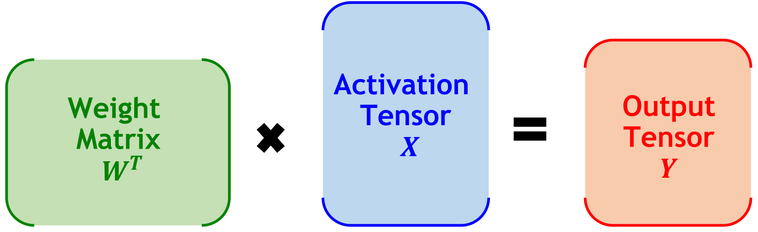
General Matrix Multiplication (GEMM)

$$Y = W^T X$$

⬆️ trainable parameters: ⬆️ expressivity ✅
⬆️ model size: ⬆️ inference cost ❌

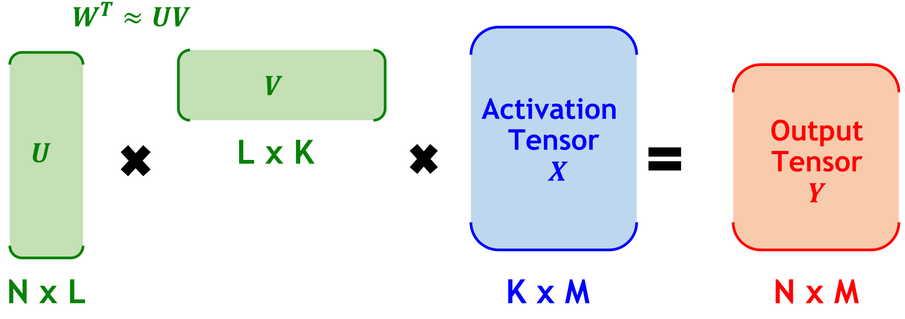


K = input dimension, N= output dimension, M = batch $*$ sequence

GEMM with Weight Decomposition
(e.g., truncated SVD)

$$Y \approx UVX$$

⬇️ trainable parameters: ⬇️ expressivity ❌
⬇️ model size: ⬇️ inference cost ✅

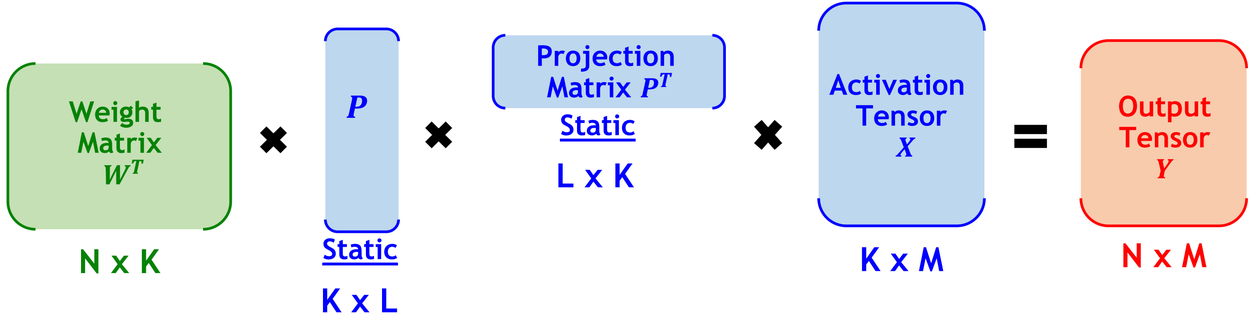


K = input dimension, N= output dimension, M = batch * sequence
L = intermediate dimension

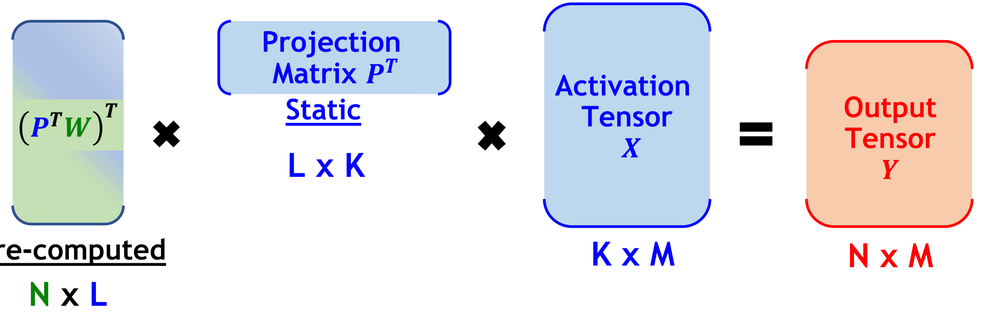
ESPACE: Activation Dimensionality Reduction Using Static Projections

