JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

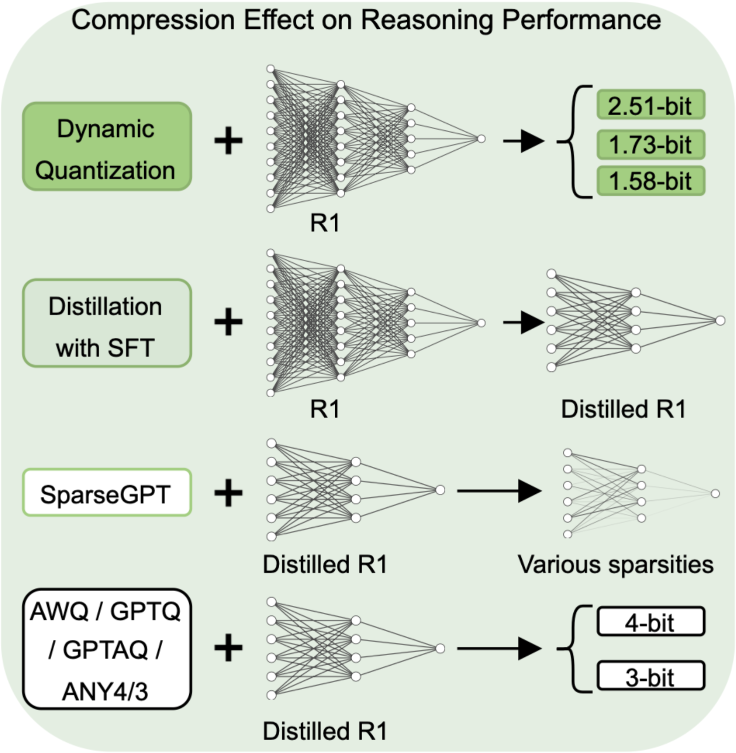
Can we preserve reasoning integrity under efficiency and precision constraints?

Dengjia Zhang, Elisabeth Fittschen

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

Zhang, et al.

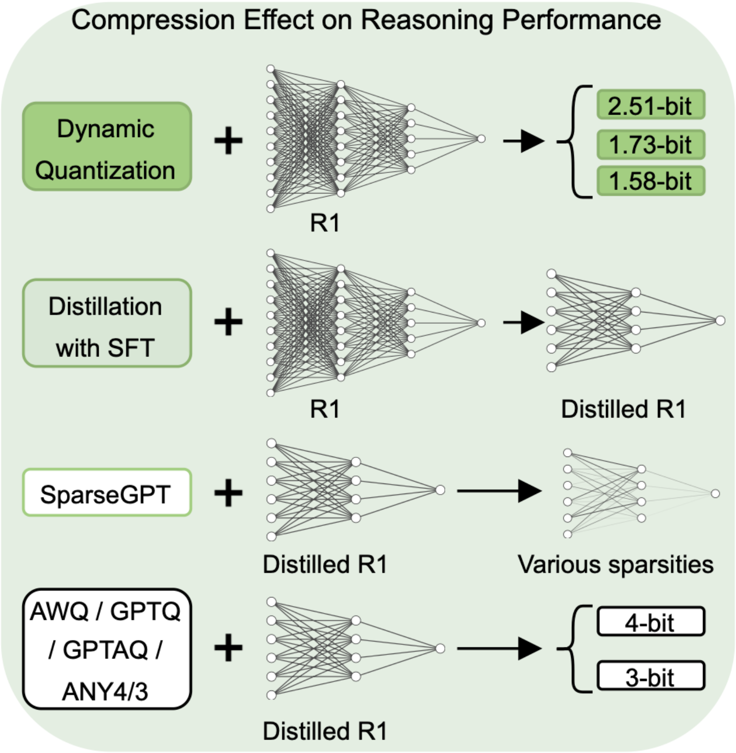
Big Picture



Two parts:

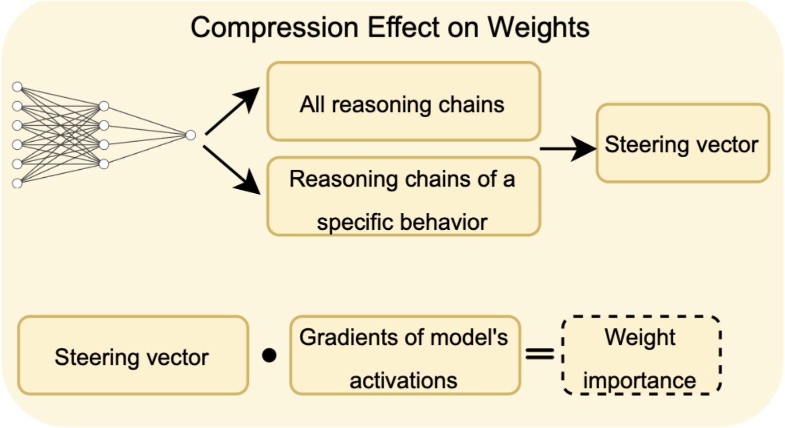
- Benchmarking compression methods

Big Picture

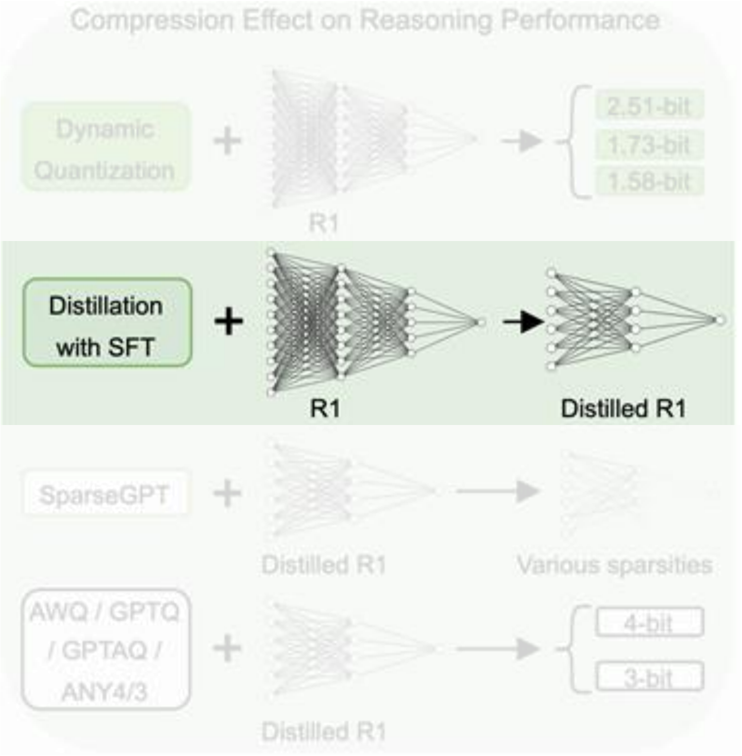


Two parts:

