COMPARISON OF LORA, DORA, AND QLORA

By Alisha Rath

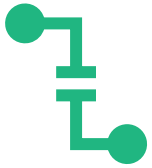
Presented to : Prof. Dr. Chris Pollett

INTRODUCTION

- **Parameter-Efficient Fine-Tuning** helps adapt large pre-trained language models for specific tasks.
- Why Compare LoRA, DORA, and QLoRA?
  - All three methods focus on reducing computational costs.
  - They introduce different techniques for updating parameters efficiently.
  - Understanding their strengths and weaknesses aids in selecting the right method for specific applications.

WHAT IS LORA (LOW-RANK ADAPTATION)?

Definition: LoRA introduces low-rank matrices for fine-tuning while keeping most of the model parameters frozen.



Key Features:

Parameter Efficiency: Reduces the number of parameters to be trained.

Training Speed: Faster training due to fewer trainable parameters.

Use Cases: NLP tasks like text classification, question answering.

Strengths: Good balance between performance and computational costs.

WHAT IS DORA?

Definition: DORA (Dynamic Offset Rank Adaptation) extends LoRA by allowing dynamic rank adjustments during training.

Key Features:

- **Adaptive Training:** Adjusts rank based on task complexity, offering flexibility.
- **Improved Generalization:** Helps the model adapt better to diverse datasets.
- **Use Cases:** Tasks requiring adaptability to data variations, such as image recognition.

Strengths: Better flexibility compared to LoRA in adapting to complex datasets.

WHAT IS QLORA?

QLoRA (Quantized Low-Rank Adaptation) combines quantization with low-rank adaptation.**Key Features:Quantization:** Compresses model weights, reducing memory usage.

Low-Rank Matrices: Similar to LoRA, it uses low-rank matrices for efficient fine-tuning.

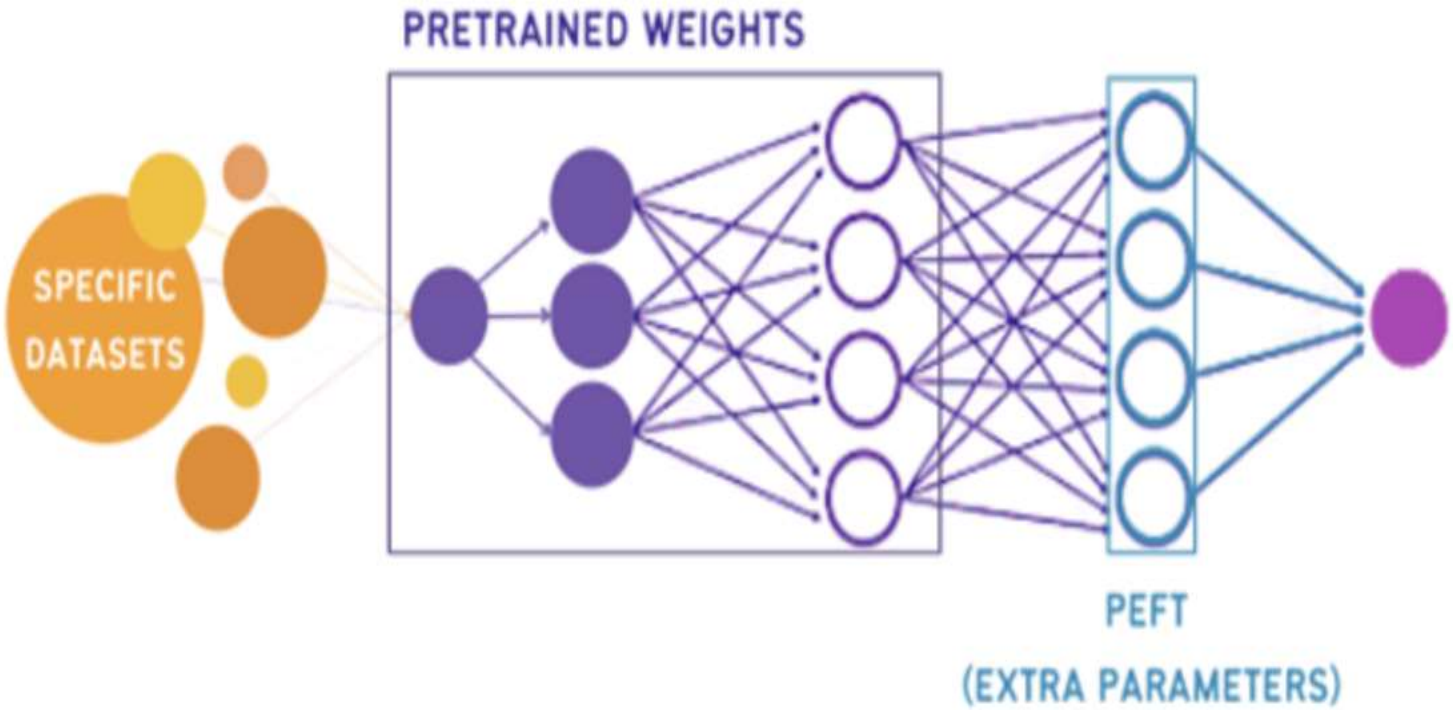
Use Cases: Resource-constrained environments like edge devices.