$$Y \approx W^T(PP^TX) = \left(P^TW\right)^T(P^TX)$$

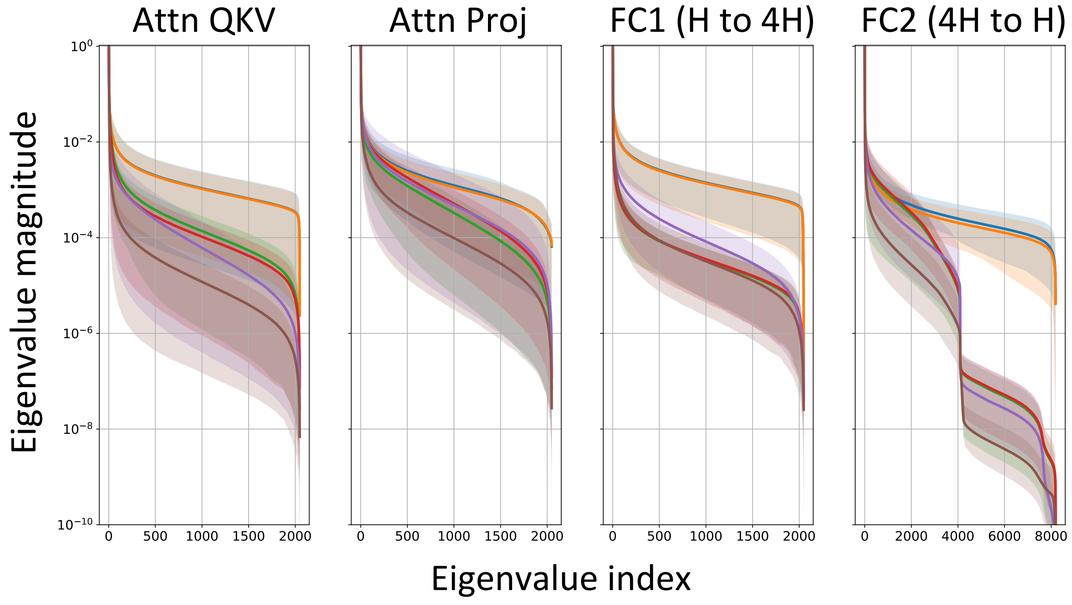
During Training: ⬆ trainable parameters: ⬆ expressivity ✅