- Benchmarking compression methods
- Estimating weight importance for reasoning



Background

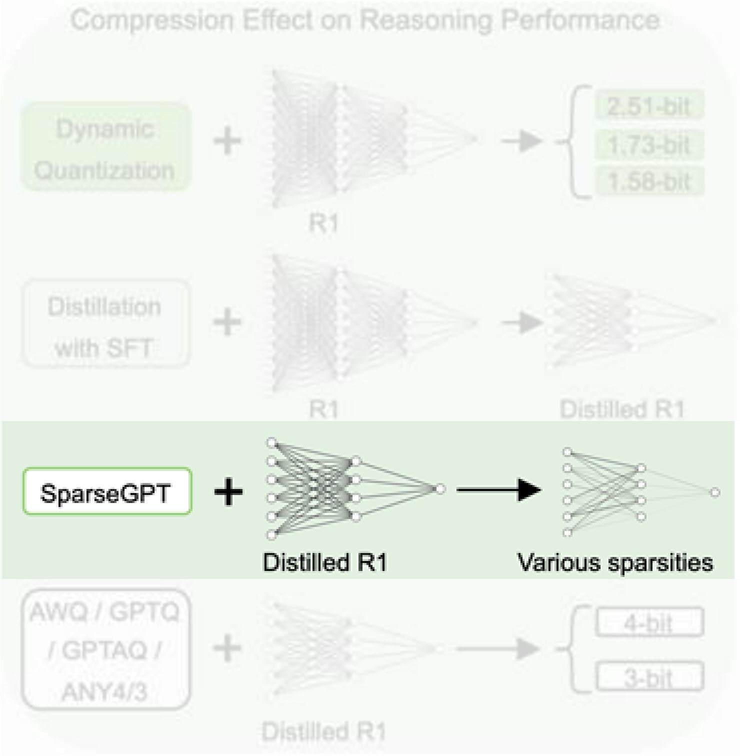


Distillation:

- Trainer/Student setup
- Teacher logit distribution is used in addition to regular token prediction loss

	Model Structure	Weights
Distillation	❌	❌
Pruning		
Quantization		

Background

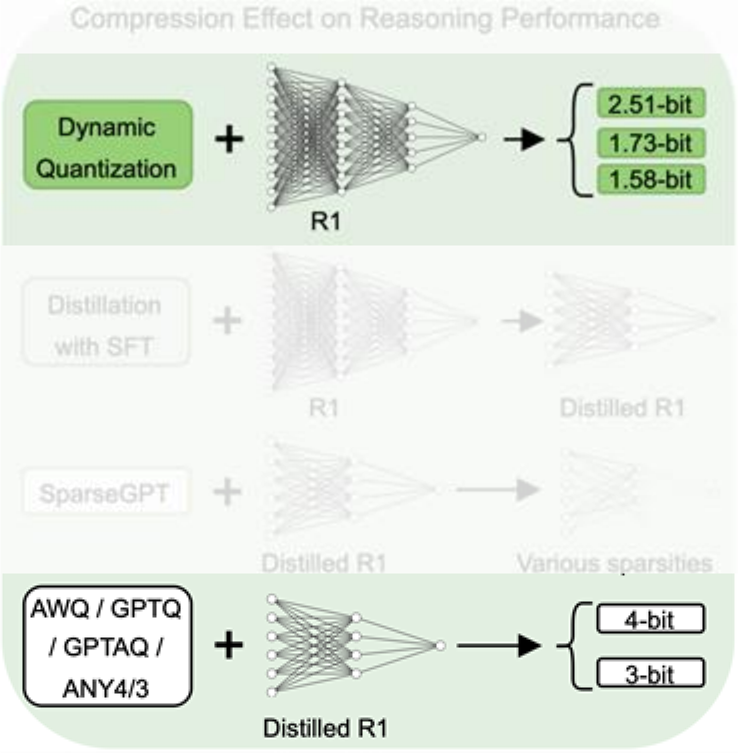


Pruning:

- Removal of weights
- Can be individual weights or larger structural components

	Model Structure	**Weights**
Distillation	❌	❌
Pruning	❌	✅
Quantization		

Background

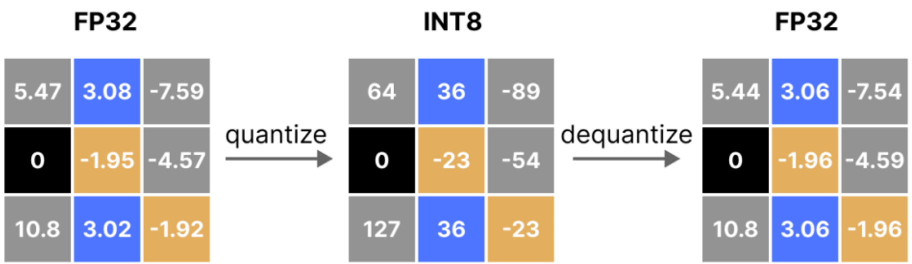


Quantization:

- Performed over original model
- Weights are compressed from FP16/32 to 4/3/... -bit

	Model Structure	Weights
Distillation	❌	❌
Pruning	❌	✅
Quantization	✅	❌

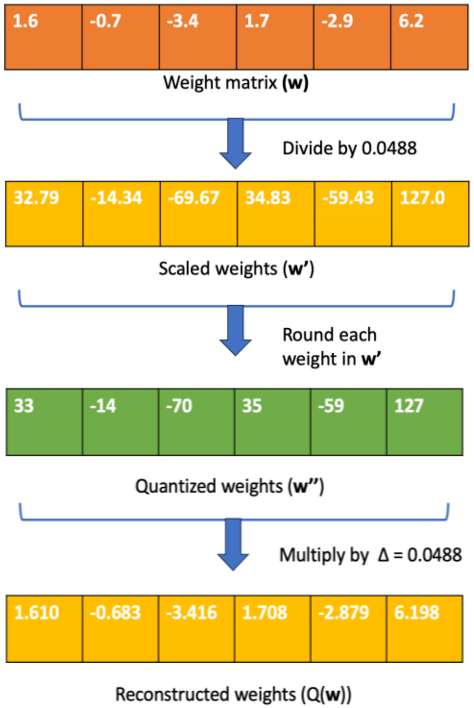
Background (Quantization)



Original layer: $y = \mathbf{w}\mathbf{x},$

Layer after Quantization: $y = Q(\mathbf{w})\mathbf{x}.$

Example function: $$Q(\mathbf{w}) = \Delta \cdot \boxed{\mathrm{Round}(\frac{\mathbf{w}}{\Delta})}, \quad \Delta = \frac{\max(|\mathbf{w}|)}{2^{N-1}}, \qquad (1)$$



AWQ

AWQ: Activation-aware weight quantization

Some weights are more important than others.

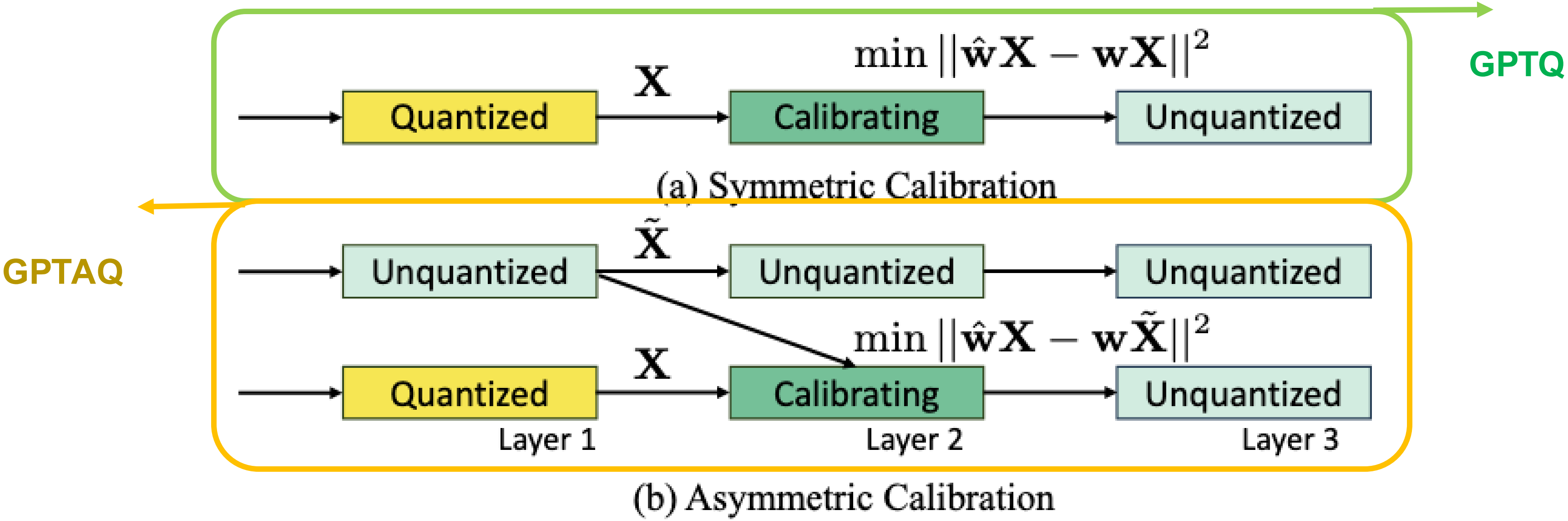
Idea: weight importance is related to activation magnitude