Strengths: Significantly reduces memory requirements, making it suitable for devices with limited resources.

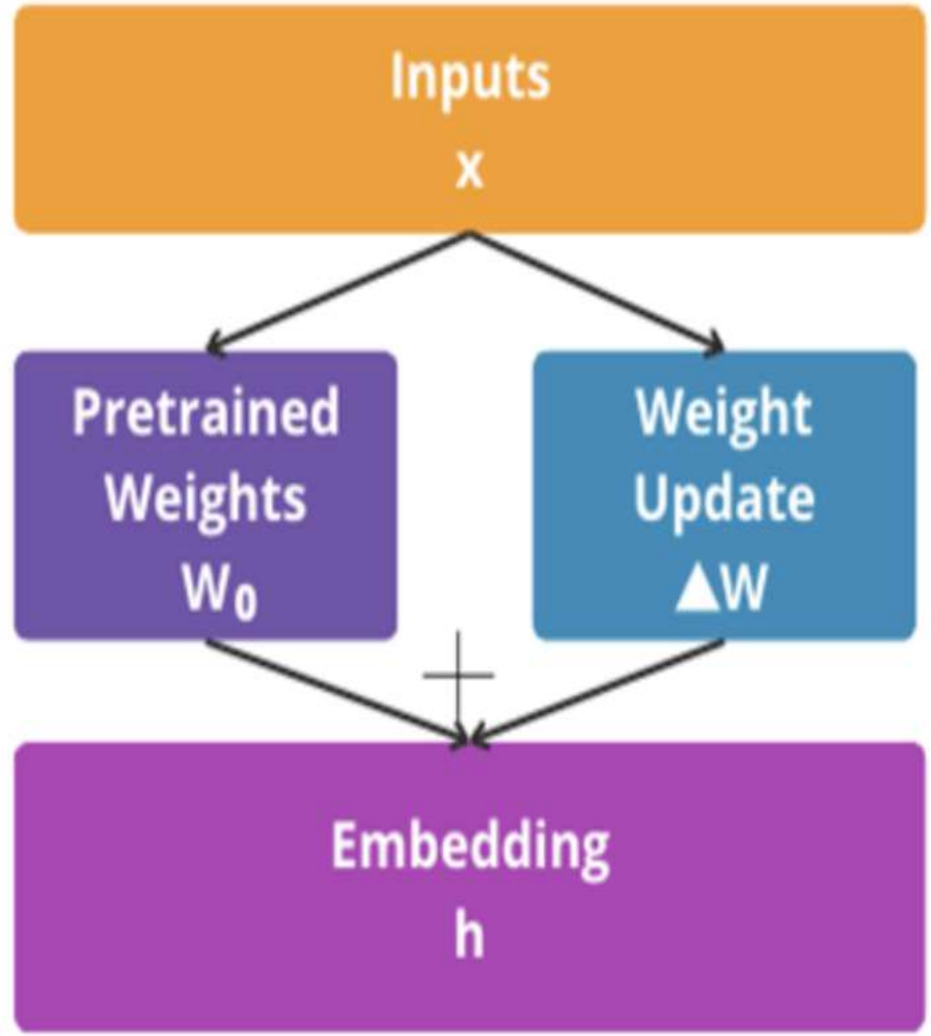
COMPARISON TABLE

Feature	LoRA	DORA	QLoRA
Adaptation	Low-rank matrices	Dynamic rank adjustment	Low-rank with quantization
Flexibility	Static rank	Adaptive rank	Static rank with quantization
Efficiency	Lower memory usage	Moderate memory usage	Lowest memory usage
Use Cases	NLP tasks	Complex data adaptation	Edge and memory-constrained tasks