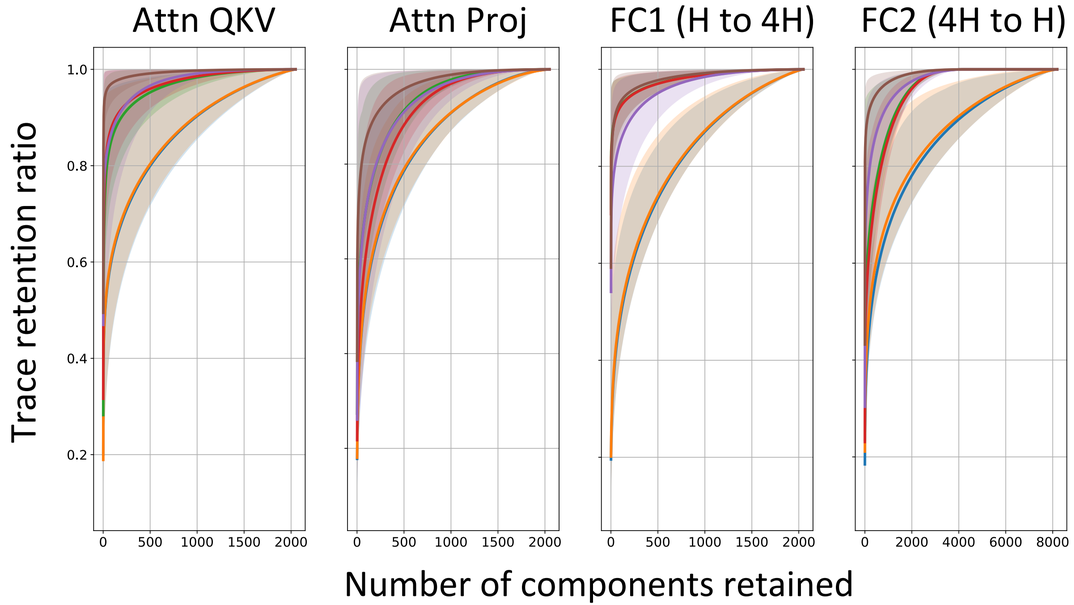


During Inference: ⬇ model size: ⬇ inference cost ✅

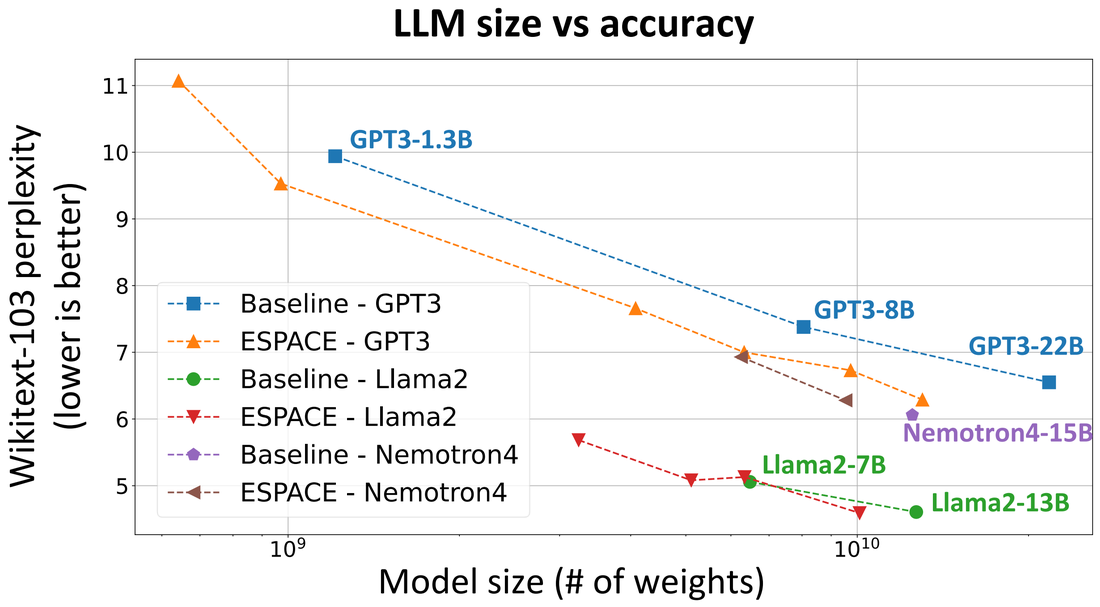


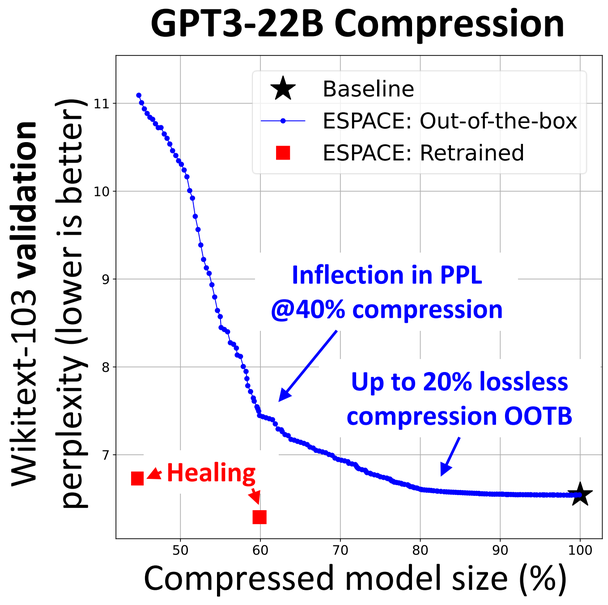
K = input dimension, N= output dimension, M = batch $*$ sequence, L = intermediate dimension

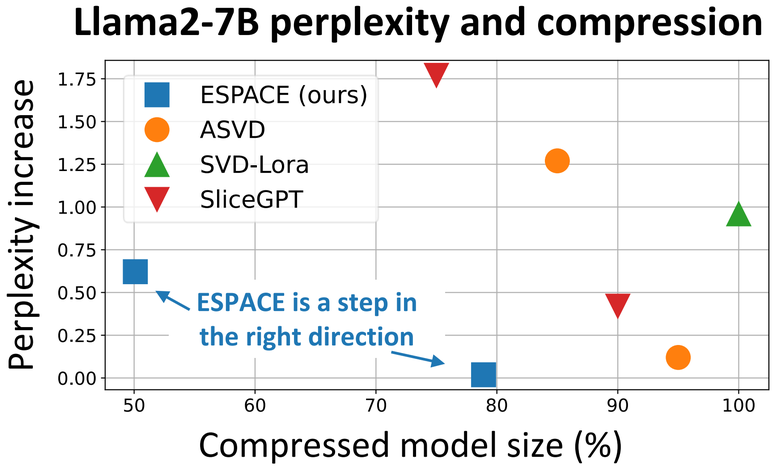

Attn QKV
Attn Proj
FC1 (H to 4H)
FC2 (4H to H)
Eigenvalue magnitude
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
0
500
1000
1500
2000
0
2000
4000
6000
8000
Eigenvalue index


Attn QKV
Attn Proj
FC1 (H to 4H)
FC2 (4H to H)
Trace retention ratio
1.0
0.8
0.6
0.4
0.2
0
500
1000
1500
2000
0
2000
4000
6000
8000
Number of components retained


MSE
NMSE
GO-MSE
GO-MSE w/ L_2-normalization
NL-MSE
NL-MSE with L_2-normalization


LLM size vs accuracy
Wikitext-103 perplexity
(lower is better)
Model size (# of weights)
11
10
9
8
7
6
5
10^9
10^{10}
GPT3-1.3B
GPT3-8B
GPT3-22B
Nemotron4-15B
Llama2-7B
Llama2-13B
Baseline - GPT3
ESPACE - GPT3
Baseline - Llama2
ESPACE - Llama2
Baseline - Nemotron4
ESPACE - Nemotron4

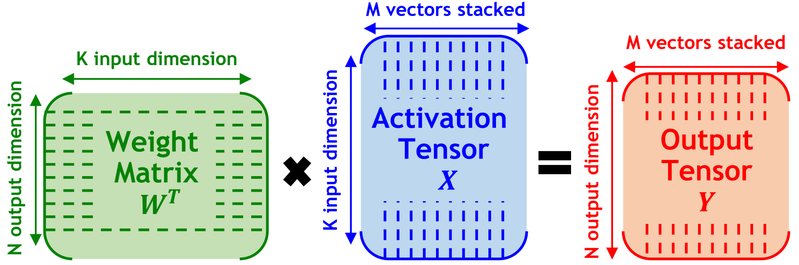

GPT3-22B Compression
Baseline
ESPACE: Out-of-the-box
ESPACE: Retrained
Inflection in PPL
@40% compression
Up to 20% lossless
compression OOTB
Healing
Wikitext-103 validation
perplexity (lower is better)
Compressed model size (%)
11
10
9
8
7
50
60
70
80
90
100