$$Q(\mathbf{w}) = \Delta \cdot \text{Round}(\frac{\mathbf{w}}{\Delta}),$$

$$Q_{int}(w_i \cdot s_i) = \text{Round}(\frac{w_i \cdot s_i}{\Delta})$$

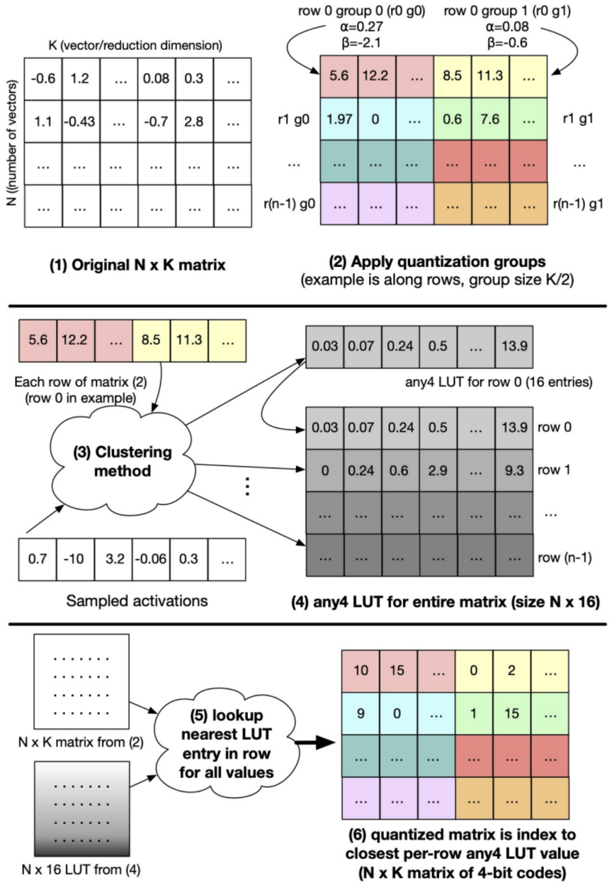
GPTQ / GPTAQ

- Quantize ever layer in order, to reduce the layer output quantization error.
- Leverages an approximation of the inverse Hessian to determine weight importance.



Any4/3

- A lightweight table for quantization
- No need to do preprocess for weights and activations
- Learn distribution from data itself, not trying to match a distribution(nf4,af4)



Dynamic

- Some weights are more important than others:
  - down_project in the first 3-6 layers
  - Shared MoE layers
  - lm_head and Embeddings
  - Layer norms and MoE router

MoE Bits	Type	Disk Size	Accuracy	Link	Details
1.58bit	UD-IQ1_S	**131GB**	Fair	Link	MoE all 1.56bit. `down_proj` in MoE mixture of 2.06/1.56bit
1.73bit	UD-IQ1_M	**158GB**	Good	Link	MoE all 1.56bit. `down_proj` in MoE left at 2.06bit
2.22bit	UD-IQ2_XXS	**183GB**	Better	Link	MoE all 2.06bit. `down_proj` in MoE mixture of 2.5/2.06bit
2.51bit	UD-Q2_K_XL	**212GB**	Best	Link	MoE all 2.5bit. `down_proj` in MoE mixture of 3.5/2.5bit

Data

Tasks	
AIME 2024	Top math chllenges
FOLIO	Logical reasoning
Temporal	Event temporal reasoning
MuSiQue	Multihop reasoning

Gpt-4o annotated outputs for:

- Uncertainty Estimation
- Adding Knowledge
- Backtracking
- Example Testing

Table 5: Dataset statistics of selected reasoning benchmarks.

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

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% sparse	23.3	71.6	97.6	64.2	(6.7, 13.49)
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% sparse	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 & 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)

Results

- Pruning deteriorates performance

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% sparse	23.3	71.6	97.6	64.2	(6.7, 13.49)
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% sparse	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 & 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)

Results

- Pruning deteriorates performance

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% sparse	23.3	71.6	97.6	64.2	(6.7, 13.49)
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)
					99.6	80.9	(12.0, 21.57)
					99.3	72.6	(4.7, 11.92)
					99.7	77.1	(5.7, 13.21)
					99.9	82.2	(2.7, 10.95)
					97.9	66.2	(2.3, 9.01)
					99.1	83.1	(3.3, 10.28)
					99.6	83.0	(4.0, 11.78)
					99.7	81.5	(2.7, 11.88)
					99.7	82.2	(5.7, 12.68)
					98.9	72.5	(4.0, 11.55)
					99.5	74.2	(2.3, 9.18)
					99.9	78.6	(3.7, 10.27)
					81.5	65.2	(0.0, 4.43)
					84.0	66.6	(0.3, 5.05)
					65.9	58.1	(0.3, 4.68)
					69.3	58.6	(0.0, 3.73)
					88.7	66.1	(0.0, 3.54)
					67.3	47.8	(0.0, 2.89)
					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)

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)

Results

- Pruning deteriorates performance
- MuSiQue experiences the most significant performance drop
- AIME 2024, collapses at 3-bit

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% sparse	23.3	71.6	97.6	64.2	(6.7, 13.49)
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% sparse	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 & 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)

Background

Difference of Means

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

Behaviors:

Uncertainty Estimation

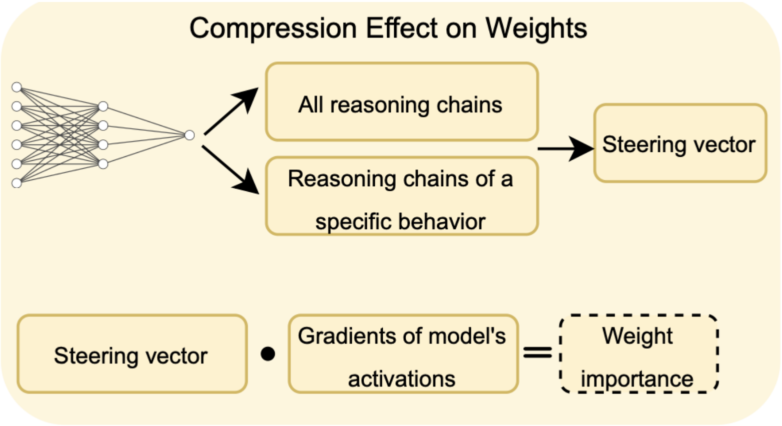
Adding Knowledge

Backtracking

Example Testing

Attribution Patching:

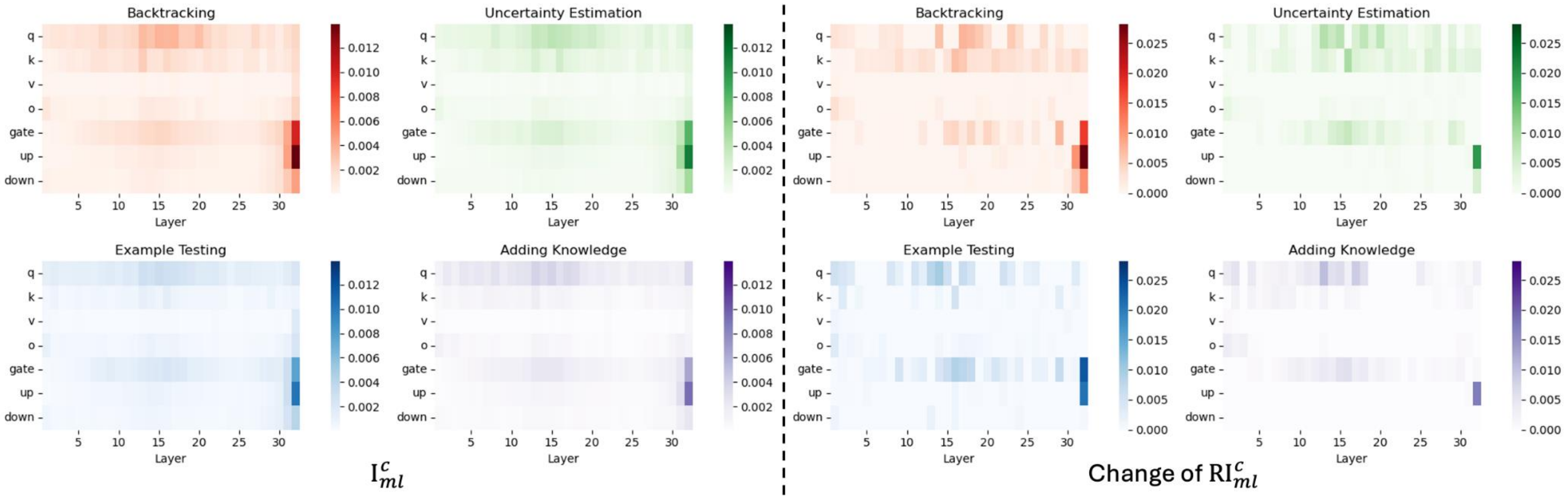
$$\mathbf{I}^c_{m\ell} \approx \frac{1}{|\mathcal{D}_+|} \left| \sum_{s_i^c \in \mathcal{D}_+} \left(\tilde{\mathbf{u}}^c_{m\ell}\right)^\top \frac{\partial}{\partial \mathbf{a}_{m\ell}} \mathcal{L}(s_i^c) \right|$$



JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

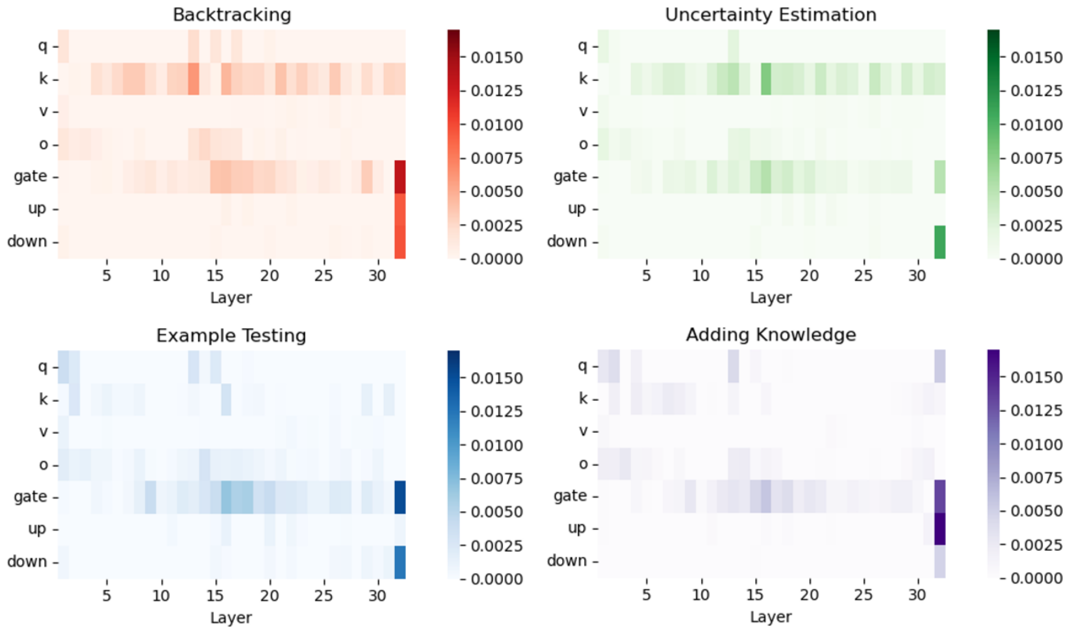
Importance Metric (compared to base)

- 32_up_project has the highest importance score for the evaluated behaviors
- experiences the biggest change in importance



Importance Metric (compared to AWQ)

- 32_up_project importance is not preserved for 'Backtracking' and 'Adding Knowledge'



JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

Verification/Results

- Selectively quantizing components: (4bit)

Quantized Component	Rank	AIME 2024	FOLIO	Temporal	Avg
32_up	1st overall	20.0	63.1	63.6	48.9
32_gate	2nd col	33.3	62.1	67.2	54.2
32_v	last col	43.3	68.0	79.6	63.6
31_up	2nd row	33.3	70.0	64.4	55.9
1_up	last row	6.7	64.5	80.4	50.5

- Altering AWQ such that the final layer MLP is preserved: (3bit)

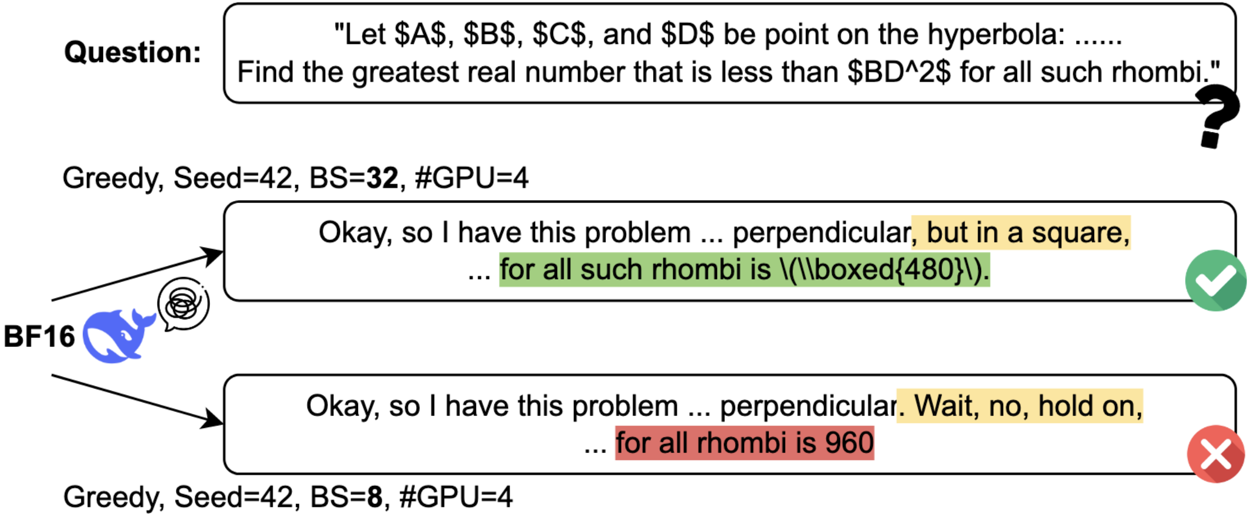
Model	Compression	Full-Precision Anywhere?	AIME 2024	FOLIO	Temporal	Avg	MuSiQue
R1-Distill-Llama-8B	3-bit AWQ	-	10.0	59.6	68.4	46.0	(0.0, 3.50)
R1-Distill-Llama-8B	3-bit AWQ	Final-layer MLP	**16.7**	**67.0**	**74.0**	**52.57**	**(1.0, 3.62)**

