Optimizing Small Language Models: An Experimental Investigation in Compressing Distilled LLaMA Architectures

Presented by Angelo Galavotti
Supervisor: Francesco Conti
Co-supervisor: Luca Bompani

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ARTIFICIAL INTELLIGENCE
Alma Mater Studiorum Università di Bologna

July 22nd 2025

The Rising Need for Local, Compact Models

In recent years, LLMs have transformed our relationship with technology, but at what cost?

- The **Environmental** Burden:
  - It's estimated that GPT-4.5 uses 20.5 Wh per large query.
  - Data centers consume 1-2% of global electricity, projected to reach 8% by 2030.
- The **Social** Burden:
  - Cloud-only access creates troubling power dynamics.
  - Service outages can affect millions simultaneously.
  - Every query can potentially become corporate training data.

How can we tackle this?

- One way is to optimize models so that they run **locally**, **efficiently**, and **sustainably**.

The LLaMA family of models

Meta's LLaMA family of Large Language Models provide state-of-the-art performance, and unlike other competitors, they adopt an **open weights** license.

- In addition, LLaMA 3.2 features a **highly compact**, distilled **1B parameters** model.

This variation was chosen as the **base** on which our **compression** techniques were applied.

The Pipeline

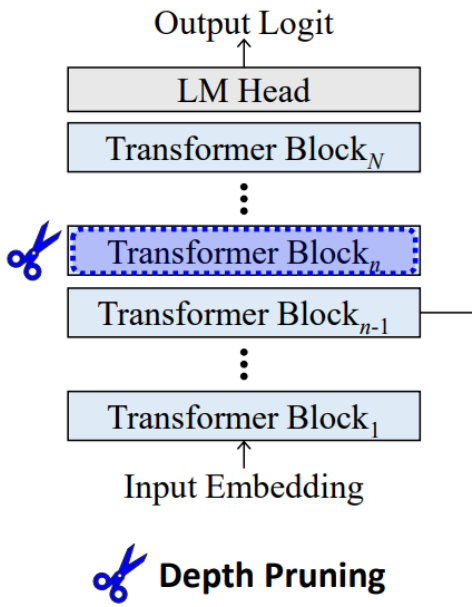
The compression pipeline is comprised of **5 stages**:

1. **Depth** pruning.
2. **Width** pruning.
3. **LoRA** fine-tuning.
4. **Quantization** using GPTQ.
5. **EoRA** performance recovery.

Depth Pruning

- In our **Depth Pruning** implementation, the **importance** of layers is evaluated by measuring **perplexity** after their removal.

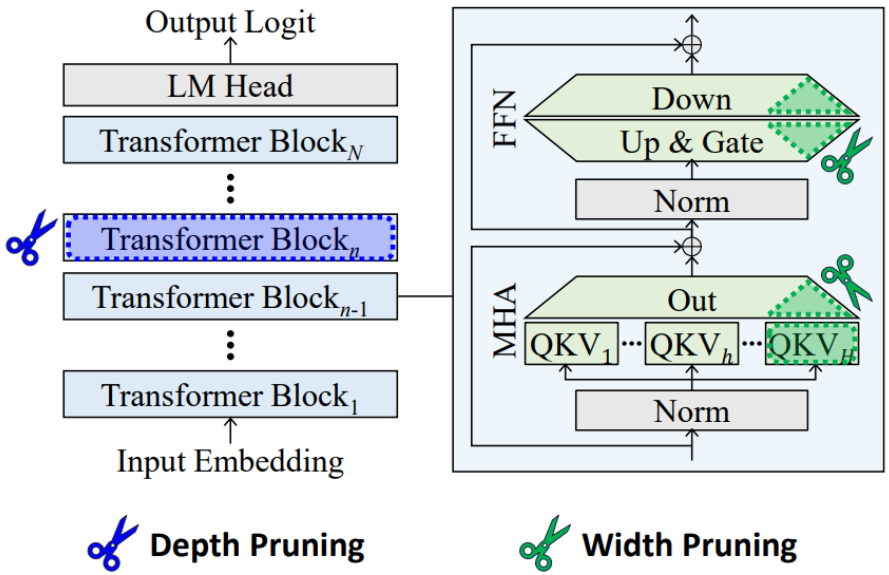
$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i \mid w_{<i})\right)$$



Width Pruning

- Width Pruning is applied in a **structured** (N:M) fashion, to **FFN** and **Attention projection** matrices.
- In our implementation, the **importance** of weights is measured by taking into account the **weight and activation magnitudes** (WANDA).

$$S_{ij} = |W_{ij}| \cdot \|X_j\|_2$$



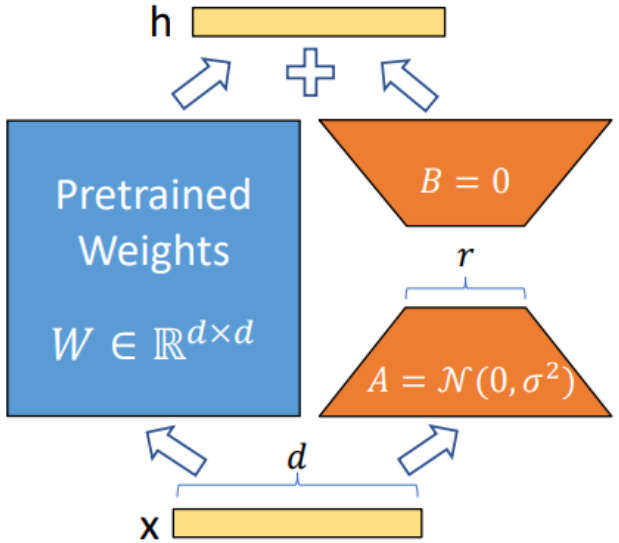
LoRA Fine-tuning

Low-Rank Adaptation (LoRA) allows for **fine-tuning** of models while being **parameter efficient**, by decomposing the weight update matrix into **two** smaller **low-rank matrices** that are learned during training.

- In this project, it has the double purpose of **performance recovery** and **task adaptation**.

$$W_0 + \Delta W = W_0 + BA$$

$$h = W_0 X + BAX$$



E.g.: 4M vs. 32K trained parameters for the W_Q matrix.

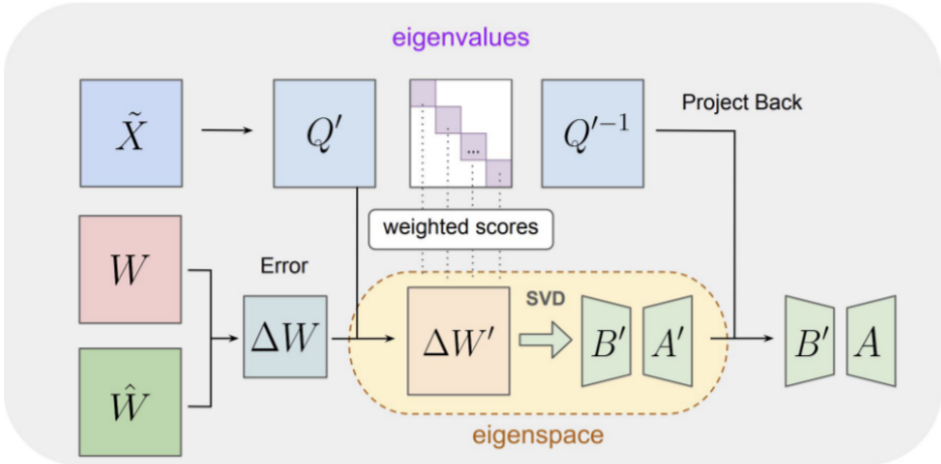
Quantization and EoRA

- **4-bit quantization** is employed using the **GPTQ** algorithm.
  - Uses **second-order Hessian information** to minimize performance loss.
  - **Reduces** memory usage to ∼**40%** with minimal accuracy loss.
- After quantization, **Eigenspace Low-Rank Approximation** (EoRA) aims at compensating the quantization error.
  - It **projects** compression errors into **task-relevant eigenspaces**.
  - Then it is **factorized** using **SVD**.
  - The resulting low-rank matrices are then **re-projected back**.

$$\Delta W' = \Delta W \cdot Q' \approx B'A'$$

$$A = A'(Q')^{-1}$$

$$h = W_0 X + B'AX$$



Datasets

The configurations were evaluated using:

- **TriviaQA Open-book** and **Closed-book Accuracy**:
  - In **open-book** testing, the question's **context** is provided.
- **WikiText-2 Perplexity Evaluation**
  - It measures how well the model **predicts** the **next token** in a sequence.

These TriviaQA and WikiText-2 datasets were also used for **LoRA** fine-tuning.

- Fine-tuning employed 8,192 samples and batch size 8.