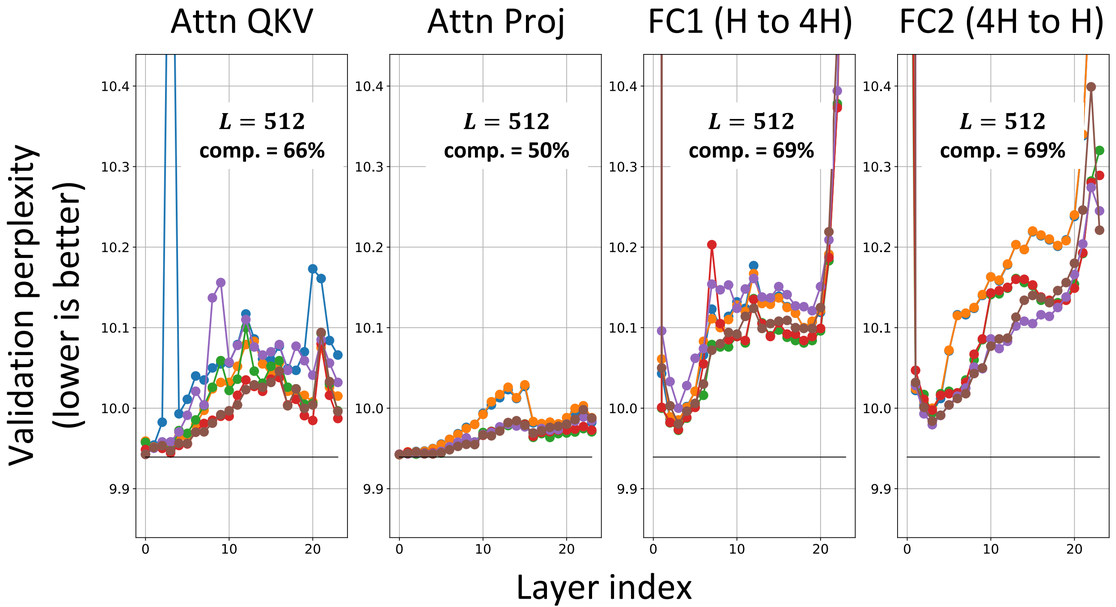

Llama2-7B perplexity and compression
Perplexity increase
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
ESPACE (ours)
ASVD
SVD-Lora
SliceGPT
ESPACE is a step in
the right direction
50
60
70
80
90
100
Compressed model size (%)

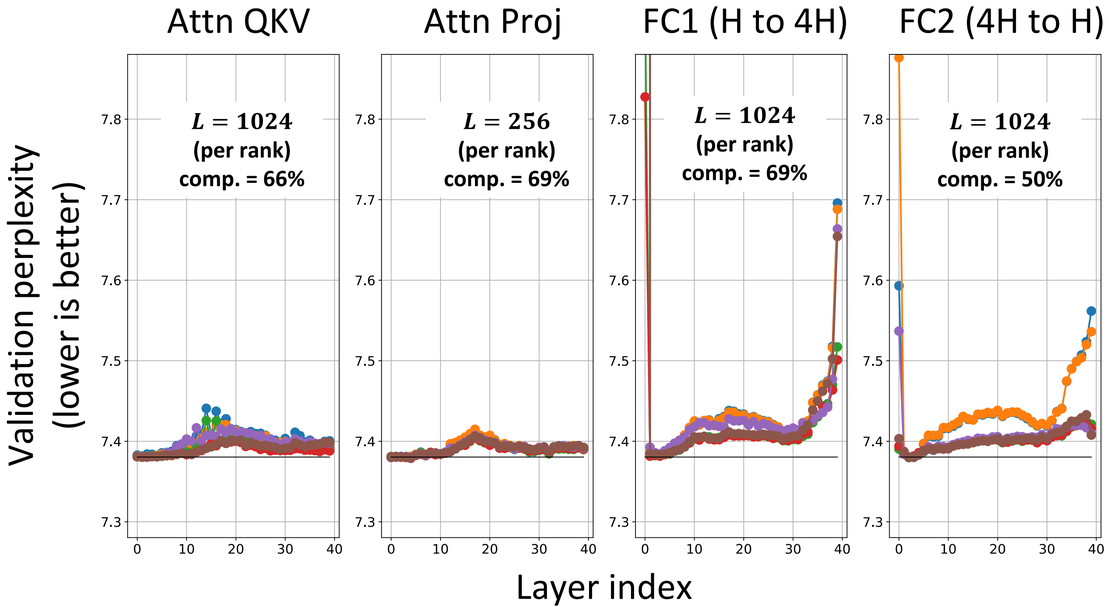
General Matrix Multiplication (GEMM)