Understanding and Mitigating Numerical Sources of Nondeterminism in LLM Inference

Yuan, Jiayi, et al.

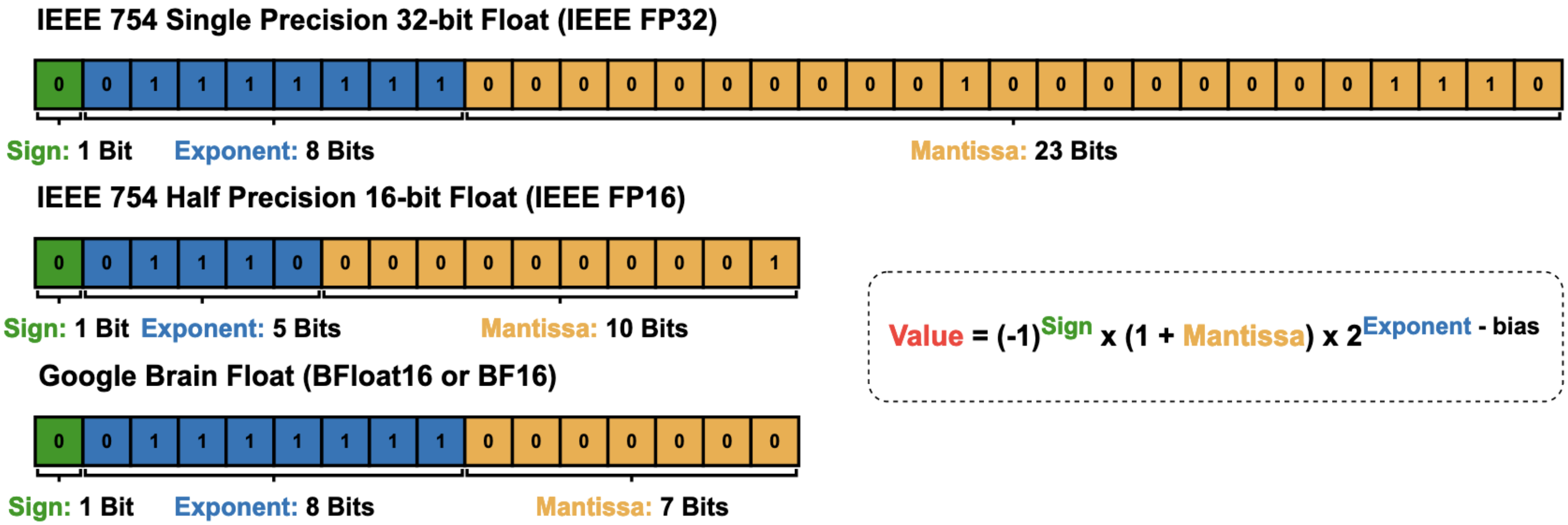
One sentence for the paper?

- Different running configurations can lead to variation of results. (Live by FP32, die by FP32)



Precision

- Larger Mantissa means **more precise** and Larger Exponent means **larger range** of values



$$\text{Value} = (-1)^{\text{Sign}} \times (1 + \text{Mantissa}) \times 2^{\text{Exponent} - \text{bias}}$$

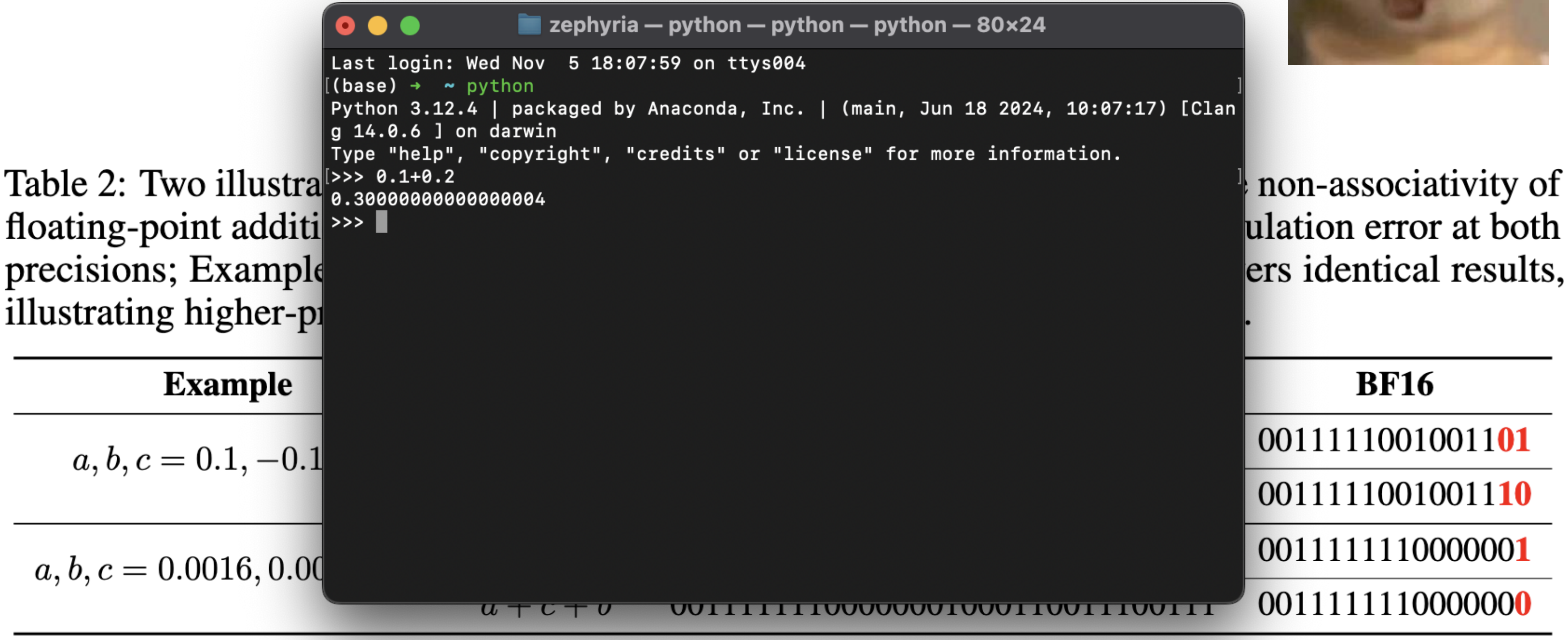
Precision

:O

Table 2: Two illustrative cases demonstrate how rounding error, together with the non-associativity of floating-point addition, can affect numerical results. Example 1 reveals accumulation error at both precisions; Example 2 exhibits a discrepancy only in BF16, while FP32 delivers identical results, illustrating higher-precision numeric types are more tolerant of rounding errors.