PEFT VISUAL REPRESENTATION



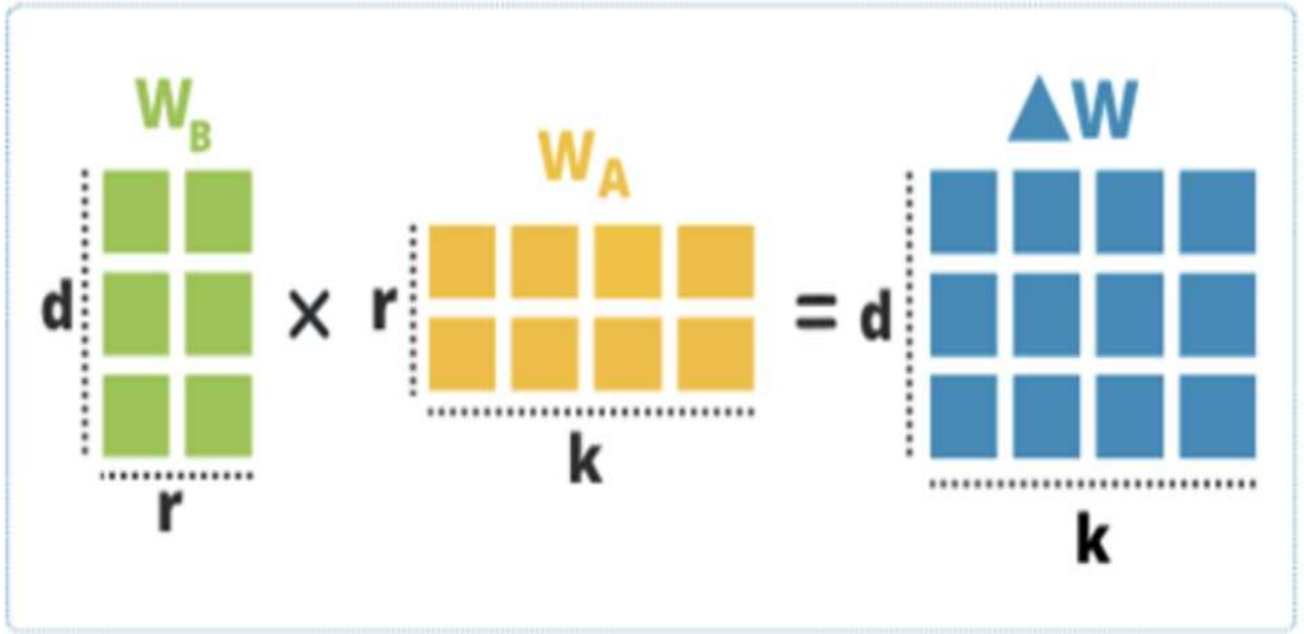
PEFT MATHEMATICAL REPRESENTATION