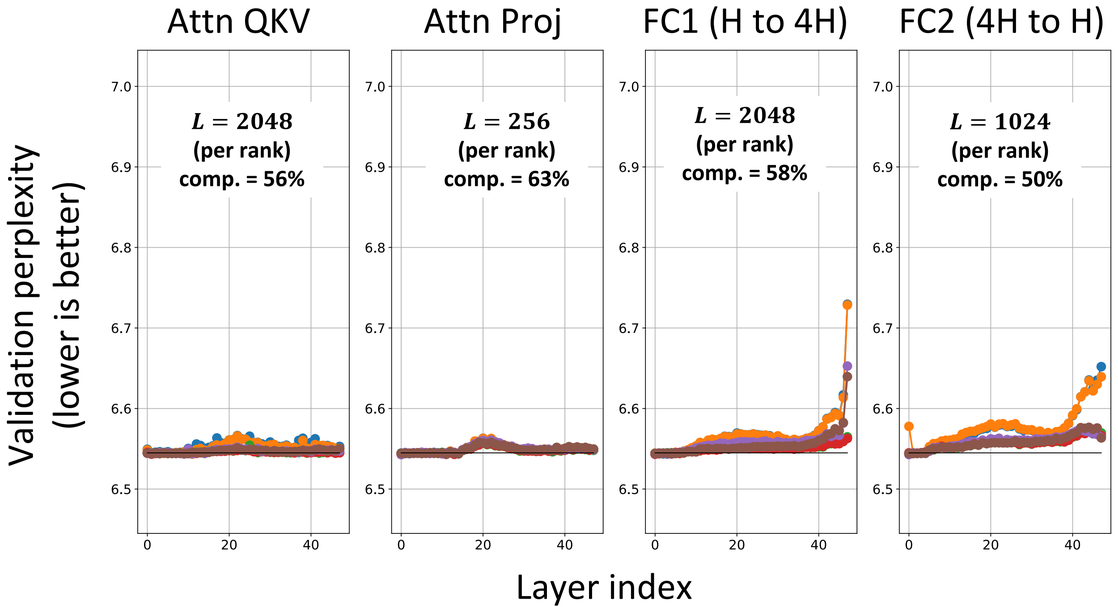
$$Y = W^T X$$

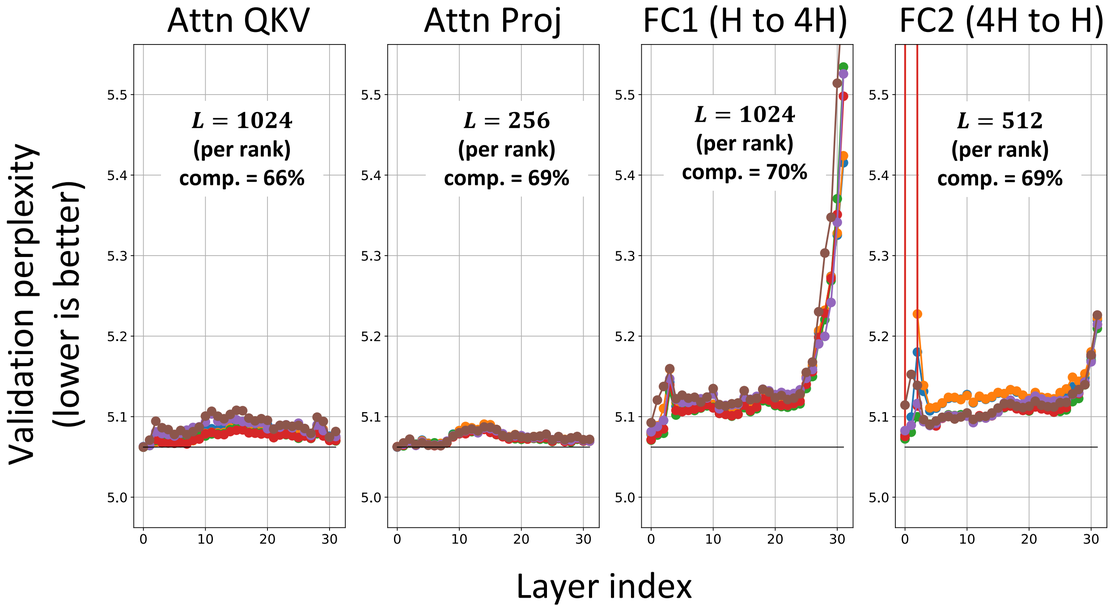
⬆️ trainable parameters: ⬆️ expressivity ✅
⬆️ model size: ⬆️ inference cost ❌