Example	Sum Order	FP32	BF16
$a, b, c = 0.1, -0.1, 0.2$	$a+b+c$	001111100100110011001100110011**01**	00111110010011**01**
	$a+c+b$	001111100100110011001100110011**10**	00111110010011**10**
$a, b, c = 0.0016, 0.0027, 1.0$	$a+b+c$	00111111100000001000110011100111	001111111000000**1**
	$a+c+b$	00111111100000001000110011100111	001111111000000**0**

Precision



zephyria — python — python — python — 80×24
Last login: Wed Nov 5 18:07:59 on ttys004
(base) → ~ python
Python 3.12.4 | packaged by Anaconda, Inc. | (main, Jun 18 2024, 10:07:17) [Clang 14.0.6] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> 0.1+0.2
0.30000000000000004
>>>
Table 2: Two illustra
non-associativity of
floating-point additi
ulation error at both
precisions; Example
ers identical results,
illustrating higher-p
Example
BF16
$a, b, c = 0.1, -0.1$
0011111001001101
0011111001001110
$a, b, c = 0.0016, 0.00$
0011111110000001
0011111110000000

Experiment Settings

- Models:
  - Reasoning Models:DeepSeek-R1-DistillQwen-7B, DeepSeek-R1-Distill-Llama-8B
  - Non-Reasoning Models: Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct

- Benchmarks: AIME'24, MATH500, LiveCodeBench-[Easy,Medium,Hard]

- **Running Configurations**
  - GPU Types: NVIDIA L40S, A100
  - GPU counts: 2, 4
  - Batch Sizes: 8, 16, 32

Experiment Metrics

- **Greedy Decoding:**
  - Std@Acc(⬇): Standard deviation of accuracy
  - Avg_Std@Output_Length(⬇): Average standard deviation of output length
  - Div_Index(⬆): Model produce identical token sequences up to a certain position, but generate different tokens after that position
  - Avg_Std@top1_prob(⬇): Average standard deviation of top-1 token prediction probability(0 to Div_Index)

- **Random Sampling:**
  - Pass@1(⬇): Standard deviation of Pass@1

Greedy Decoding

- FP32 helps a lot
- BF16 exhibits substantial instability.

Table 3: Std@Acc of greedy decoding across 12 different settings (GPU types, GPU counts, and batch sizes) under BF16, FP16, and FP32 Numerical Precisions. Reasoning models also exhibit larger variance compared to non-reasoning counterparts. More results can be found in Appendix C.

	AIME'24			MATH500			LiveCodeBench-Easy		
	BF16	**FP16**	**FP32**	**BF16**	**FP16**	**FP32**	**BF16**	**FP16**	**FP32**
DeepSeek-R1-Distill-Qwen-7B	9.15%	5.74%	0	1.04%	1.12%	0.12%	1.67%	1.28%	0.37%
DeepSeek-R1-Distill-Llama-8B	4.60%	6.00%	5.8e-17	1.59%	0.73%	0.23%	2.31%	1.92%	0.29%
Qwen2.5-7B-Instruct	1.71%	1.45e-17	1.45e-17	0.83%	0.36%	1.16e-16	0.79%	0.48%	1.16e-16
Llama-3.1-8B-Instruct	1.92%	1.30%	0	0.94%	0.34%	0.13%	1.00%	0.67%	0.25%

Table 4: Standard deviation of output length of greedy decoding across 12 different settings (GPU types, GPU counts, and batch sizes) under BF16, FP16, and FP32 numerical precisions. The output length of reasoning models exhibit large variance. More results can be found in Appendix C.

	AIME'24			MATH500			LiveCodeBench-Easy		
	BF16	**FP16**	**FP32**	**BF16**	**FP16**	**FP32**	**BF16**	**FP16**	**FP32**
DeepSeek-R1-Distill-Qwen-7B	9189.53	5990.32	0	2774.28	2090.46	138.75	5507.52	4282.78	262.55
DeepSeek-R1-Distill-Llama-8B	9348.59	7822.43	0	4015.00	2518.38	146.03	4732.85	3652.16	105.85
Qwen2.5-7B-Instruct	211.47	48.14	0	52.61	15.37	0	7.79	0.71	0
Llama-3.1-8B-Instruct	119.21	49.73	0	124.43	40.57	2.76	31.03	4.70	0.49

Greedy Decoding

- Answer from BF16 will diverge more quickly
- FP32 doesn't diverge too much

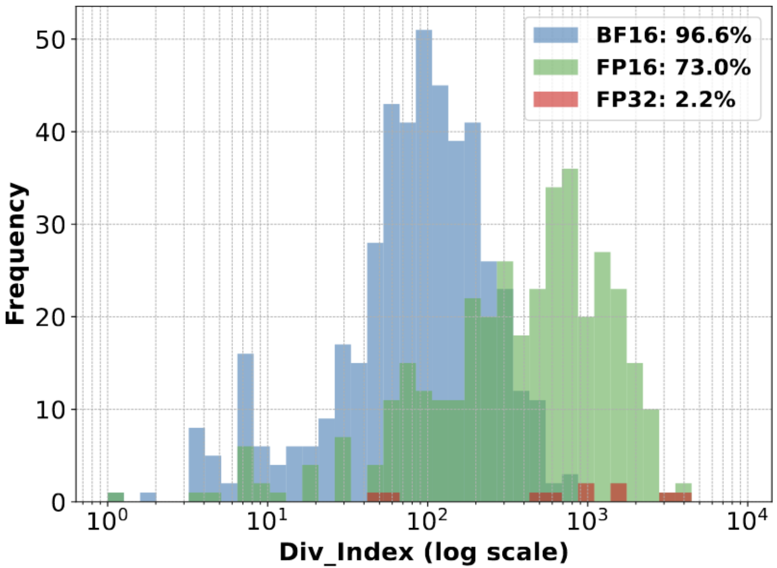


Figure 5: Distribution of Div_Index for DeepSeek-R1-Distill-Qwen-7B on MATH500 under BF16, FP16, and FP32. Higher numerical precisions lead to fewer divergent examples and a shift of divergence point to later token positions.

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

Greedy Decoding

- FP32 helps a lot
- BF16 exhibits substantial instability.

