When Reasoning Meets Compression: Understanding the Effects of LLMs Compression on Large Reasoning Models

Nan Zhang, Eugene Kwek, Yusen Zhang, Ngoc-Hieu Nguyen, Prasenjit Mitra, Rui Zhang

Overview

- Compression methods, including quantization, distillation, and pruning, improve the computational efficiency of large reasoning models (LRMs).

- However, existing studies either fail to sufficiently compare all three compression methods on LRMs or lack in-depth interpretation analysis.

- To understand LLMs compression, we investigate how the reasoning capabilities of LRMs are compromised during compression, through performance benchmarking and mechanistic interpretation.

Our Benchmarking and Interpretation Pipeline

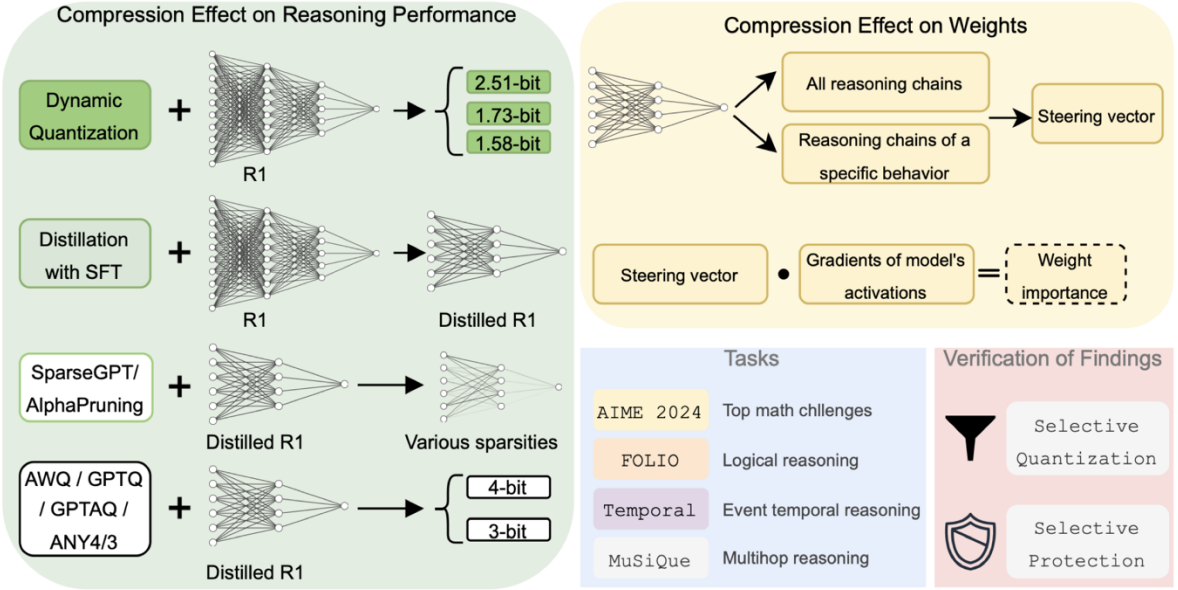


Figure 1: An overview of our pipeline. On the left, we benchmark compressed R1 variants on various reasoning tasks. On the right, by computing weight importance towards a specific reasoning behavior (a dot product of the steering vector and gradients with respect to an LRM's activations), we study the compression effects on individual weight matrices. We empirically verify our findings on weight importance by selectively quantizing or protecting a module to test its importance.

Performance Benchmarking

- Benchmarked various compressed LRMs over 3 passes.

- Selected reasoning benchmarks with varying levels of difficulty.
  - **AIME 2024**: Top math challenges.
  - **FOLIO**: Logical reasoning.
  - **Temporal Sequences of Big-Bench Hard**: Event temporal reasoning.
  - **MuSiQue**: Multihop reasoning.