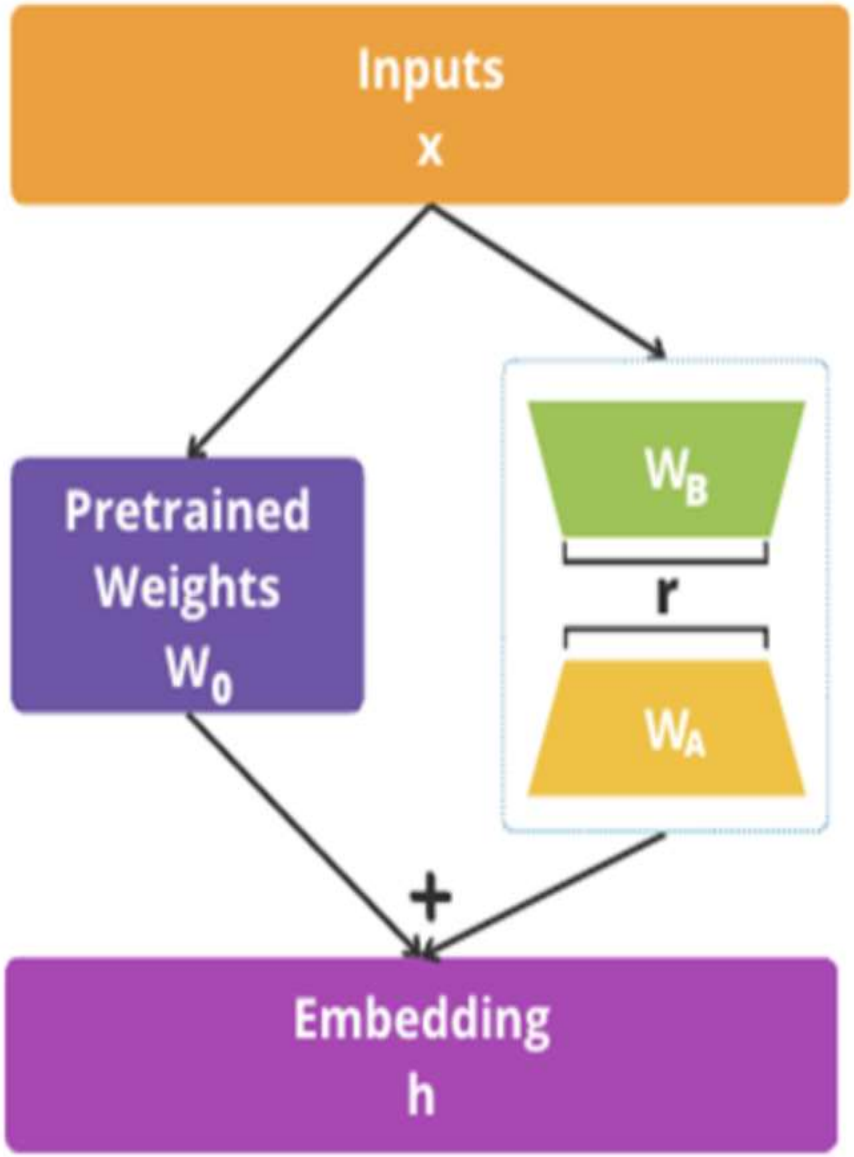
$$h = W_0 x + \Delta W x$$

LORA WEIGHTS ADDITION

LORA MATRIX DECOMPOSITION

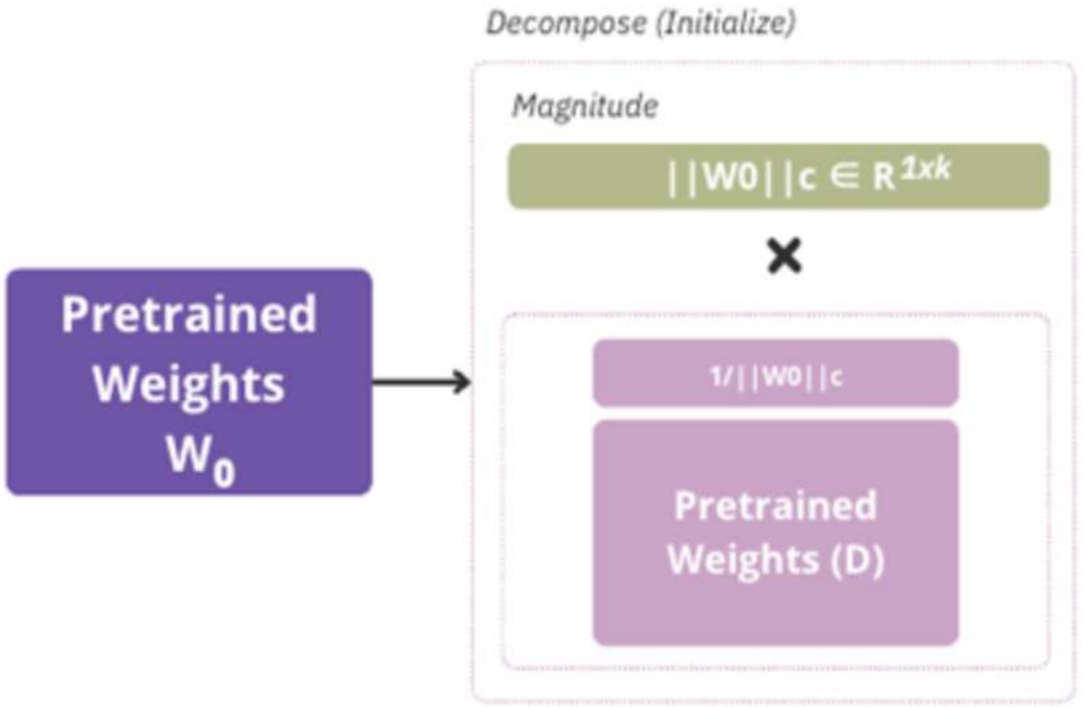