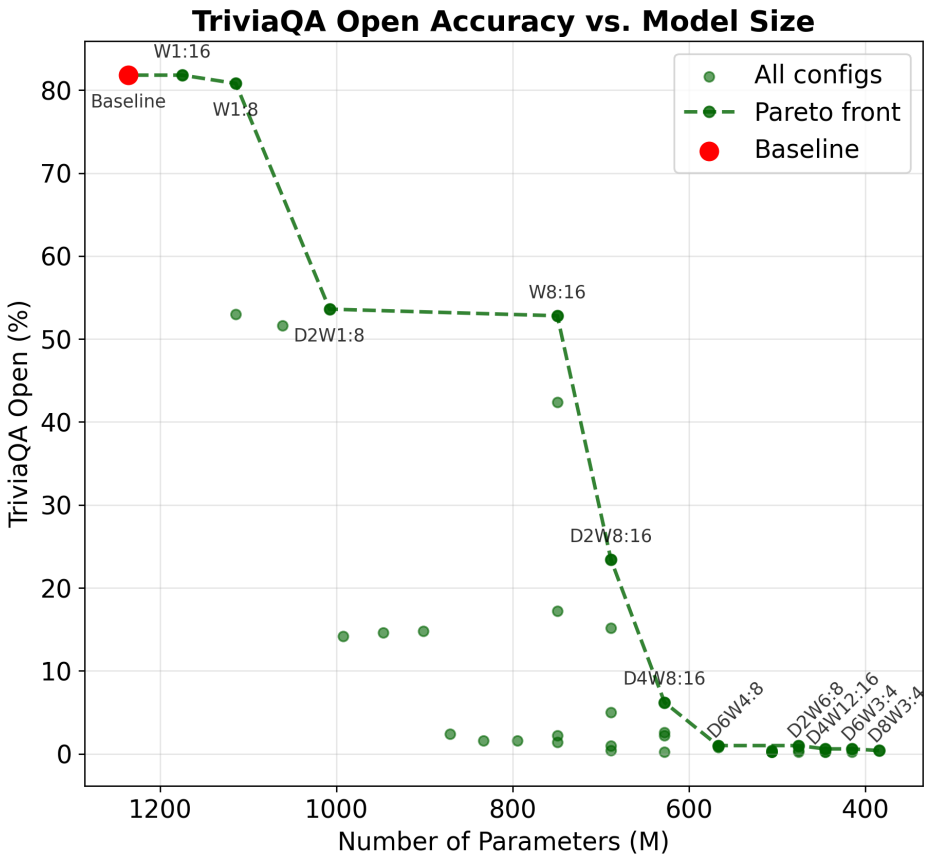
Evaluation Details

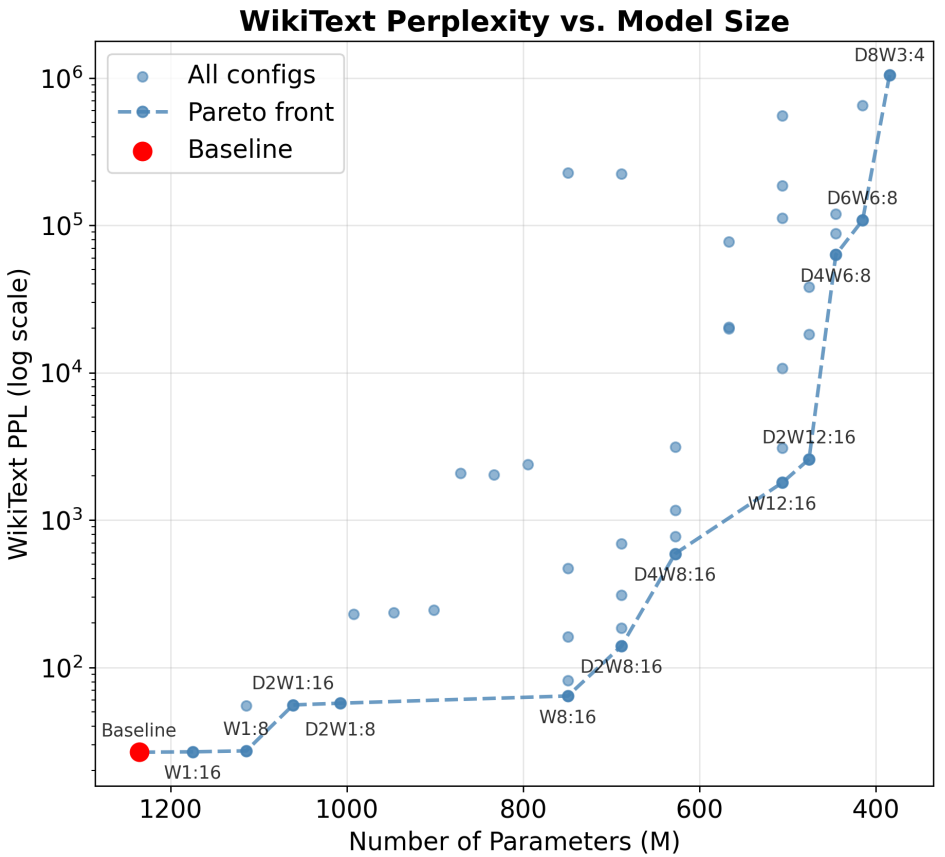
In total:

- We evaluated **42 pruned** configurations, meaning they either:
  - Underwent depth pruning (Depth 2→D2, Depth 4→D4...)
  - Underwent width pruning (Width 2:4→W2:4...)
  - Or a combination of both (Depth 2 + Width 2:4→D2W2:4...)

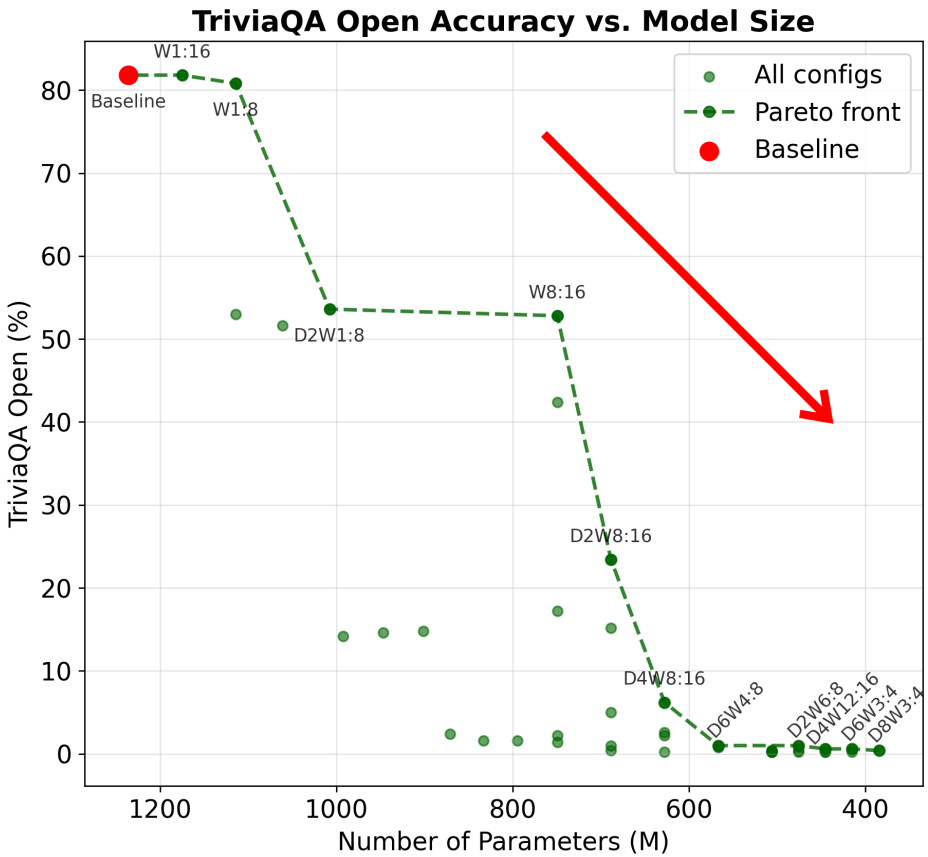
- Of these 42, the most interesting **27** were selected for independent **LoRA fine-tuning** on the two datasets, and were then **quantized**.
  - **10** of them also underwent **EoRA** compensation.

