Compressing Large Language Models to Any Size Without Re-Computation

Martin Genzel*, Patrick Putzky*, Pengfei Zhao*‡, Sebastian Schulze,
Mattes Mollenhauer, Robert Seidel, Stefan Dietzel, Thomas Wollmann

Merantix Momentum GmbH, Berlin, Germany

*equal contribution ‡now at Understandable Machine Intelligence Lab, ATB Potsdam, Germany

MERANTIX MOMENTUM

MORE WITH LESS

With funding from the: Federal Ministry of Research, Technology and Space

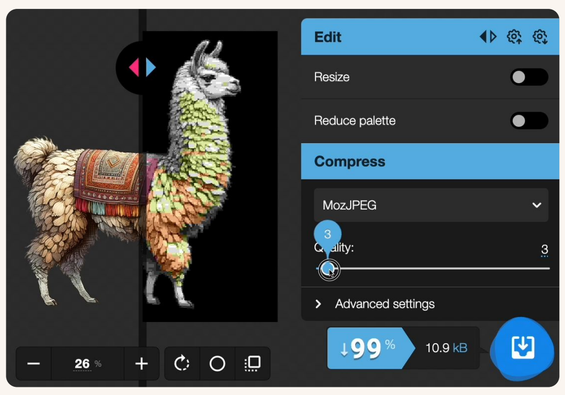
ICML International Conference On Machine Learning

Making model compression as simple as image compression

Is there a way to get instant access to compressed models of any size without re-computation? Yes! We advocate for **Any Compression,** which puts all expensive computations before the user's choice of compression rate. Our algorithm ACIP achieves this by building a score map from low-rank parametrizations of linear layers.

Problem

Image compression is very intuitive



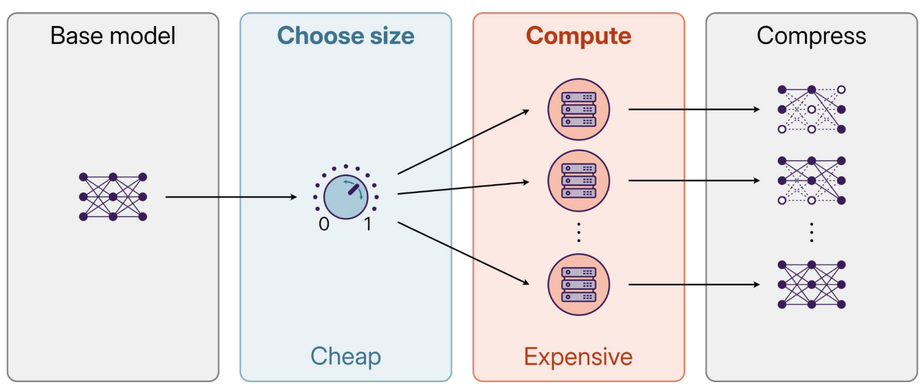
Users can adjust the **compression slider** based on **their needs.** They are not required to have a deep understanding of the underlying algorithm.

Model compression reality check: Quantization

Post-training compression for LLMs is effective, but creates a fundamental **trade-off between size and performance.** The process can feel like a **black box** for users.

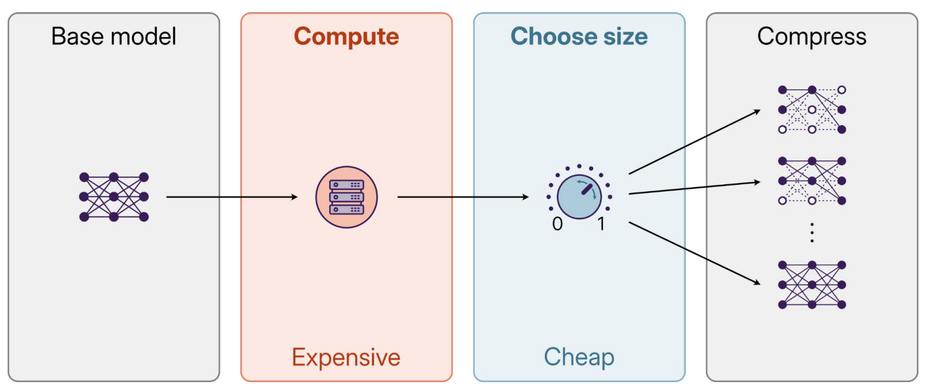
Solution

Conventional approaches require re-computation …



The usual process is inefficient: Pick one of **few preset target sizes**, run a costly calibration, and **repeat the entire process** for every new compression rate.