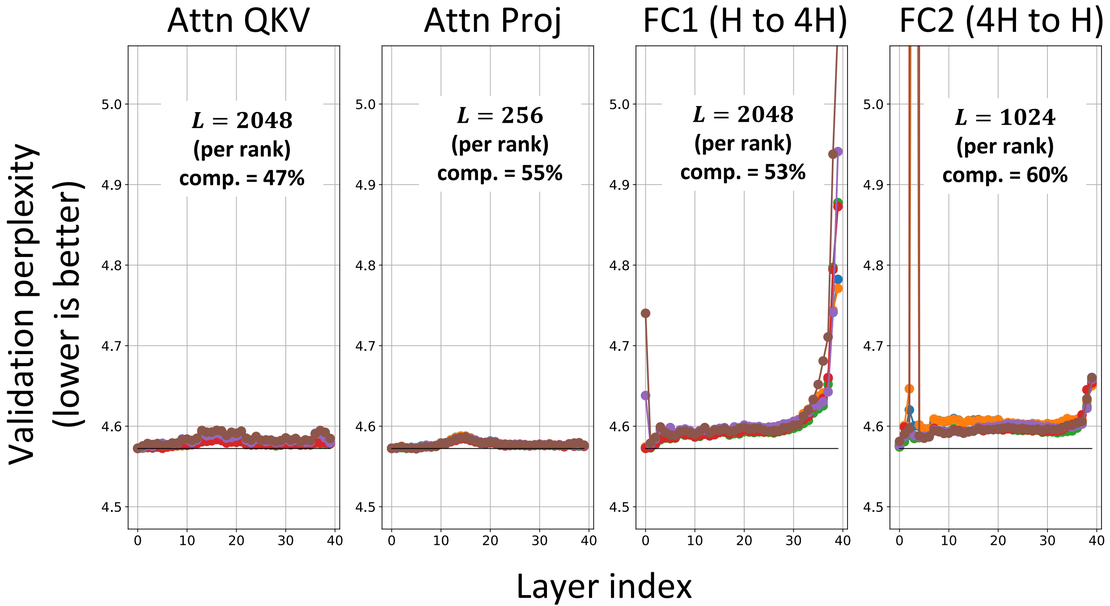


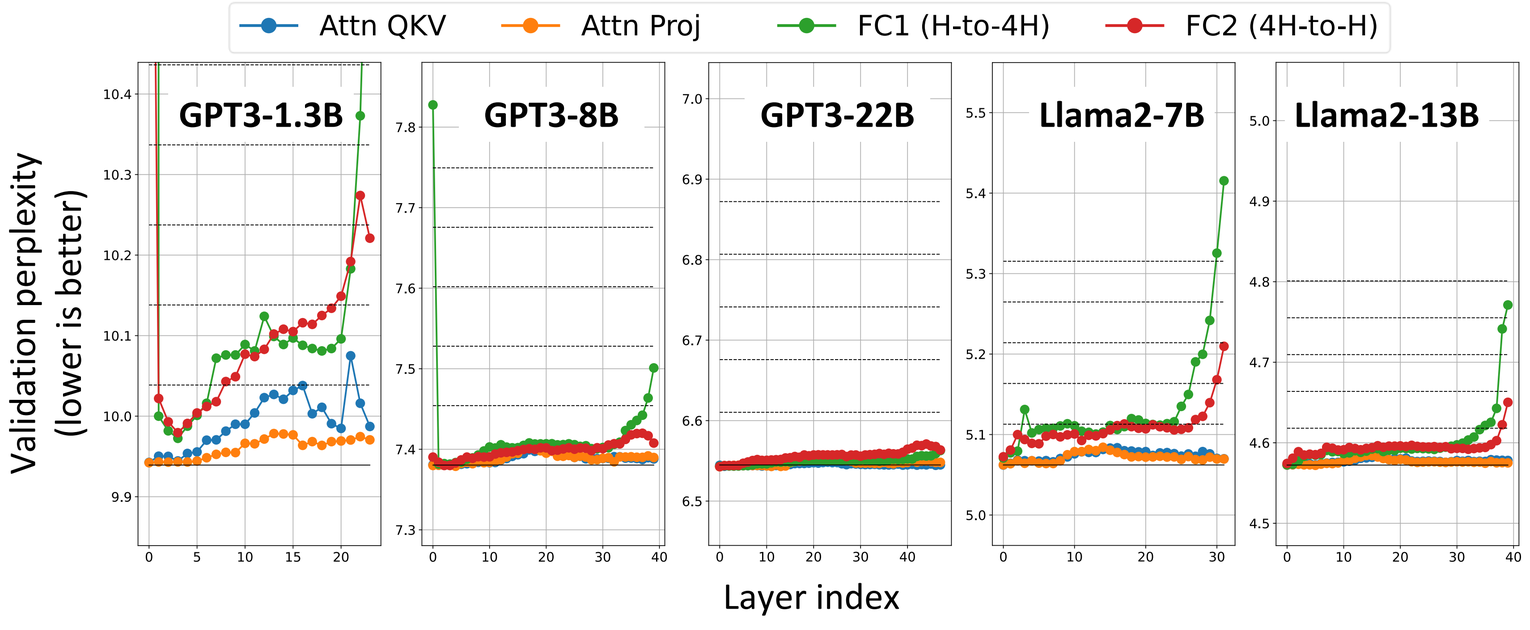

Attn QKV
Attn Proj
FC1 (H to 4H)
FC2 (4H to H)
Validation perplexity
(lower is better)
$L = 512$
comp. = 66%
$L = 512$
comp. = 50%
$L = 512$
comp. = 69%
$L = 512$
comp. = 69%
10.4
10.3
10.2
10.1
10.0
9.9
0
10
20
Layer index

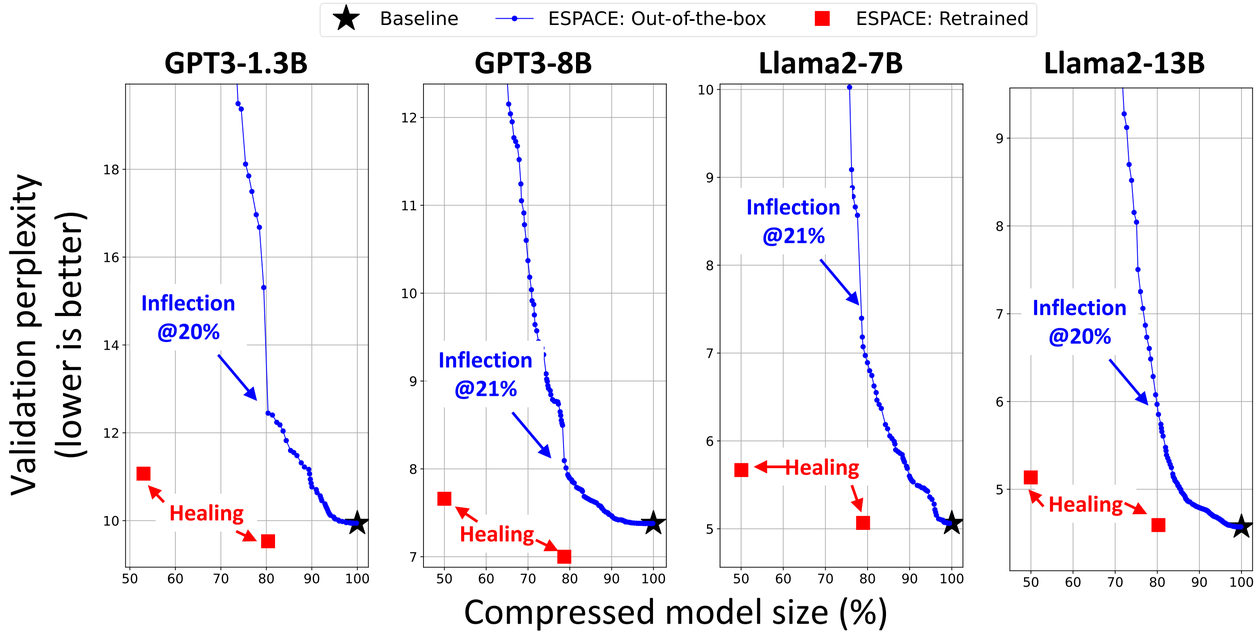

Attn QKV
Attn Proj
FC1 (H to 4H)
FC2 (4H to H)
Validation perplexity
(lower is better)
$L = 1024$
(per rank)
comp. = 66%
$L = 256$
(per rank)
comp. = 69%
$L = 1024$
(per rank)
comp. = 69%
$L = 1024$
(per rank)
comp. = 50%
7.8
7.7
7.6
7.5
7.4
7.3
0
10
20
30
40
Layer index


Attn QKV
Attn Proj
FC1 (H to 4H)
FC2 (4H to H)
Validation perplexity
(lower is better)
$L = 2048$
(per rank)
comp. = 56%
$L = 256$
(per rank)
comp. = 63%
$L = 2048$
(per rank)
comp. = 58%
$L = 1024$
(per rank)
comp. = 50%
7.0
6.9
6.8
6.7
6.6
6.5
0
20
40
Layer index