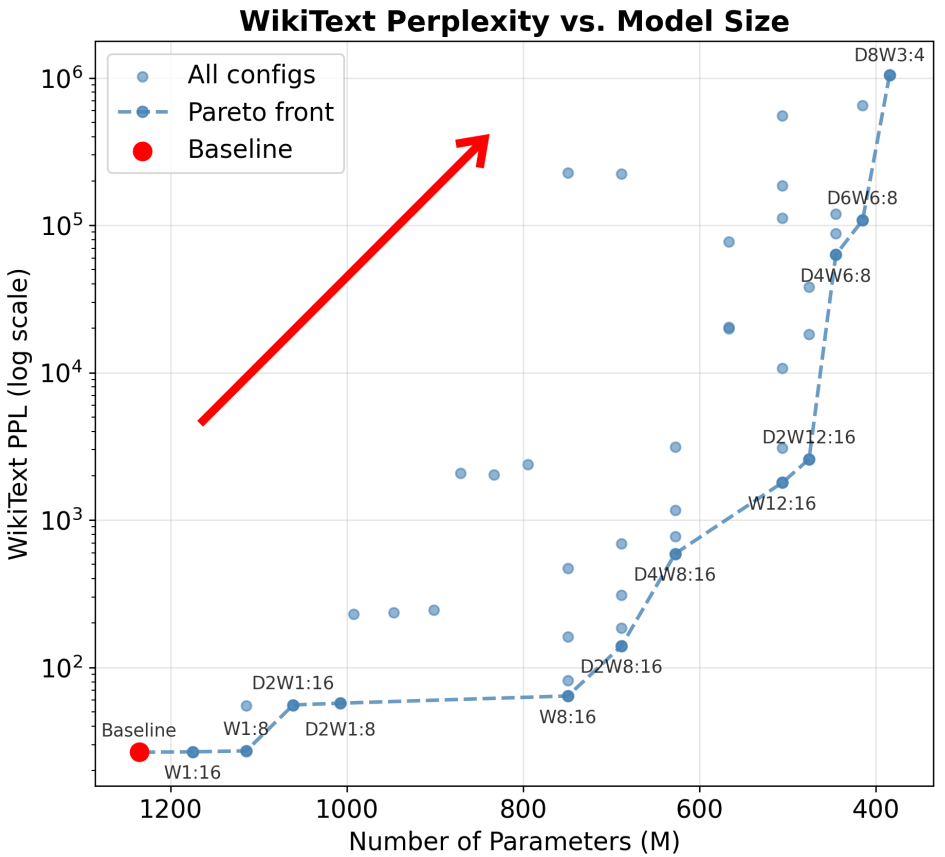
Pruned-only Results



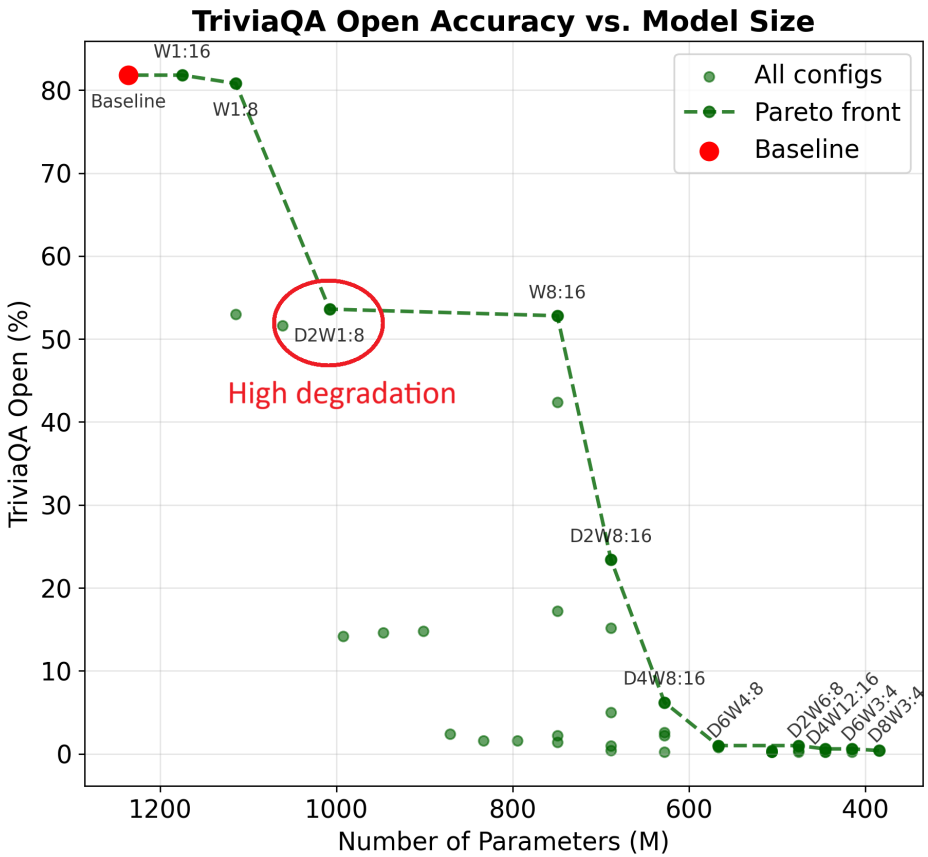


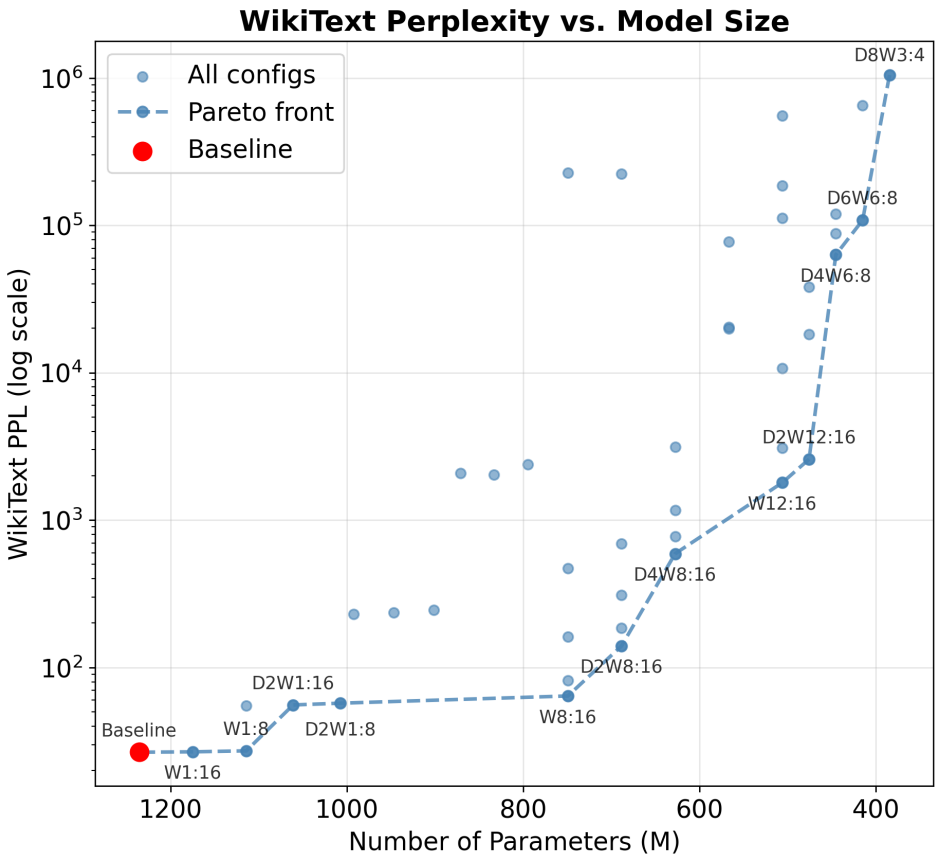
Pruned-only Results



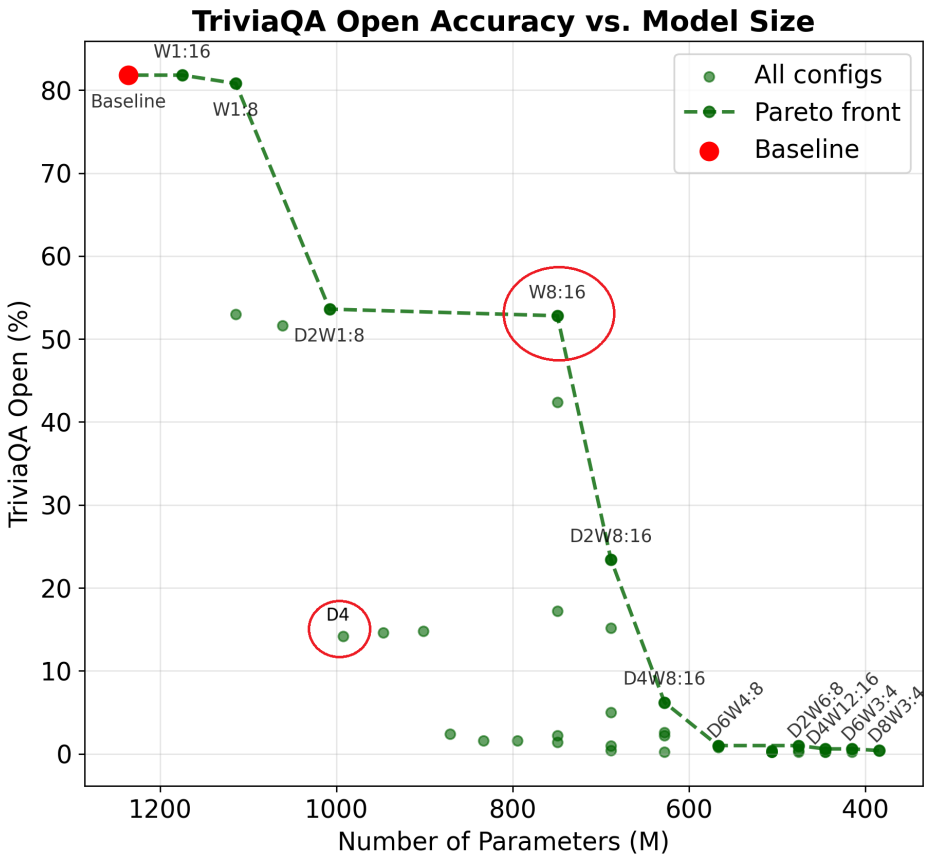


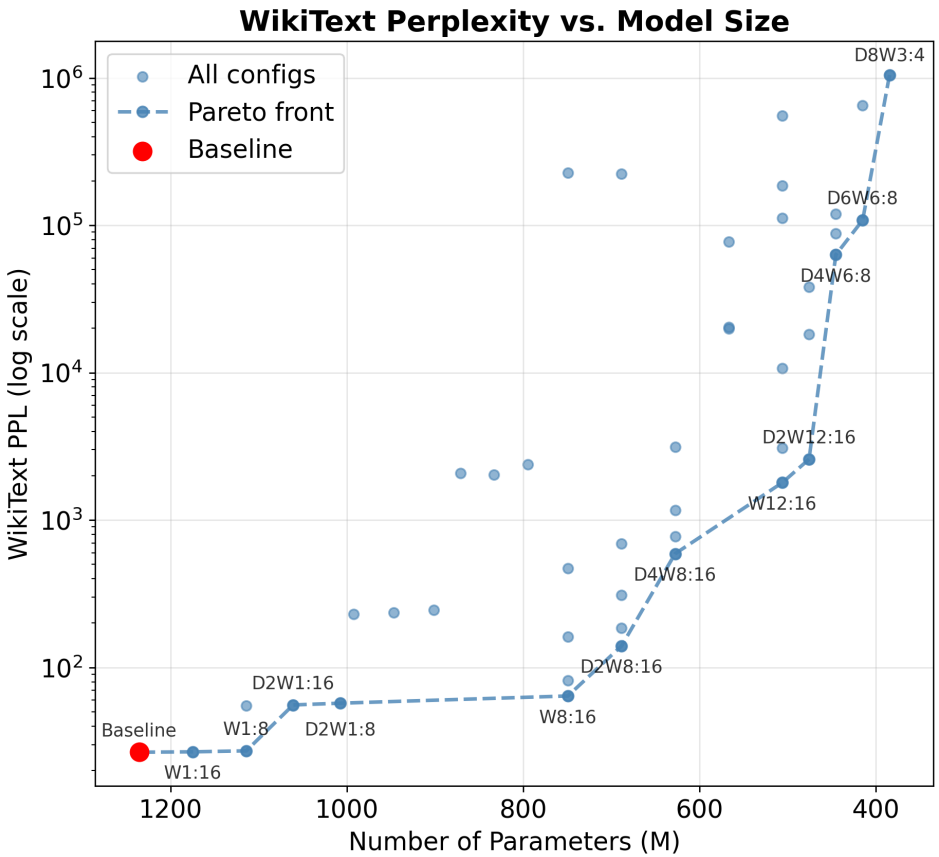