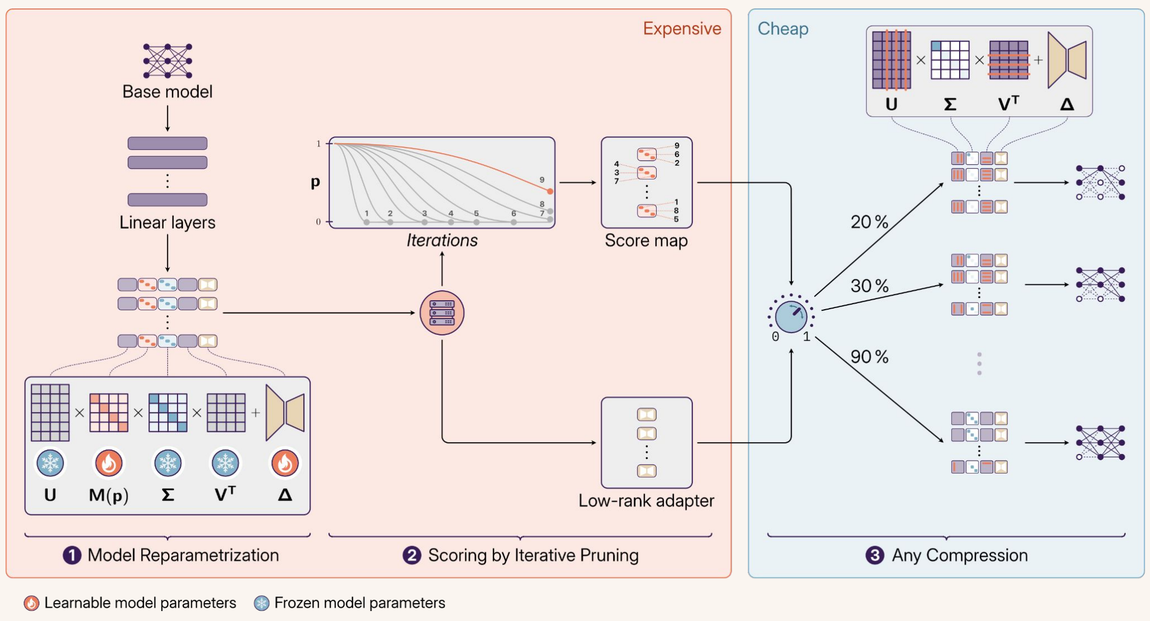
… Any Compression does not!



We propose **Any Compression:** Perform a single, **upfront computational step** that empowers users to generate a model at any size in **real-time,** without extra cost.

Method

Any Compression via Iterative Pruning (ACIP)



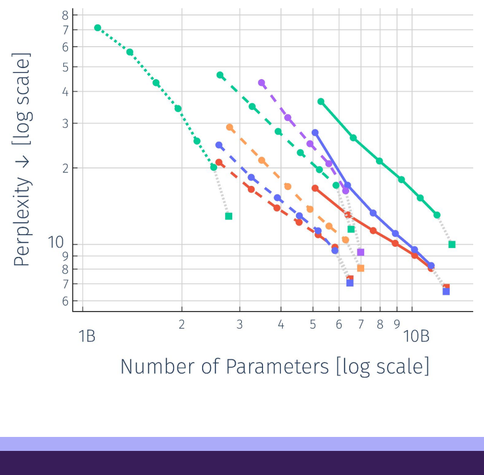
Key Idea: **Decouple** the pruning stage (calibration) from the compression stage.

Result: Instant access to models of any size, like a **slider** in image compression!

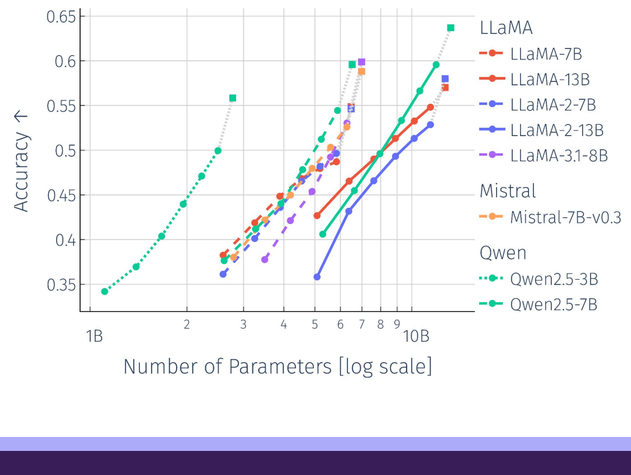
Results

ACIP produces consistent compression-performance trade-offs

Perplexity on C4



⌀ Accuracy on LM-Eval tasks



ACIP × Quantization (LLaMA-7B on C4)

Benchmark

Size	Method	C4 ↓	WikiText-2 ↓
100%	Original	**7.34**	**5.68**
80%	ASVD	15.93	11.14
	SVD-LLM	15.84	**7.94**
	ACIP (ours)	**10.92**	8.83
70%	ASVD	41.00	51.00
	SVD-LLM	25.11	**9.56**
	ACIP (ours)	**12.22**	10.35
60%	ASVD	1109.00	1407.00
	SVD-LLM	49.83	13.11
	ACIP (ours)	**13.91**	**12.46**
50%	ASVD	27925.00	15358.00
	SVD-LLM	118.57	23.97
	ACIP (ours)	**16.47**	**16.16**
40%	ASVD	43036.00	57057.00
	SVD-LLM	246.89	42.30
	ACIP (ours)	**21.05**	**23.99**

Paper

GitHub

Hugging Face