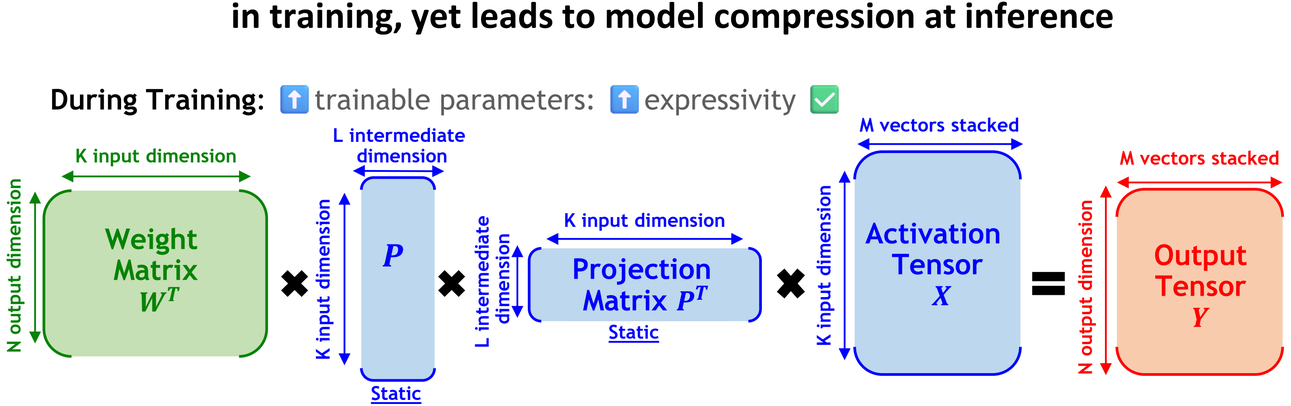

Attn QKV
Attn Proj
FC1 (H to 4H)
FC2 (4H to H)
Validation perplexity
(lower is better)
$L = 1024$
(per rank)
comp. = 66%
$L = 256$
(per rank)
comp. = 69%
$L = 1024$
(per rank)
comp. = 70%
$L = 512$
(per rank)
comp. = 69%
5.5
5.4
5.3
5.2
5.1
5.0
0
10
20
30
Layer index

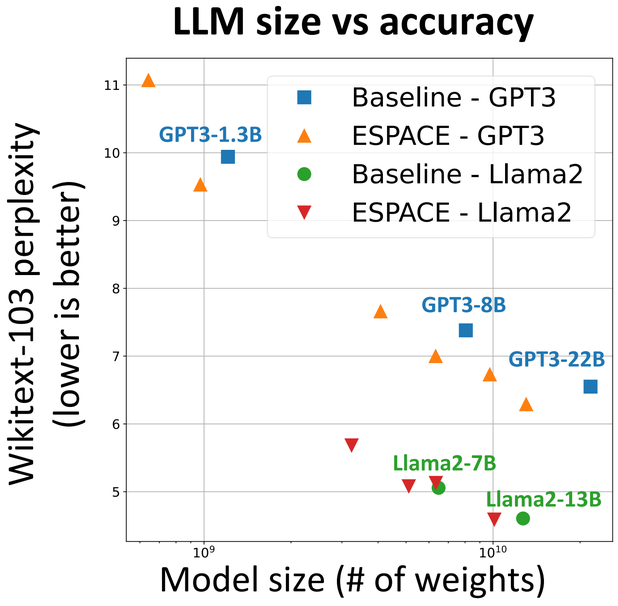

Attn QKV
Attn Proj
FC1 (H to 4H)
FC2 (4H to H)
Validation perplexity
(lower is better)
$L = 2048$
(per rank)
comp. = 47%
$L = 256$
(per rank)
comp. = 55%
$L = 2048$
(per rank)
comp. = 53%
$L = 1024$
(per rank)
comp. = 60%
5.0
4.9
4.8
4.7
4.6
4.5
0
10
20
30
40
Layer index

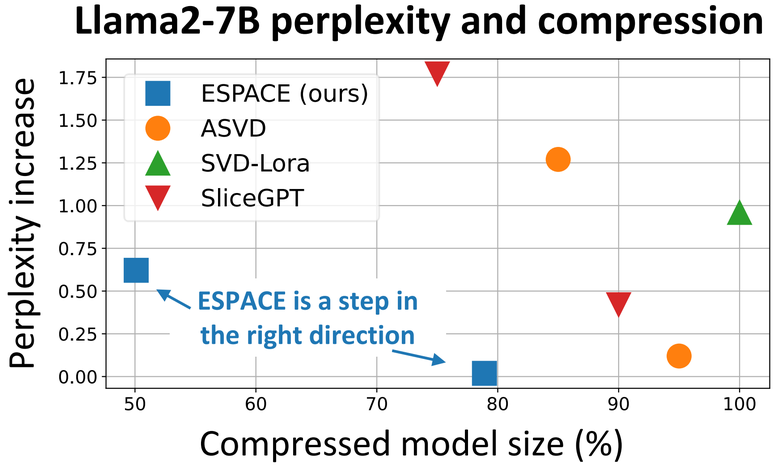

Attn QKV
Attn Proj
FC1 (H-to-4H)
FC2 (4H-to-H)
GPT3-1.3B
GPT3-8B
GPT3-22B
Llama2-7B
Llama2-13B
Validation perplexity
(lower is better)
Layer index


Baseline
ESPACE: Out-of-the-box
ESPACE: Retrained
GPT3-1.3B
GPT3-8B
Llama2-7B
Llama2-13B
Validation perplexity
(lower is better)
Inflection
@20%
Inflection
@21%
Inflection
@21%
Inflection
@20%
Healing
Healing
Healing
Healing
Compressed model size (%)

Activation projection does not interfere with weight learnability in training, yet leads to model compression at inference