Pruned-only Results



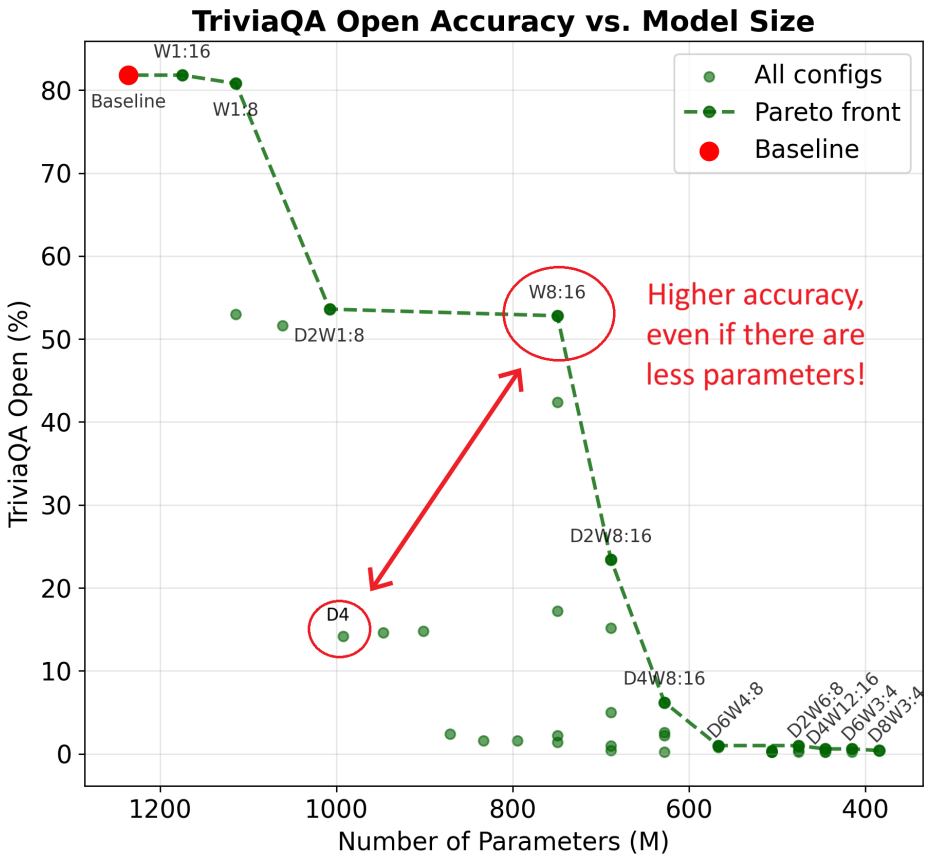


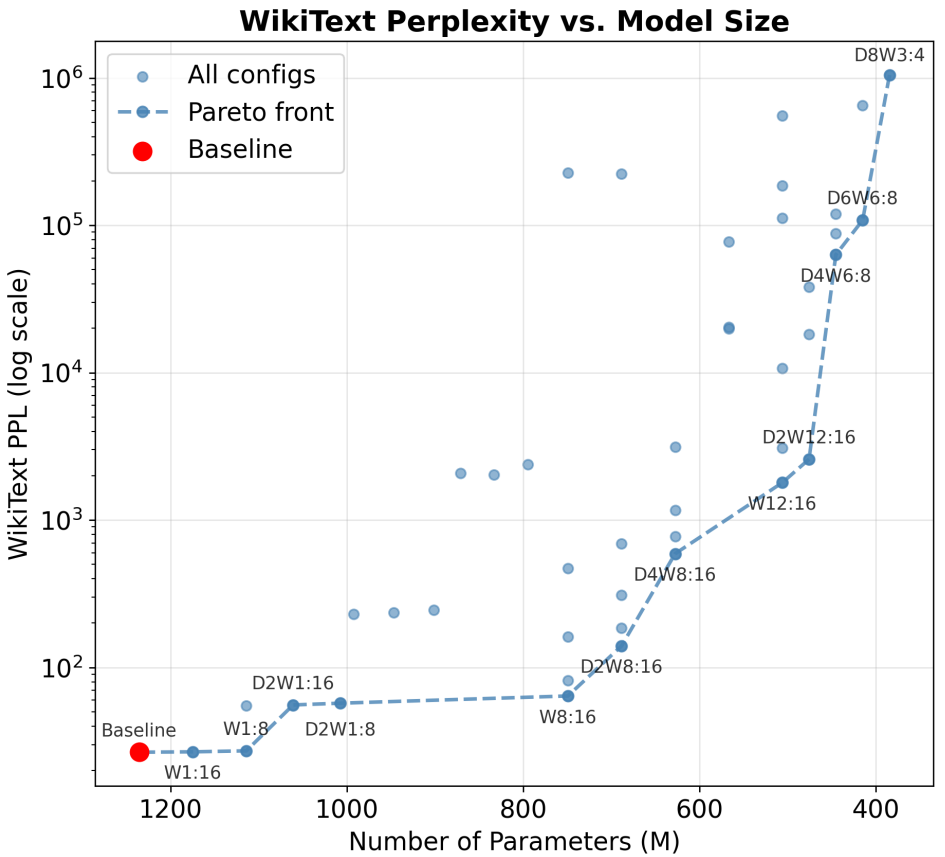
Pruned-only Results



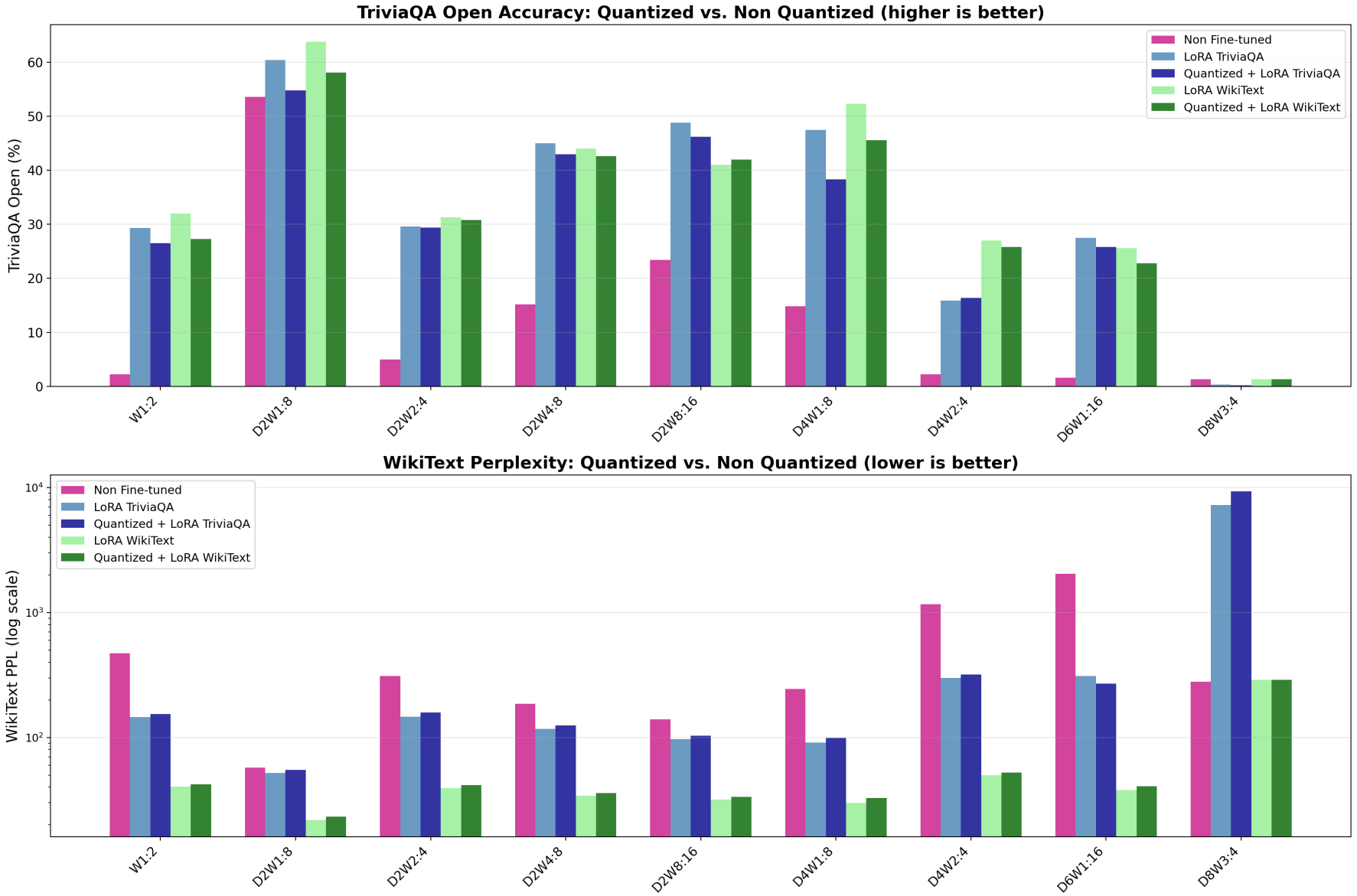


Pruned-only Results

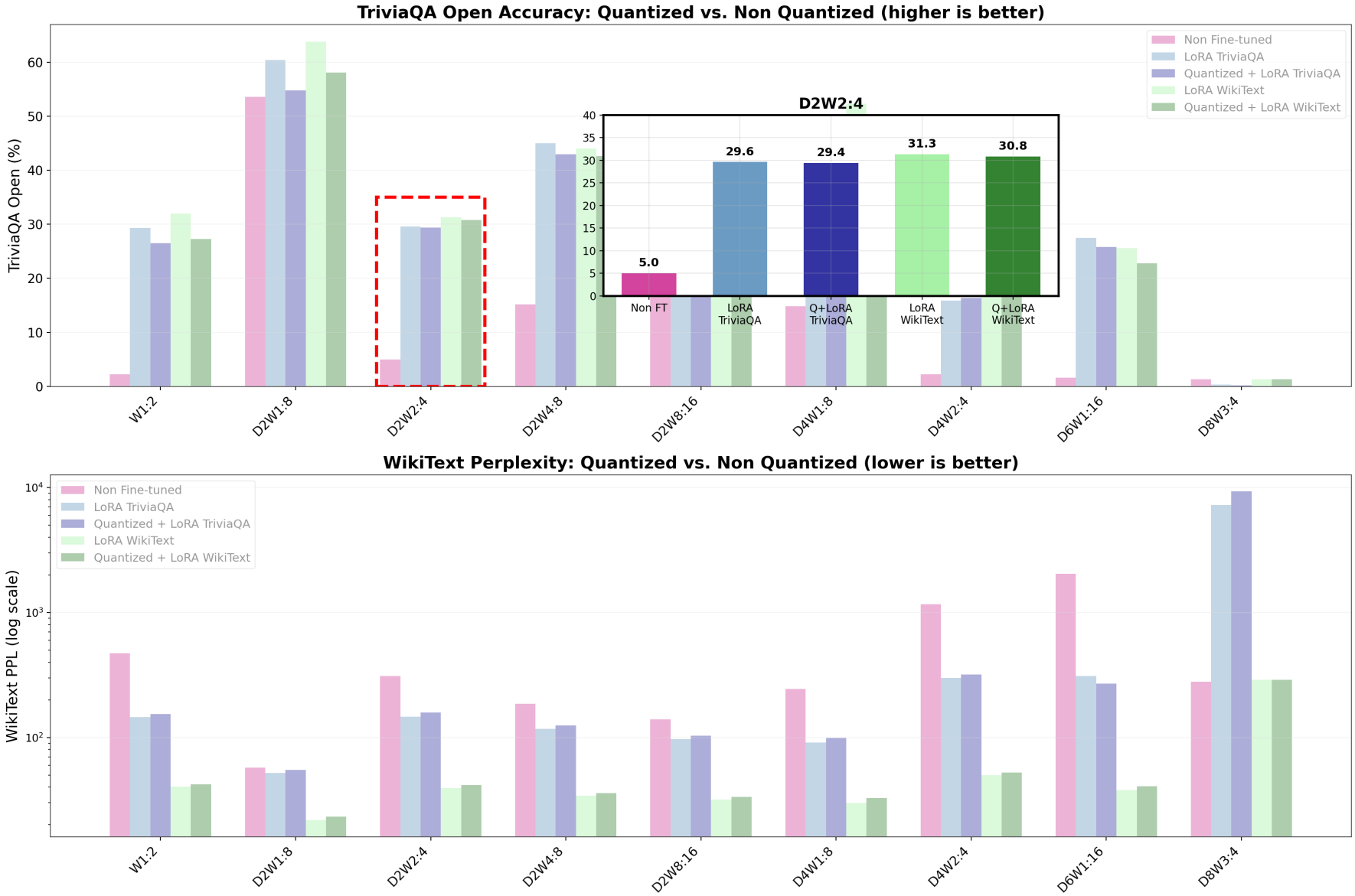