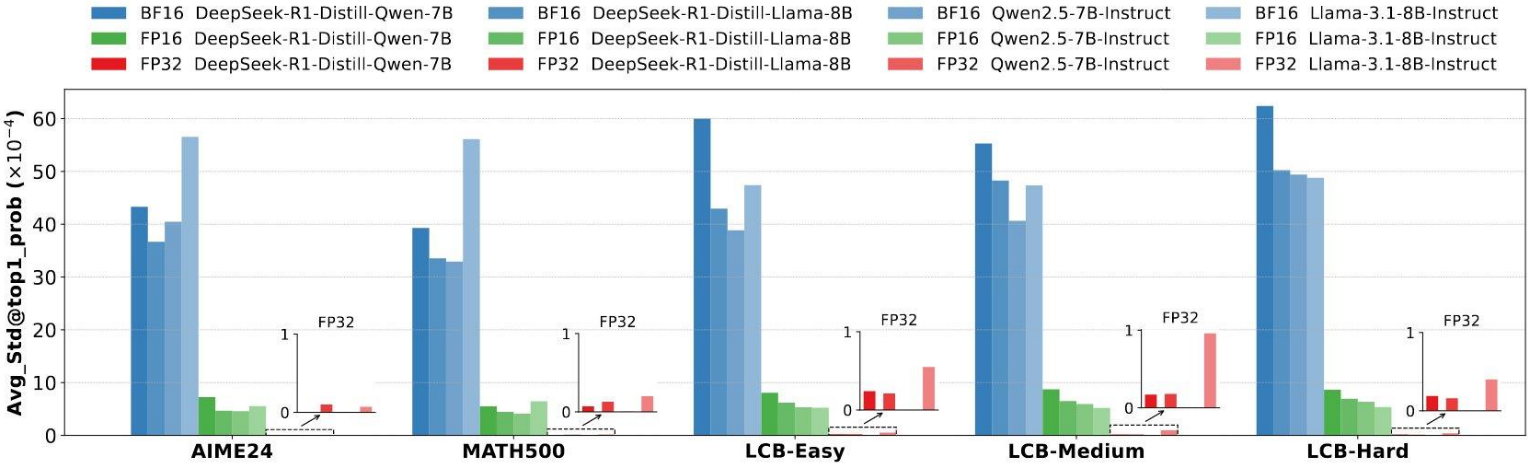


Figure 4: Avg_Std@top1_prob across 12 different settings for 4 models and 5 tasks, under BF16, FP16 and FP32. FP16 shows significantly lower variance compared to BF16. FP32 yields near-zero variance, demonstrating strong robustness to floating-point rounding errors.

Random Sampling

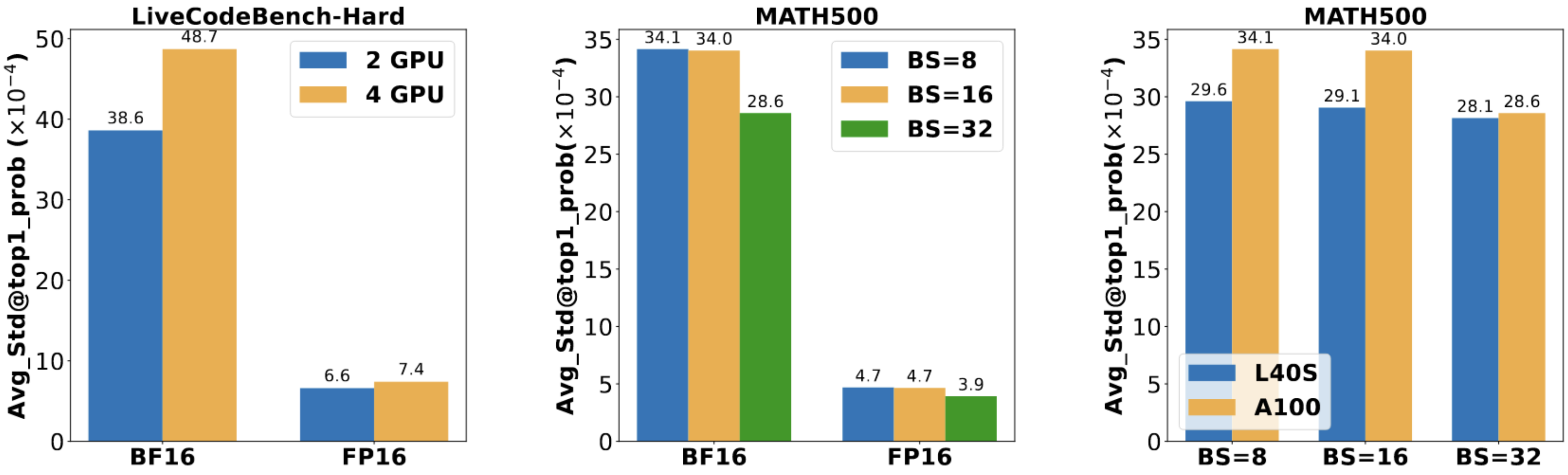
- Authors: result of dataset size and sampling dynamics, not contradiction

Table 5: Standard deviation of Pass@1 performance (%) under different GPU counts and precisions. We emphasize that the reported values reflect **the variability of Pass@1 performance across 6 different system configurations** (3 batch sizes × 2 GPU counts), *not* across repeated runs of the same configuration.

	MATH500 (n=4)			AIME'24 (n=16)			AIME'24 (n=64)		
	BF16	**FP16**	**FP32**	**BF16**	**FP16**	**FP32**	**BF16**	**FP16**	**FP32**
DeepSeek-R1-Distill-Qwen-7B	0.3158	0.1463	0.1021	1.7151	0.8273	1.1785	0.3749	0.5391	0.7377
DeepSeek-R1-Distill-Llama-8B	0.3602	0.3371	0.1211	1.5124	1.8792	0.8606	0.8774	0.8539	0.5034
Qwen2.5-7B-Instruct	0.4663	0.1686	0.0274	0.7056	0.2523	0	0.1784	0.1382	0
Llama-3.1-8B-Instruct	0.6020	0.1725	0.3293	0.5992	0.2282	0.7759	0.4216	0.2898	0.1296

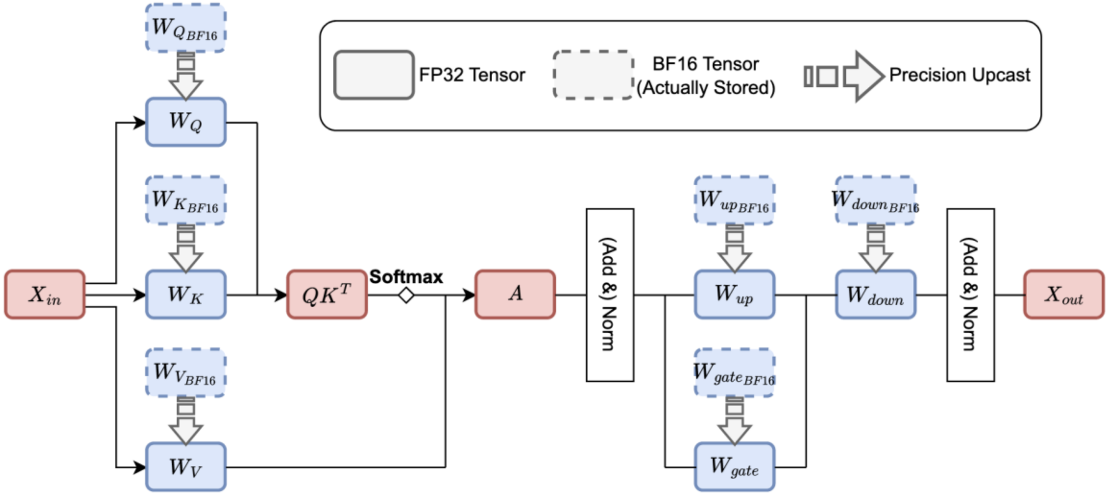
How Runtime Configurations Affect Reproducibility

- More GPUs means more parallel computation
- Smaller BS means more sequential processing steps
- Hardware-level implementations and memory hierarchies count