Mechanistic Interpretation

- **Difference of Means**: compute the numerical representation in activation space of each reasoning behavior.

$$\mathbf{u}^c_{m\ell} = \frac{1}{|\mathcal{D}_+|} \sum_{s^c_i \in \mathcal{D}_+} \bar{\mathbf{a}}^c_{m\ell}(s^c_i) - \frac{1}{|\mathcal{D}_-|} \sum_{s_j \in \mathcal{D}_-} \bar{\mathbf{a}}_{m\ell}(s_j), \quad \text{with} \quad \bar{\mathbf{a}}^c_{m\ell}(s^c_i) = \frac{1}{|s^c_i|} \sum_{t \in s^c_i} \mathbf{a}_{m\ell}(t)$$

- **Attribution Patching**: incorporate gradient to find the causally relevant LRMs components with respect to each reasoning behavior.

$$\mathbf{I}^c_{m\ell} \approx \frac{1}{|\mathcal{D}_+|} \left| \sum_{s^c_i \in \mathcal{D}_+} \left(\tilde{\mathbf{u}}^c_{m\ell}\right)^\top \frac{\partial}{\partial \mathbf{a}_{m\ell}} \mathcal{L}(s^c_i) \right|$$

Benchmarking Results

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	65.6	**79.8**	**99.9**	**81.8**	(**13.3**, **21.57**)
R1-Distill-Llama	70B	Distillation & 50% SparseGPT	23.3	71.6	97.6	64.2	(6.7, 13.49)
R1-Distill-Llama	70B	Distillation & 50% AlphaPruning	26.7	74.2	97.7	66.2	(5.3, 12.39)
R1-Distill-Llama	70B	Distillation & 4-bit AWQ	63.4	78.5	99.3	80.4	(10.7, 19.23)
R1-Distill-Llama	70B	Distillation & 4-bit GPTQ	**66.7**	77.0	**99.9**	81.2	(10.3, 18.17)
R1-Distill-Llama	70B	Distillation & 4-bit GPTAQ	64.4	78.8	99.6	80.9	(12.0, **21.57**)
R1-Distill-Llama	70B	Distillation & 3-bit GPTQ	46.7	71.8	99.3	72.6	(4.7, 11.92)
R1-Distill-Llama	70B	Distillation & 3-bit GPTAQ	54.4	77.3	99.7	77.1	(5.7, 13.21)
R1-Distill-Qwen	32B	Distillation	64.4	82.3	**99.9**	82.2	(2.7, 10.95)
R1-Distill-Qwen	32B	Distillation & 50% SparseGPT	25.6	75.1	97.9	66.2	(2.3, 9.01)
R1-Distill-Qwen	32B	Distillation & 4-bit AWQ	67.8	82.3	99.1	**83.1**	(3.3, 10.28)
R1-Distill-Qwen	32B	Distillation & 4-bit GPTQ	**68.9**	80.6	99.6	83.0	(4.0, 11.78)
R1-Distill-Qwen	32B	Distillation & 4-bit GPTAQ	63.3	81.5	99.7	81.5	(2.7, 11.88)
R1-Distill-Qwen	32B	Distillation & 4-bit ANY4	**68.9**	78.0	99.7	82.2	(**5.7**, **12.68**)
R1-Distill-Qwen	32B	Distillation & 3-bit GPTQ	44.4	74.2	98.9	72.5	(4.0, 11.55)
R1-Distill-Qwen	32B	Distillation & 3-bit GPTAQ	45.6	77.5	99.5	74.2	(2.3, 9.18)
R1-Distill-Qwen	32B	Distillation & 3-bit ANY3	53.3	**82.6**	**99.9**	78.6	(3.7, 10.27)
R1-Distill-Llama	8B	Distillation	42.2	**71.9**	81.5	65.2	(0.0, 4.43)
R1-Distill-Llama	8B	Distillation & 30% AlphaPruning	41.1	68.9	82.1	64.1	(0.3, 4.51)
R1-Distill-Llama	8B	Distillation & 50% AlphaPruning	6.7	61.7	79.6	49.3	(0.0, 2.95)
R1-Distill-Llama	8B	Distillation & 4-bit AWQ	**47.8**	68.0	84.0	**66.6**	(**0.3**, **5.05**)
R1-Distill-Llama	8B	Distillation & 4-bit GPTQ	42.2	66.2	65.9	58.1	(**0.3**, 4.68)
R1-Distill-Llama	8B	Distillation & 4-bit GPTAQ	40.0	66.4	69.3	58.6	(0.0, 3.73)
R1-Distill-Llama	8B	Distillation & 4-bit ANY4	41.1	68.5	**88.7**	66.1	(0.0, 3.54)
R1-Distill-Llama	8B	Distillation & 3-bit GPTQ	11.1	65.0	67.3	47.8	(0.0, 2.89)
R1-Distill-Llama	8B	Distillation & 3-bit GPTAQ	7.8	65.5	57.2	43.5	(0.0, 3.45)
R1-Distill-Llama	8B	Distillation & 3-bit ANY3	3.3	50.1	34.9	29.4	(0.7, 2.35)
R1-Distill-Qwen	7B	Distillation	46.7	**78.0**	75.6	**66.8**	(0.0, 3.57)
R1-Distill-Qwen	7B	Distillation & 4-bit AWQ	46.6	75.5	74.9	65.7	(0.0, 3.14)
R1-Distill-Qwen	7B	Distillation & 4-bit GPTQ	38.9	72.9	70.3	60.7	(**1.0**, **4.27**)
R1-Distill-Qwen	7B	Distillation & 4-bit GPTAQ	**47.8**	74.4	67.7	63.3	(0.0, 3.96)
R1-Distill-Qwen	7B	Distillation & 4-bit ANY4	**47.8**	75.6	**77.1**	**66.8**	(0.0, 3.05)
R1-Distill-Qwen	7B	Distillation & 3-bit GPTQ	17.8	65.7	31.7	38.4	(0.0, 3.12)
R1-Distill-Qwen	7B	Distillation & 3-bit GPTAQ	24.4	64.5	48.7	45.9	(0.0, 3.06)
R1-Distill-Qwen	7B	Distillation & 3-bit ANY3	32.2	69.3	30.1	43.9	(0.0, 3.89)