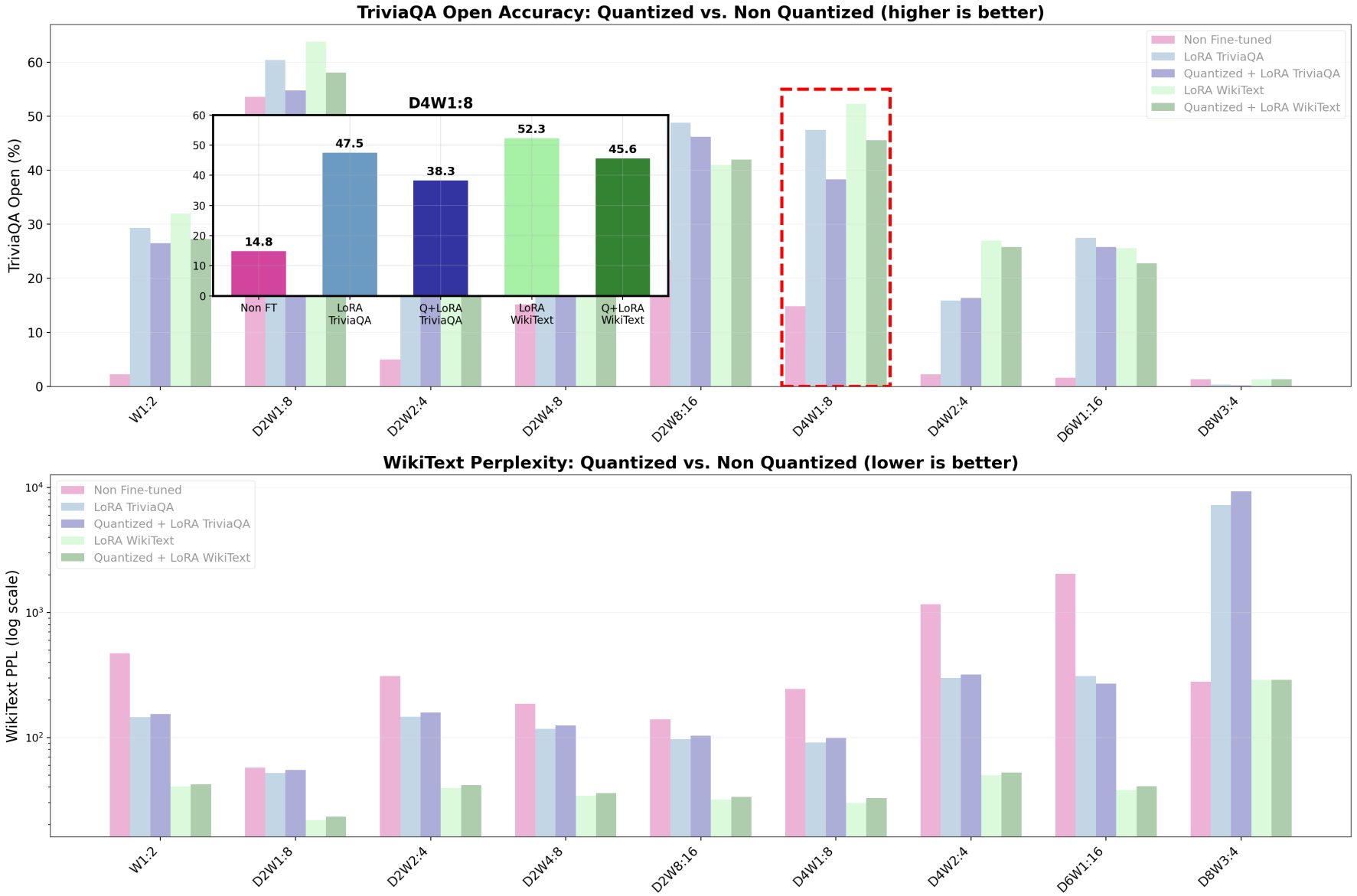
LoRA and Quantization



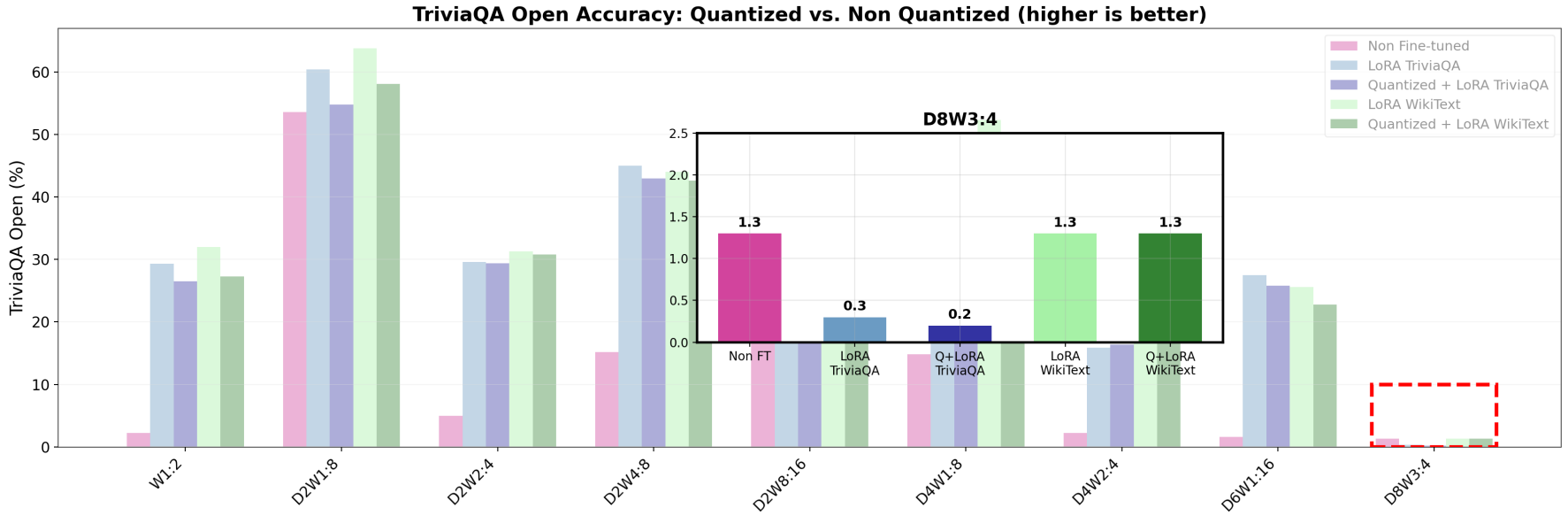
LoRA and Quantization

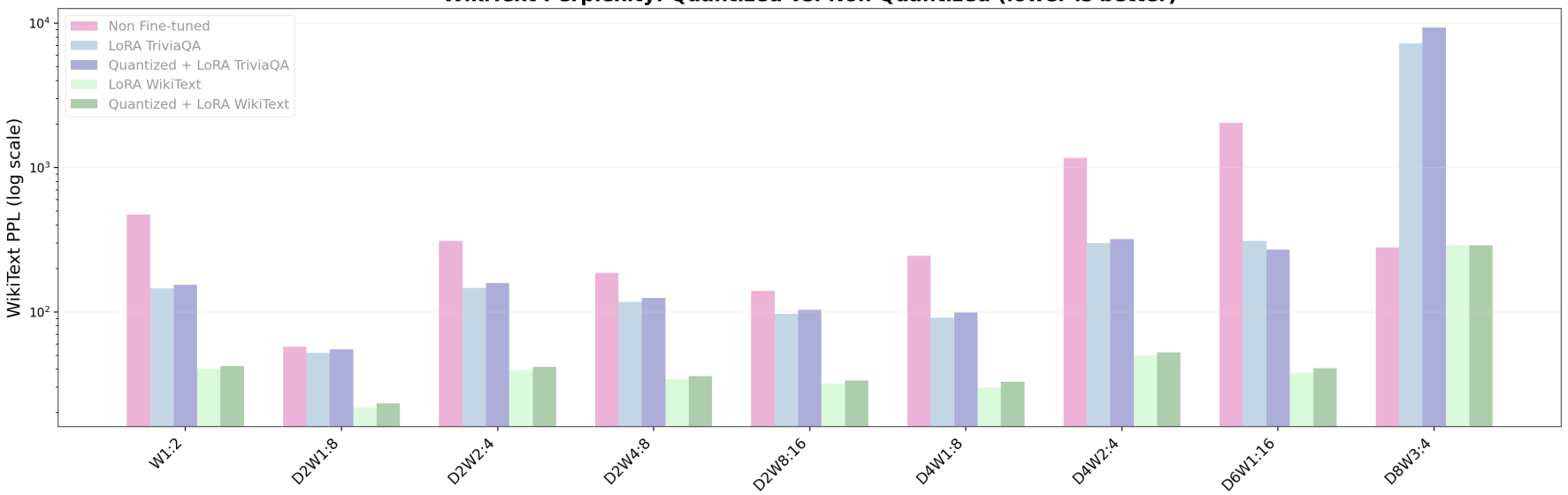


LoRA and Quantization

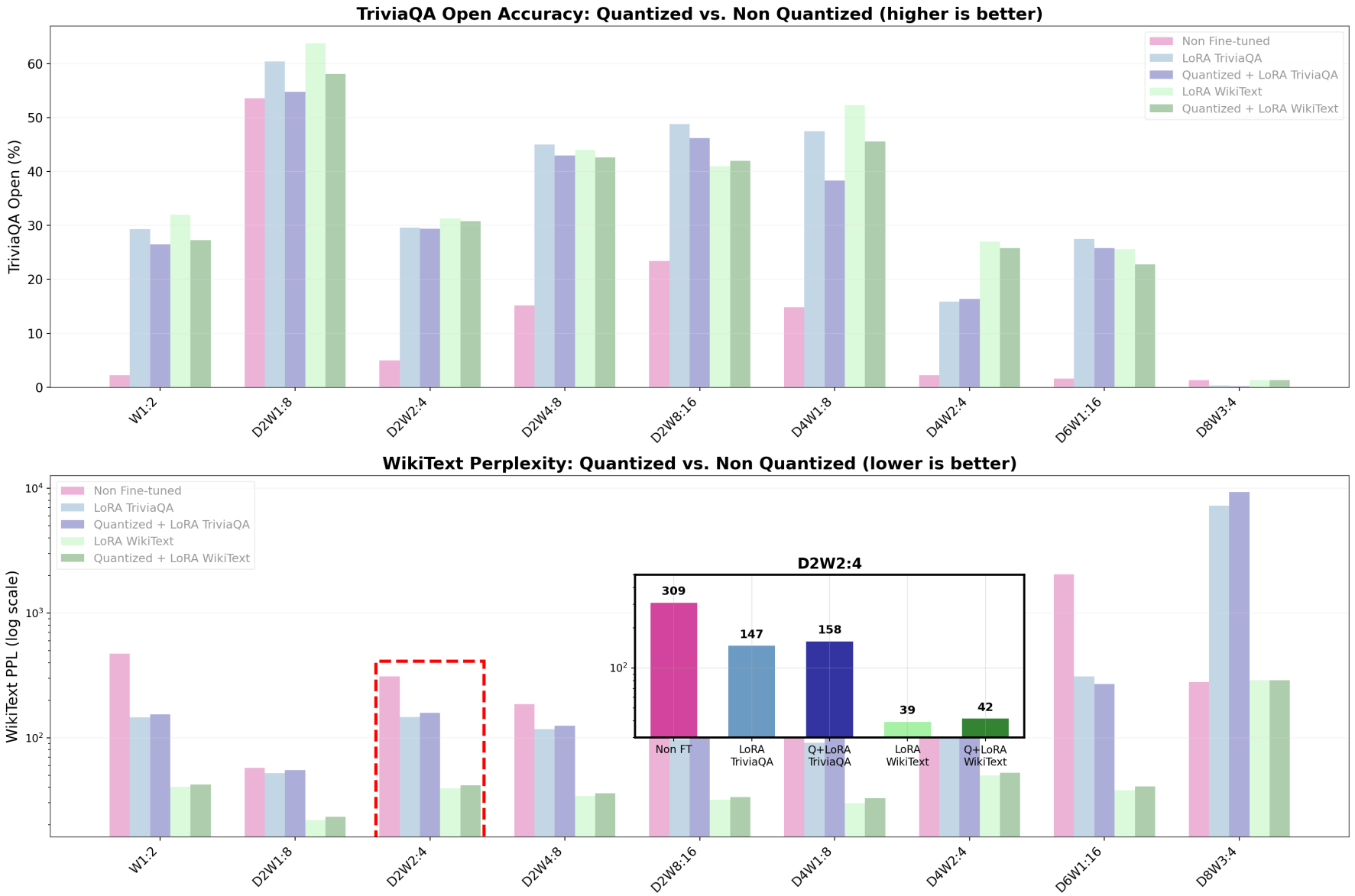