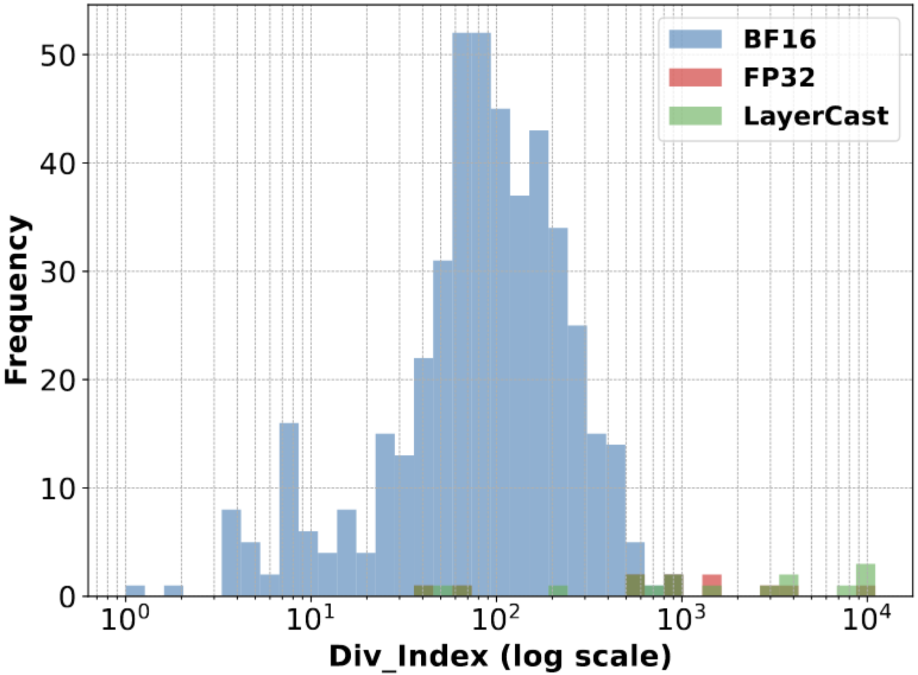
LayerCast

- Store as BF16 and Compute as FP32

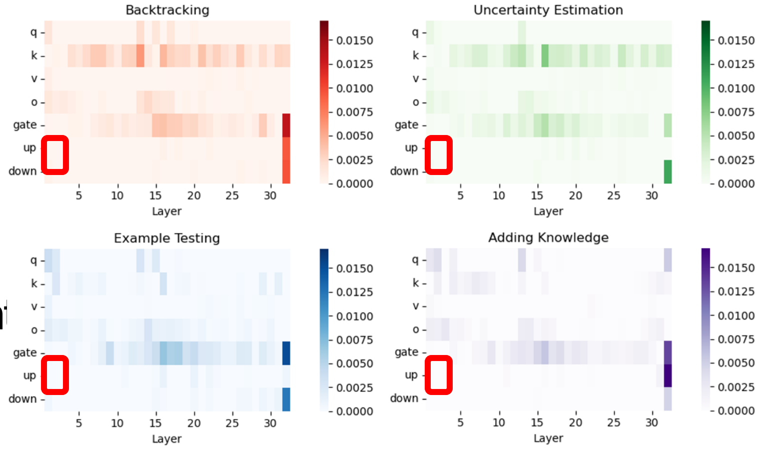


LayerCast

- It is as good as FP32!



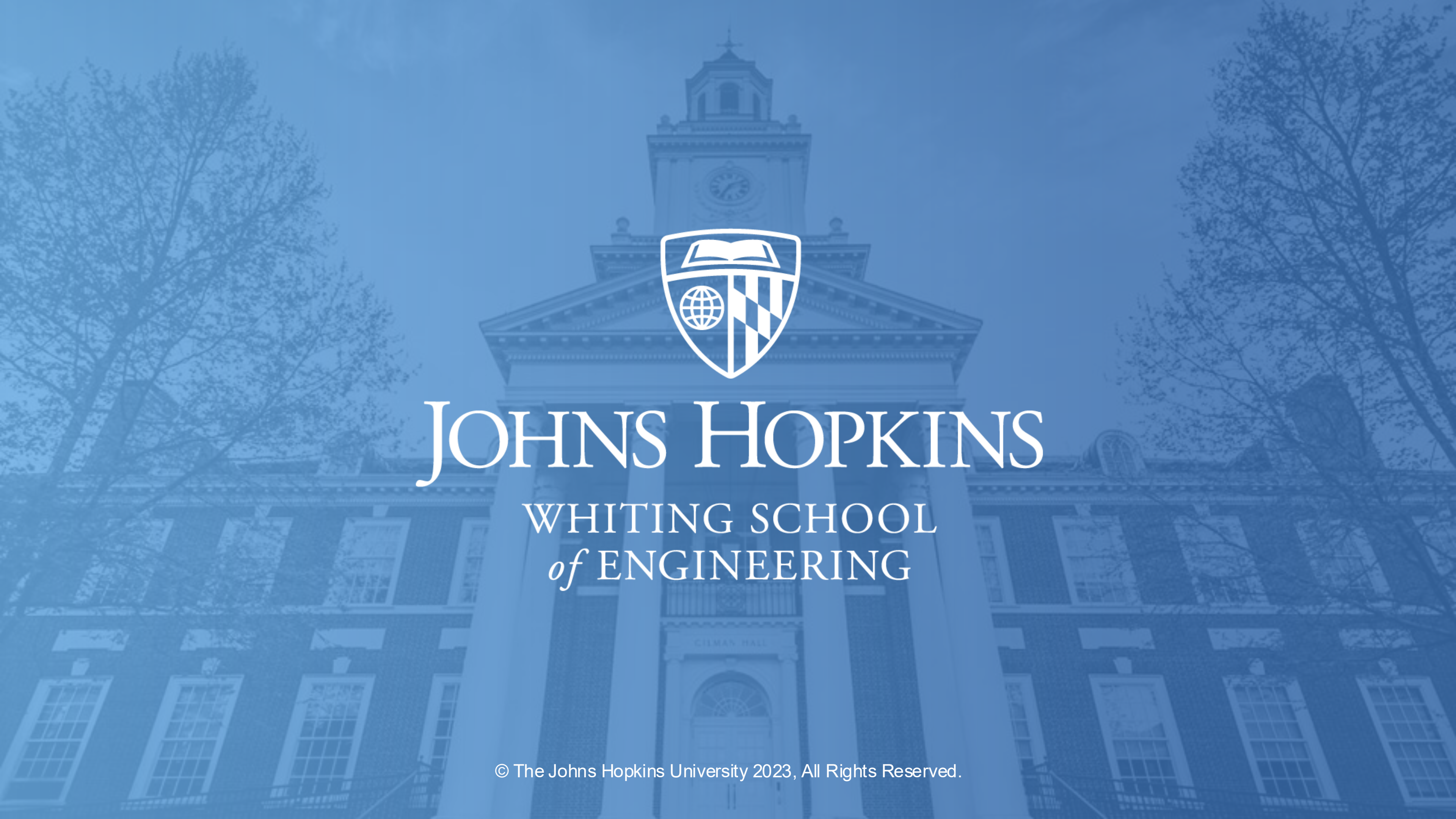
Disccusion



- FP32, FP16 or BF16
  - What will be the result of INT8 or other quant
- Compression
  - Does the compression metric make sense?
    - Does 'thinking' text actually represent what the model is doing.
    - Incorrect answers have longer chains of thought.
  - Calculate metric on the 4-bit model, apply change only to the 3-bit model.

Quantized Component	Rank	AIME 2024	FOLIO	Temporal	Avg
32_up	1st overall	20.0	63.1	63.6	48.9
32_gate	2nd col	33.3	62.1	67.2	54.2
32_v	last col	43.3	68.0	79.6	63.6
31_up	2nd row	33.3	70.0	64.4	55.9
1_up	last row	6.7	64.5	80.4	50.5

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING



JOHNS HOPKINS

WHITING SCHOOL of ENGINEERING