Distillation Effect

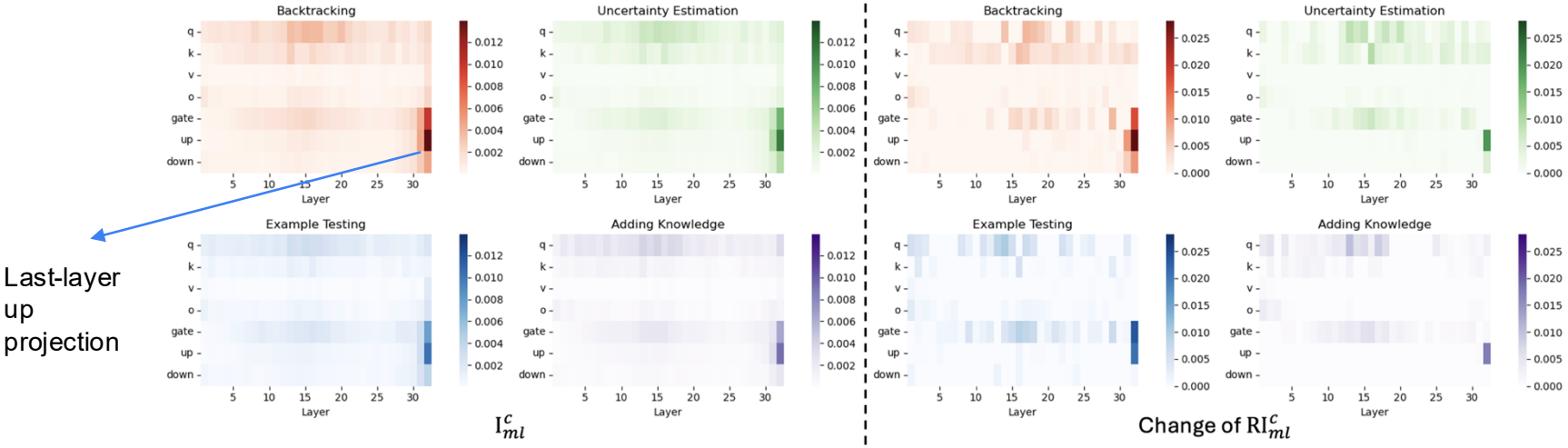
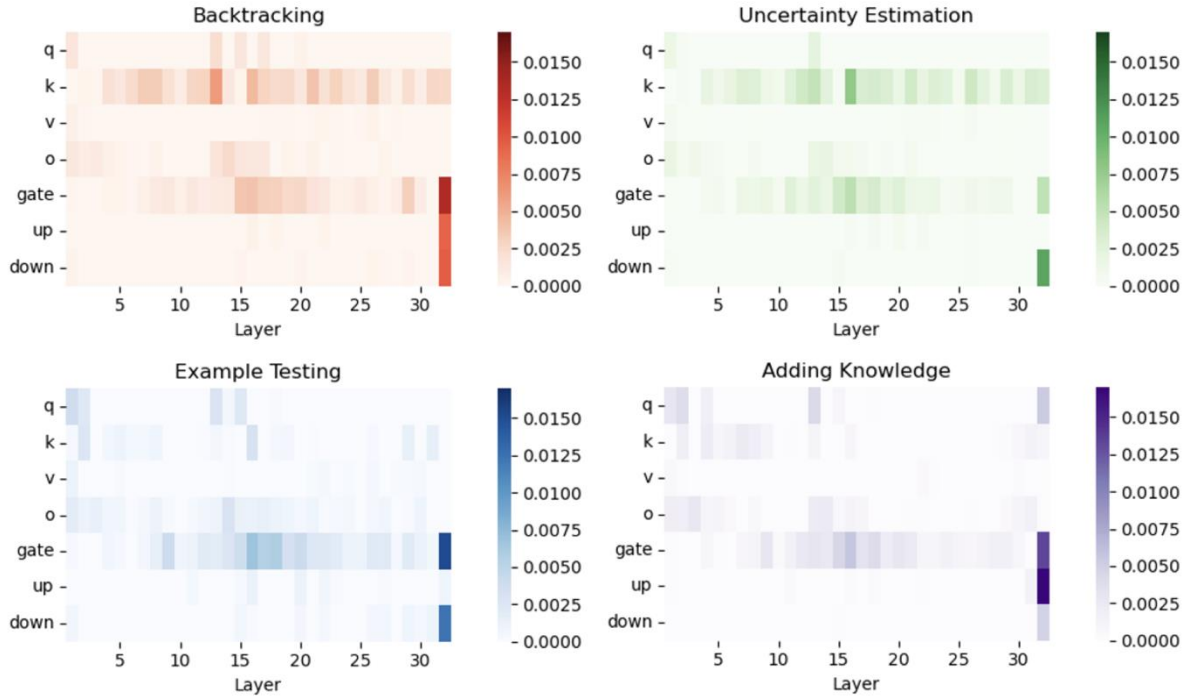