LoRA and Quantization

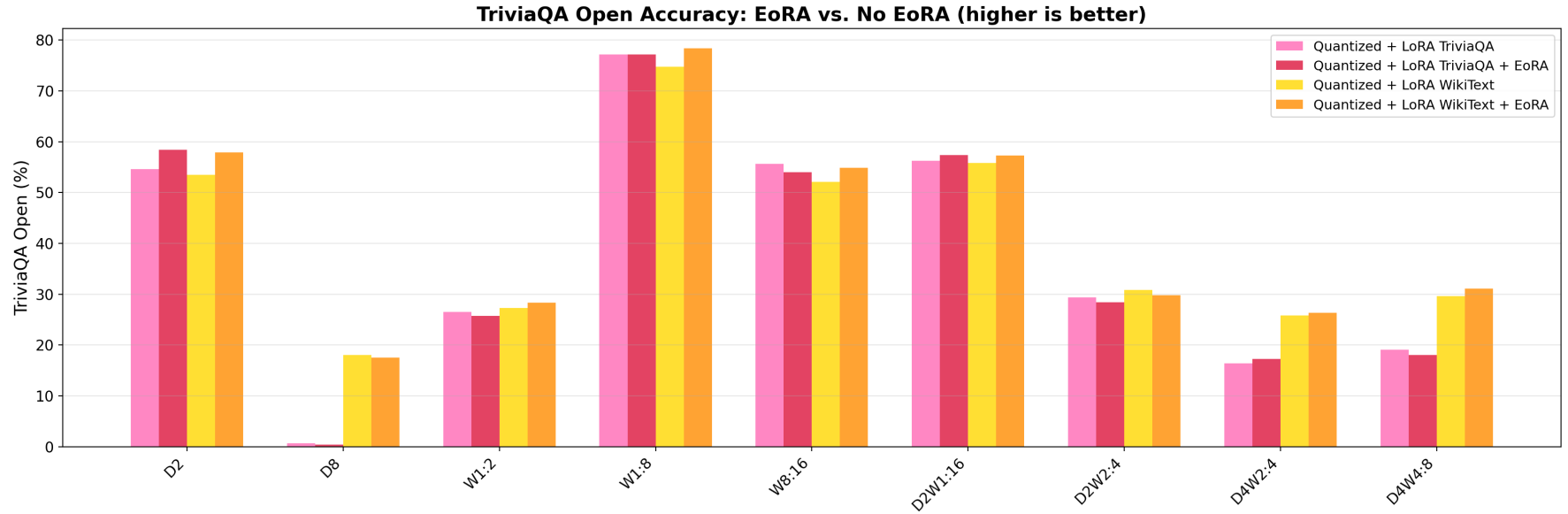


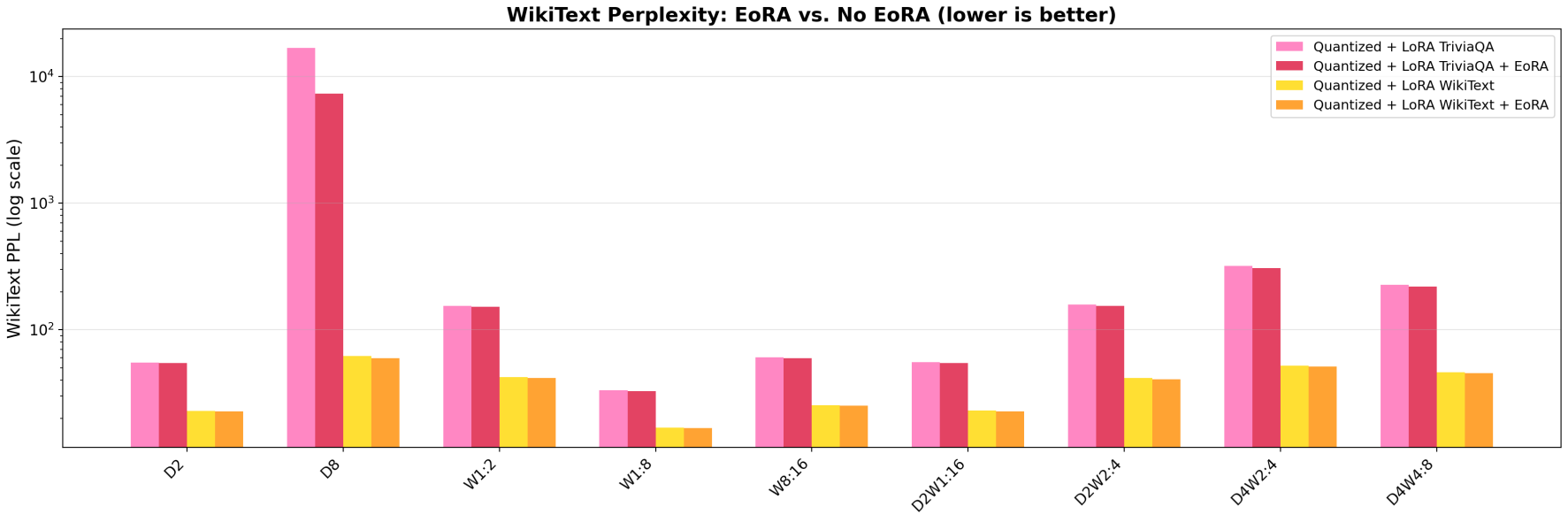


LoRA and Quantization

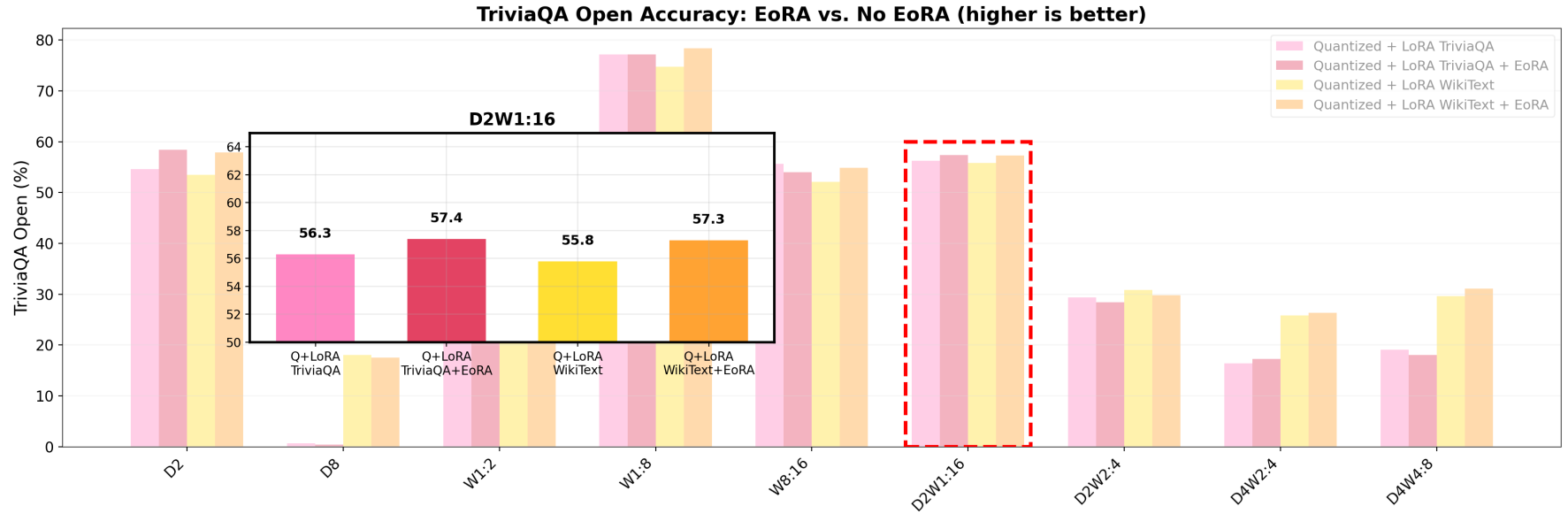


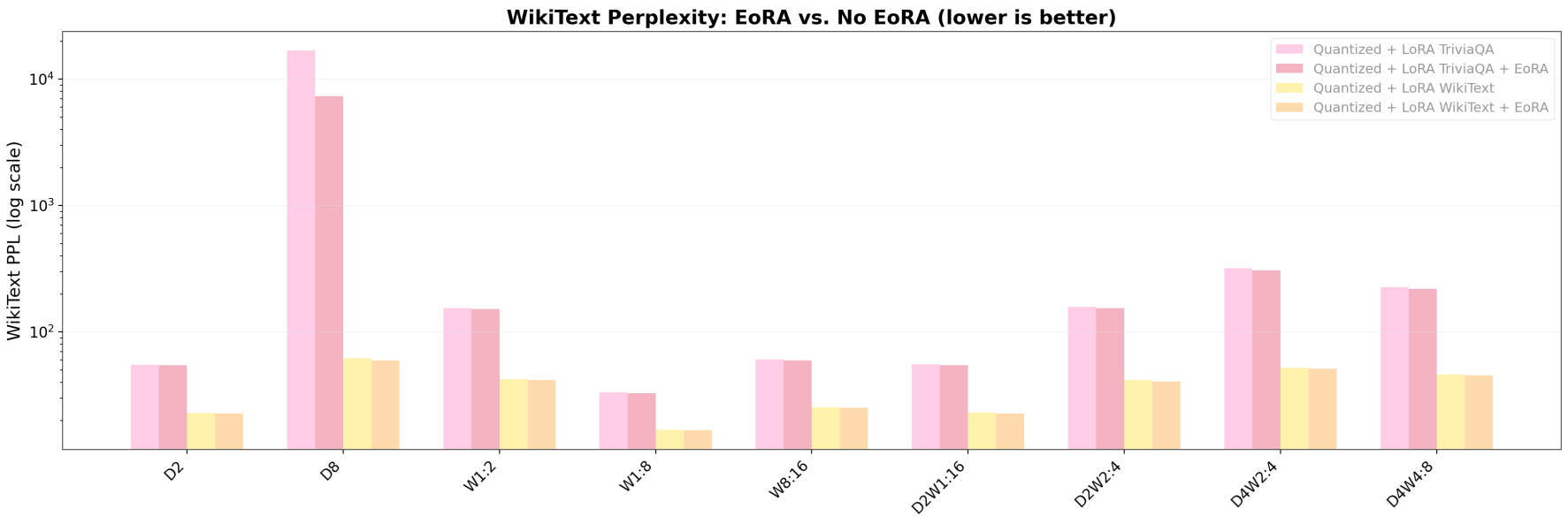
To EoRA or Not To EoRA?