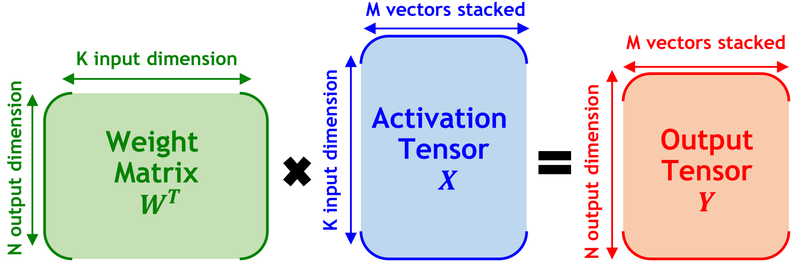

LLM size vs accuracy
Wikitext-103 perplexity
(lower is better)
Model size (# of weights)
Baseline - GPT3
ESPACE - GPT3
Baseline - Llama2
ESPACE - Llama2
GPT3-1.3B
GPT3-8B
GPT3-22B
Llama2-7B
Llama2-13B
11
10
9
8
7
6
5
10^9
10^{10}


Llama2-7B perplexity and compression
Perplexity increase
Compressed model size (%)
ESPACE (ours)
ASVD
SVD-Lora
SliceGPT
ESPACE is a step in
the right direction
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
50
60
70
80
90
100

General Matrix Multiplication (GEMM)

$$Y = W^T X$$

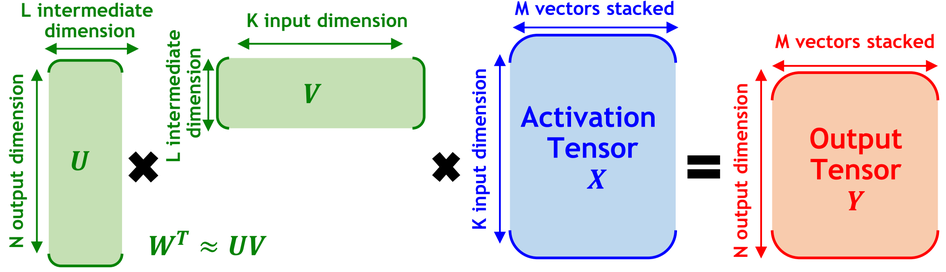
⬆️ trainable parameters: ⬆️ expressivity ✅
⬆️ model size: ⬆️ inference cost ❌



GEMM with Weight Decomposition (e.g., truncated SVD)

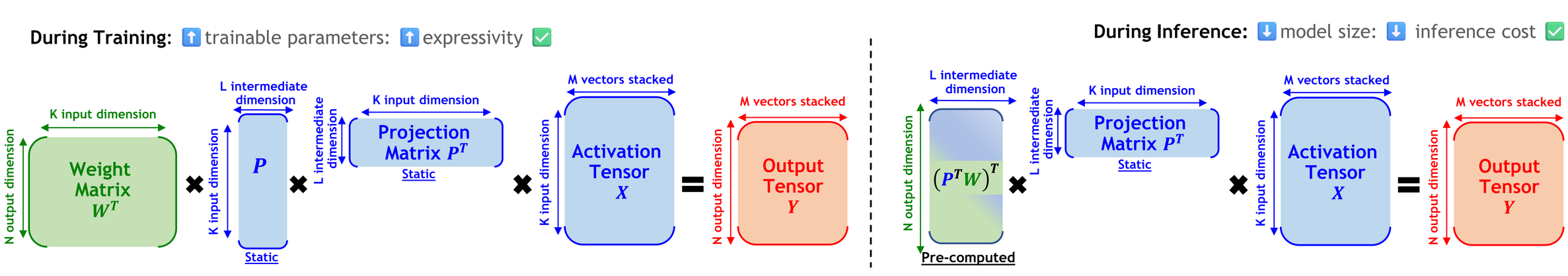
$$Y \approx UVX$$

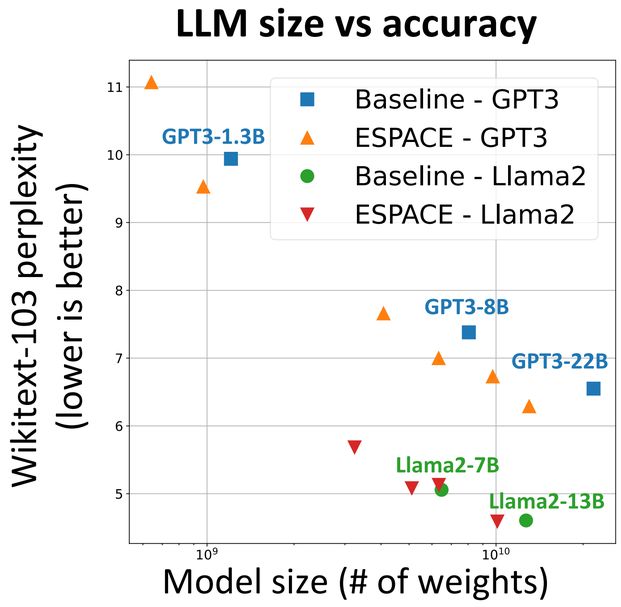
⬇️ trainable parameters: ⬇️ expressivity ❌
⬇️ model size: ⬇️ inference cost ✅



ESPACE: Activation Dimensionality Reduction Using Static Projections

$$Y \approx W^T(PP^TX) = \left(P^TW\right)^T(P^TX)$$




LLM size vs accuracy
Wikitext-103 perplexity
(lower is better)
Model size (# of weights)
Baseline - GPT3
ESPACE - GPT3
Baseline - Llama2
ESPACE - Llama2
GPT3-1.3B
GPT3-8B
GPT3-22B
Llama2-7B
Llama2-13B
11
10
9
8
7
6
5
10^9
10^{10}