Figure 2: $\mathbf{I}^c_{m\ell}$ of `DeepSeek-R1-Distill-Llama-8B` (**left**) and change of $\mathbf{RI}^c_{m\ell}$ from `DeepSeek-R1-Distill-Llama-8B` to `Llama-3.1-8B` (**right**). Each heatmap displays scores of importance (or importance shift) of every module at each layer, providing a fine-grained analysis of weight contributions to the corresponding reasoning capability. On the right, increases in $\mathbf{RI}^c_{m\ell}$ are set to 0, as they only offset decreases elsewhere as discussed in 2.3. Every cluster of 4 side-by-side heatmaps (including those displayed below) follow the same scaling to show the precise magnitude of each weight module.

Quantization Effect



Common outliers: Last-layer modules and MLP gate projections

Figure 3: Change of $\mathbf{RI}^{c}_{m\ell}$ from `DeepSeek-R1-Distill-Llama-8B` to its 4-bit AWQ variant.

Pruning Effect

Figure 10: Change of $\mathbf{RI}^{c}_{m\ell}$ from `DeepSeek-R1-Distill-Llama-8B` to its AlphaPruning variant (at 30% sparsity).

Similar findings as quantization effect: outliers at last-layer modules and MLP gate projections

Generalizability and Empirical Verification

- More analysis on non-R1 families (e.g., gpt-oss), quantization, and pruning.

- Selective quantization and selective protection
  - Our selective protection boosts 3-bit AWQ on all benchmarks by protecting only 2.2% of all weights, with an average accuracy improvement of 6.57%.

Key Findings

- Weight count has a greater impact on LRMs' knowledge memorization than reasoning, highlighting the risks of pruning and distillation.
- The MLP up projection in the final layer of R1 distilled LRMs is one of the most important components, offering a new perspective on locating critical weights — a fundamental problem in model compression.
- Current quantization and pruning methods overly compress the final-layer modules and MLP gate projections.

Questions?

- Paper: https://arxiv.org/abs/2504.02010
- Interpretation code: https://github.com/psunlpgroup/Compression-Effects
- Email: njz5124@psu.edu
- Homepage: https://zn1010.github.io