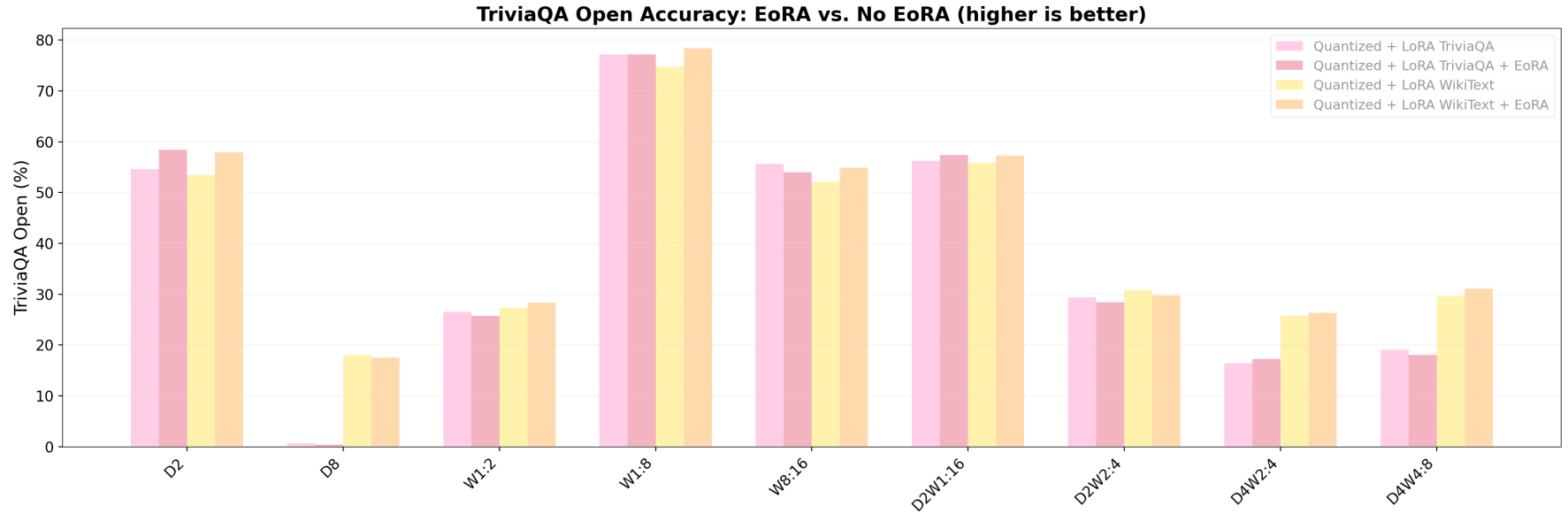


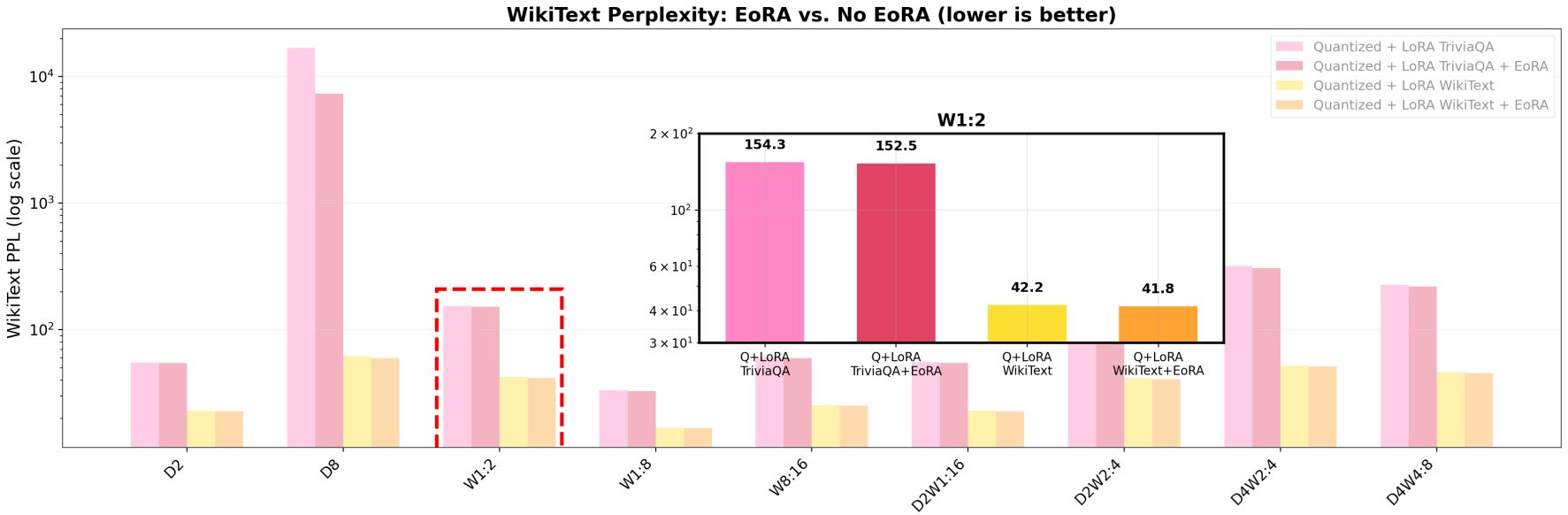
To EoRA or Not To EoRA?





To EoRA or Not To EoRA?





Conclusions and Future Work

- The models generated are **small**, while maintaining **good performance**!
- However...
  - The models with the **least parameters** cannot produce coherent text, and are simply **unusable**...
  - ...and in those particular cases, there's not much that can be done with LoRA and EoRA.
- Nevertheless, this work demonstrates that **optimization of distilled Small Language Models** is indeed **possible**, even with **notable compression rates**.

- Future Work:
  - Exploration of distillation techniques.
  - Support for a wider range of open-weight LLMs.
  - Enhancements to the evaluation framework.

Thank you for your attention!

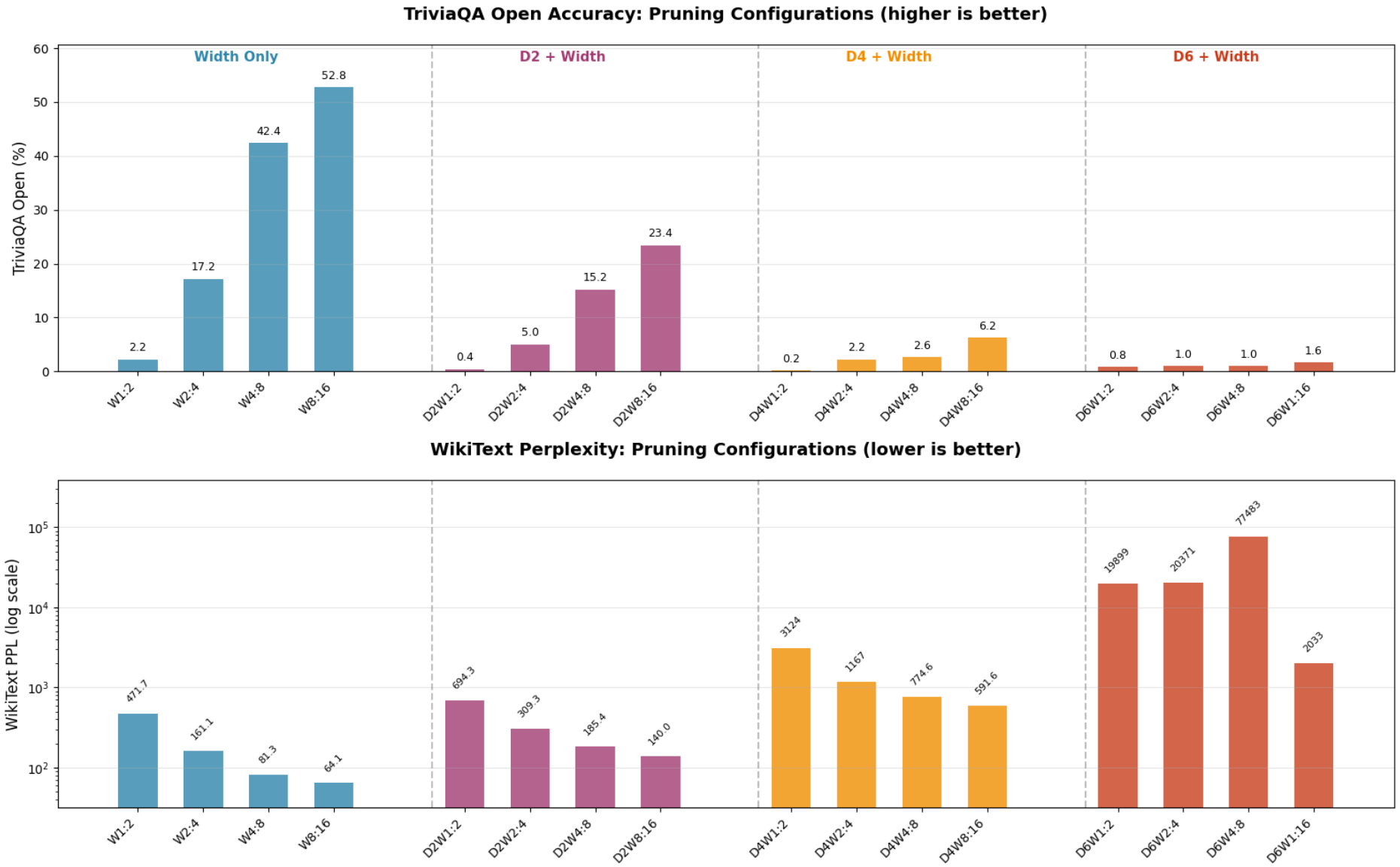
Adapter Size

- The generated LoRA adapter has rank 8.
- The EoRA adapter has rank 32.
- They are applied to the **attention projection matrices** of each layer:
  - W_Q is 2048×2048 → For LoRA, A is 8×2048, B is 8×2048.
  - W_K is 2048×512 → For LoRA, A is 8×2048, B is 8×512...
  - W_V is 2048×512
  - W_0 is 2048×2048

As such:

- LoRA parameters per layer are $8\times6\times2048 + 8\times2\times512 = 106'496$.
- EoRA parameters per layer are $32\times6\times2048 + 32\times2\times512 = 425'984$.

