543>.

Daya Guo, Dejian Yang, Huawei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1229. URL <https://aclanthology.org/2024.emnlp-main.1229/>.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.830. URL <https://aclanthology.org/2023.acl-long.830/>.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *Advances in Neural Information Processing Systems*, 37:1270–1303, 2024.

Yukun Huang, Yanda Chen, Zhou Yu, and Kathleen McKeown. In-context learning distillation: Transferring few-shot learning ability of pre-trained language models. *arXiv preprint arXiv:2212.10670*, 2022.

Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? *arXiv preprint arXiv:2411.16489*, 2024.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37:59532–59569, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Ming Li, Lichang Chen, Juhai Chen, Shuai He, Jiuxiang Gu, and Tianyi Zhou. Selective reflection-tuning: Student-selected data recycling for LLM instruction-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 16189–16211, Bangkok, Thailand and virtual meeting, August 2024a. Association for Computational Linguistics. URL <https://aclanthology.org/2024.findings-acl.958>.

Shiyang Li, Jianshu Chen, yelong shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhui Chen, and Xifeng Yan. Explanations from large language models make small reasoners better. In *2nd Workshop on Sustainable AI*, 2024b. URL <https://openreview.net/forum?id=rH8ZUcfL9r>.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: task-aware layer-wise distillation for language model compression. In *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org, 2023.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100, 2024.

Shih-yang Liu, Zechun Liu, Xijie Huang, Pingcheng Dong, and Kwang-Ting Cheng. LLM-FP4: 4-bit floating-point quantized transformers. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 592–605, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.39. URL <https://aclanthology.org/2023.emnlp-main.39>.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. Llm-qat: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 467–484, 2024a.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Forty-first International Conference on Machine Learning*, 2024b. URL <https://openreview.net/forum?id=L057s2Rq80>.

Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era of 1-bit llms: All large language models are in 1.58 bits, 2024. URL <https://arxiv.org/abs/2402.17764>.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*, 2023.

Mathematical Association of America. MAA Invitational Competitions. URL <https://maa.org/maa-invitational-competitions/>. Accessed: 2025-03-19.

Meta AI. Meta Llama 3.3 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md. Accessed: 2025-03-19.

Gianluca Mondillo, Simone Colosimo, Alessandra Perrotta, Vittoria Frattolillo, and Mariapia Masino. Comparative evaluation of advanced ai reasoning models in pediatric clinical decision support: Chatgpt o1 vs. deepseek-r1. *medRxiv*, pp. 2025-01, 2025.

OpenAI, ;, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Lederc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL <https://arxiv.org/abs/2412.16720>.

Qwen, ;, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang

Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omnipoint: Omnidirectionally calibrated quantization for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=8Wuvhh0LYW>.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=ozX92bu8VA>.

Nathan Srebro and Tommi Jaakkola. Weighted low-rank approximations. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 720–727, 2003.

Gaurav Srivastava, Shuxiang Cao, and Xuan Wang. Towards reasoning ability of small language models. *arXiv preprint arXiv:2502.11569*, 2025.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022. URL <https://arxiv.org/abs/2210.09261>.

Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown. *Hugging Face*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL <https://arxiv.org/abs/2307.09288>.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition, 2022. URL <https://arxiv.org/abs/2108.00573>.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. Bitnet: Scaling 1-bit transformers for large language models, 2023. URL <https://arxiv.org/abs/2310.11453>.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared LLaMA: Accelerating language model pre-training via structured pruning. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=09i0dae0zp>.

Yuzhuang Xu, Xu Han, Zonghan Yang, Shuo Wang, Qingfu Zhu, Zhiyuan Liu, Weidong Liu, and Wanxiang Che. Onebit: Towards extremely low-bit large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=ZwiG9KjfHV>.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in Neural Information Processing Systems*, 35:27168–27183, 2022.

Mengxia Yu, De Wang, Qi Shan, Colorado Reed, and Alvin Wan. The super weight in large language models, 2024. URL <https://arxiv.org/abs/2411.07191>.

Nan Zhang, Yanchi Liu, Xujiang Zhao, Wei Cheng, Runxue Bao, Rui Zhang, Prasenjit Mitra, and Haifeng Chen. Pruning as a domain-specific LLM extractor. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1417–1428, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.91. URL <https://aclanthology.org/2024.findings-naacl.91/>.

Nan Zhang, Prafulla Kumar Choubey, Alexander Fabbri, Gabriel Bernadett-Shapiro, Rui Zhang, Prasenjit Mitra, Caiming Xiong, and Chien-Sheng Wu. SireRAG: Indexing similar and related information for multihop reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=yp95goUAT1>.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Yusen Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. Verbosity ≠ veracity: Demystify verbosity compensation behavior of large language models, 2024b. URL <https://arxiv.org/abs/2411.07858>.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577, 2024. doi: 10.1162/tacl_a_00704. URL <https://aclanthology.org/2024.tacl-1.85/>.

A Additional Details of Reasoning Benchmarks

AIME 2024⁵ (parts I and II) represents top match challenges, and its answers are integers. FOLIO⁶ requires logical deductions to determine whether the provided conclusion is true, false, or uncertain based on premise. In Temporal Sequences⁷, models are asked to use a provided timeline to determine what time a person might be free to perform another activity. Since each of its questions comes with four options, we expect our models to output the index (the letter) of the selected option. The answers of MuSiQue questions are in a few words.

⁵https://huggingface.co/datasets/Maxwell-Jia/AIME_2024

⁶<https://huggingface.co/datasets/yale-nlp/FOLIO>

⁷https://github.com/suzgunmirac/BIG-Bench-Hard/blob/main/bbh/temporal_sequences.json