Granularity of Width Pruning



Pruned-only Configurations Results

Config	Closed (%) ↑	Open (%) ↑	PPL ↓	#Params (M)
Baseline	*50.6*	*81.8*	*26.6*	*1235.8*
D2	20.0	53.0	55.2	1114.2
D4	5.6	14.2	230.8	992.5
D6	2.0	2.4	2.1e3	870.9
D8	1.2	1.4	2.3e5	749.2
W1:2	1.8	2.2	471.7	749.3
W2:4	7.0	17.2	161.1	749.3
W1:8	47.8	80.8	27.2	1114.2
W4:8	12.6	42.4	81.3	749.3
W1:16	**49.2**	**81.8**	**26.8**	1175.0
W8:16	16.2	52.8	64.1	749.3
W12:16	0.2	0.2	1.8e3	506.0
D2 + W1:2	0.4	0.4	694.3	688.4

Config	Closed (%) ↑	Open (%) ↑	PPL ↓	#Params (M)
D2 + W2:4	3.8	5.0	309.3	688.4
D2 + W1:8	20.2	53.6	57.2	1007.7
D2 + W4:8	6.4	15.2	185.4	688.4
D2 + W1:16	19.6	51.6	55.7	1061.0
D2 + W8:16	7.6	23.4	140.0	688.4
D4 + W1:8	5.2	14.8	244.7	901.3
D4 + W4:8	3.0	2.6	774.6	627.6
D4 + W1:16	4.6	14.6	235.3	946.9
D4 + W8:16	2.6	6.2	591.6	627.6
D6 + W1:8	2.0	1.6	2.4e3	794.9
D6 + W4:8	0.8	1.0	7.7e4	566.8
D6 + W1:16	1.8	1.6	2.0e3	832.9
D8 + W4:8	0.8	0.4	5.5e5	506.0
D8 + W3:4	0.4	0.4	1.0e6	**384.3**

Optimized Configurations Results

Config	LoRA	Quant	Closed (%) ↑	Open (%) ↑	PPL ↓	Size (GB)
Baseline	*None*	*No*	*50.6*	*81.8*	*26.6*	*2.30*
D2	TriviaQA	No	27.8	60.2	51.1	2.08
D2	TriviaQA	Yes	23.9	54.6	55.0	0.92
D2	TriviaQA	EoRA	24.4	58.4	54.6	0.92
W1:2	WikiText	No	10.0	32.0	40.3	2.30
W1:2	WikiText	Yes	9.1	27.3	42.2	0.98
W1:2	WikiText	EoRA	8.8	28.3	41.8	0.98
W1:2	TriviaQA	EoRA	11.7	25.7	152.5	0.98
W1:8	WikiText	No	**47.8**	**80.8**	**15.7**	2.30
W1:8	WikiText	Yes	39.0	74.8	16.9	0.98
W1:8	TriviaQA	EoRA	42.0	77.2	32.9	0.98
W4:8	WikiText	No	17.0	51.1	26.0	2.30
W4:8	WikiText	Yes	15.7	45.3	27.4	0.98
W8:16	TriviaQA	Yes	20.8	55.7	60.7	0.98
W8:16	WikiText	Yes	17.5	52.1	25.4	0.98
W8:16	WikiText	EoRA	17.8	54.9	25.1	0.98
D2 + W1:8	WikiText	EoRA	22.3	59.6	23.0	0.92
D2 + W1:8	TriviaQA	No	27.4	60.4	51.9	2.08
D2 + W1:16	TriviaQA	Yes	25.2	56.3	55.5	0.92
D2 + W2:4	WikiText	Yes	10.6	30.8	41.5	0.92
D2 + W2:4	WikiText	EoRA	10.4	29.8	40.7	0.92
D4 + W2:4	WikiText	Yes	7.2	25.8	52.2	0.86
D4 + W2:4	WikiText	EoRA	7.6	26.3	51.2	0.86
D4 + W4:8	TriviaQA	No	10.5	20.8	210.9	1.85
D6 + W1:16	WikiText	No	11.5	25.6	37.8	1.62
D6 + W1:16	TriviaQA	No	16.8	27.5	310.5	1.62
D8 + W3:4	WikiText	No	0.2	1.3	278.4	1.40
D8 + W3:4	WikiText	Yes	0.5	1.3	289.5	**0.74**
D8 + W3:4	TriviaQA	No	0.3	0.3	7.3e3	1.40

Text Generation Examples

Config	Generated Text
Baseline Instruct	Ross Bagdasarian Sr. was an American singer, songwriter, and record producer who was known for his hit song "The Chipmunk Song"
W1:2 + W	The Chipmunks are the only group of the world's most popular and influential rock bands. The group is known for its unique...
W1:2 + W + Q	The Chipmunks are a group of children who are born in the early 20th century, who are the descendants of the family...
W1:8 + T	Ross Bagdasarian Jr. (disambiguation) - film producer and singer. He was born Ross Bagdasarian, Jr. in 1936 and died...
W1:8 + W	Ross Bagdasarian Jr., also known as David Seville, is the voice behind the Chipmunks. He was born on May 24, 1943
W1:8 + W + Q	The Chipmunk Man, also known as The Chipmunk Man, is a 2005 American comedy film directed by Chris Weigman
W4:8 + W	The Chipmunks were created by the British cartoonist and writer Arie Bellosi
W8:16 + W + Q	The Chipmunks were created by the British animation studio Aardust Productions. The studio was founded in 1991 by Pet...
D2 + W	The Chipmunks were created by Tom Pollock, a British animator and cartoonist. Pollock was inspired by the 1950s cartoons
D2 + W1:16 + W	The Chipmunks were created by Tom Pollock, a British animator and cartoonist. Pollock was inspired by the 1950s cartoons
D4 + W	The Chipmunks were created by Canadian comedian and musician John L. Linder, who was a member of the Canadian comedy...
D4 + W + Q	The Chipmunks were the men behind the Chipmunks cartoon series. Chipy, the Chipmunks'main character, was a chipster w...
D6 + W1:16 + W	The Chipmunks, the original version of the cartoon, was created by the Canadian cartoonist John H. Griffith
D6 + W1:16 + W + Q	The Chipmunks, the original version of the cartoon, was created by the Canadian cartoonist Jeff Sarnick and the Ameri...

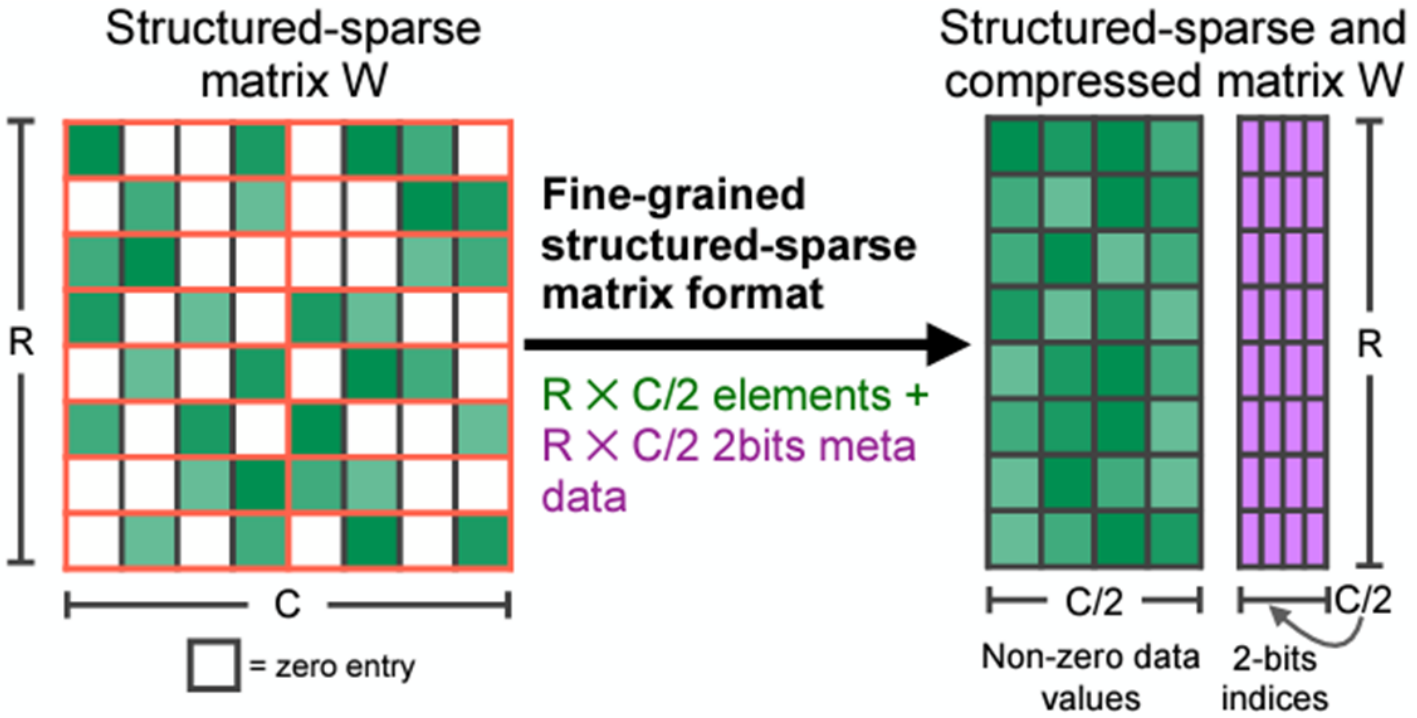
Text Generation Examples

Config	Generated Text
D8 + W + Q	The Chipmunks were the first to record the first recording of the song "The Chipmunks". The song was recorded by th...
D8 + W1:8 + W	The Chipmunks were the first to be recorded in the 1980s by the BBC. The Chipmunks were the first to be recorded in t...
D2 + W3:4 + W	The Chipmunks's first answer, the first answer,
D4 + W3:4 + W + Q	The first question, the question was the question of the question, and the question was not not so much of the questi...
D6 + W3:4 + T + Q	The World World World World World
D8 + W3:4 + W + Q	the 11thththththththththththththththth

Depth→LoRA→Width Experiment

Config	LoRA Type	D→L→W			D→W→L		
		TriviaQA (%) ↑		WikiText	TriviaQA (%) ↑		WikiText
		Closed	Open	PPL ↓	Closed	Open	PPL ↓
D2 + W1:8	WikiText	26.3	58.8	21.93	26.5	**63.8**	21.76
D2 + W1:8	TriviaQA	**28.2**	59.3	52.51	27.4	60.4	51.89
D2 + W1:16	WikiText	27.6	59.4	**21.42**	**28.1**	61.8	**21.42**
D2 + W1:16	TriviaQA	27.2	**59.6**	51.29	28.0	59.5	50.99
D2 + W4:8	WikiText	8.3	29.9	49.33	11.8	44.0	34.10
D2 + W4:8	TriviaQA	9.8	29.3	139.97	14.7	45.0	116.95
D6 + W1:16	WikiText	11.4	23.6	37.97	11.5	25.6	37.82
D6 + W1:16	TriviaQA	15.6	26.1	347.78	16.8	27.5	310.53
D6 + W4:8	WikiText	5.3	12.4	85.89	0.9	9.6	55.46
D6 + W4:8	TriviaQA	5.9	5.1	513.98	8.9	16.2	568.93
D6 + W6:8	WikiText	0.1	0.3	3623.92	1.7	5.8	139.97
D6 + W6:8	TriviaQA	0.1	0.1	87116.36	2.2	0.7	3273.94
D8 + W1:8	WikiText	5.1	15.3	58.80	5.8	16.8	58.58
D8 + W1:8	TriviaQA	8.6	0.4	2610.21	8.7	1.3	132836.59

Notes on Width Pruning



Notes on Width Pruning

Our model loading framework **cannot fully leverage** the benefits of Width Pruning, and instead requires **full dense allocation** regardless of zero weights.

- In our testing scenario, we used a cluster comprised of 4 $\times$ 1080 Ti, which have **Pascal** architectures.
- NVIDIA's structured sparsity optimization are compatible only with **Ampere and Hopper** architectures onwards.

- In addition, structured sparsity model loading is available only through **PyTorch nightlies** and by tinkering with the architecture using the **TorchAO** library.