LORA'S MATHEMATICAL REPRESENTATION



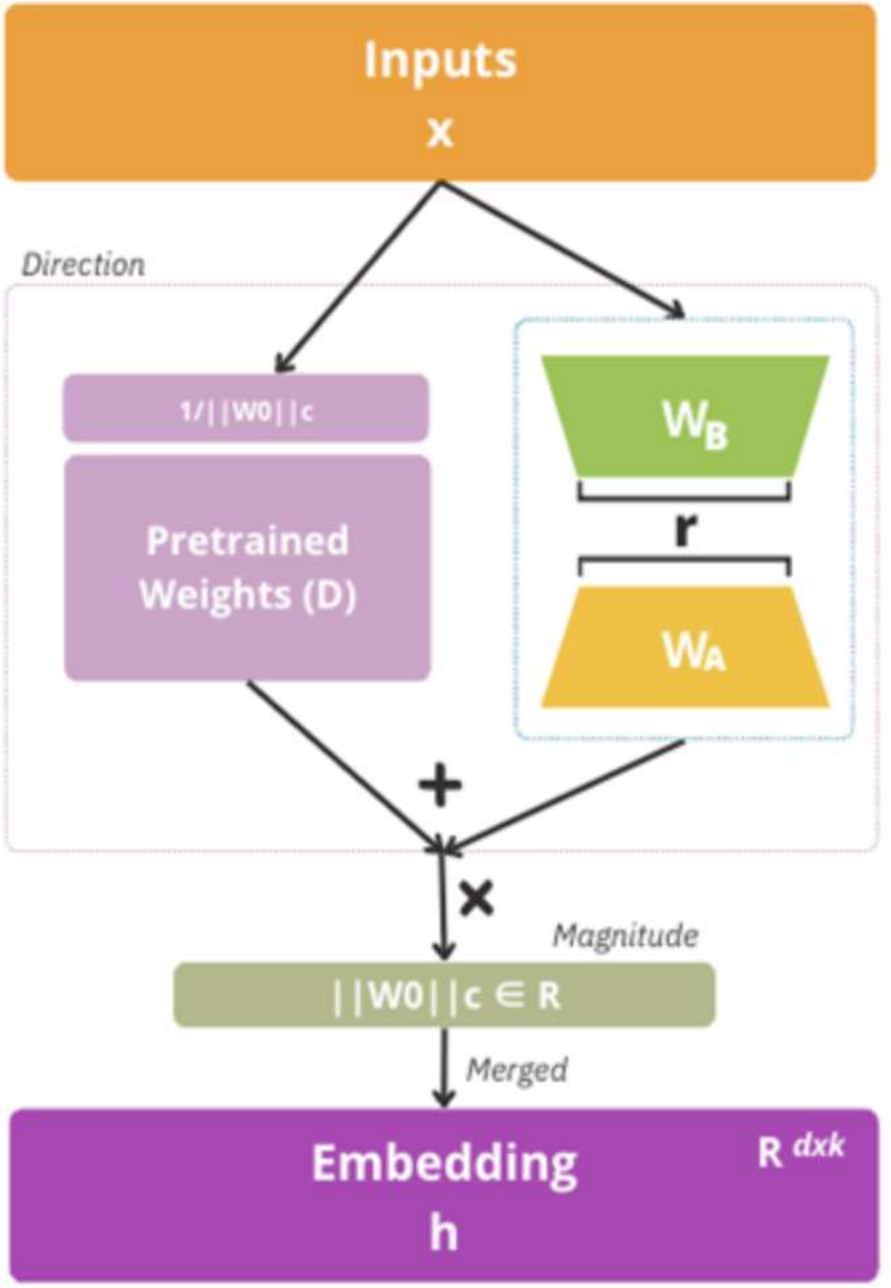
$$h = W_0 x + W_A W_B x$$

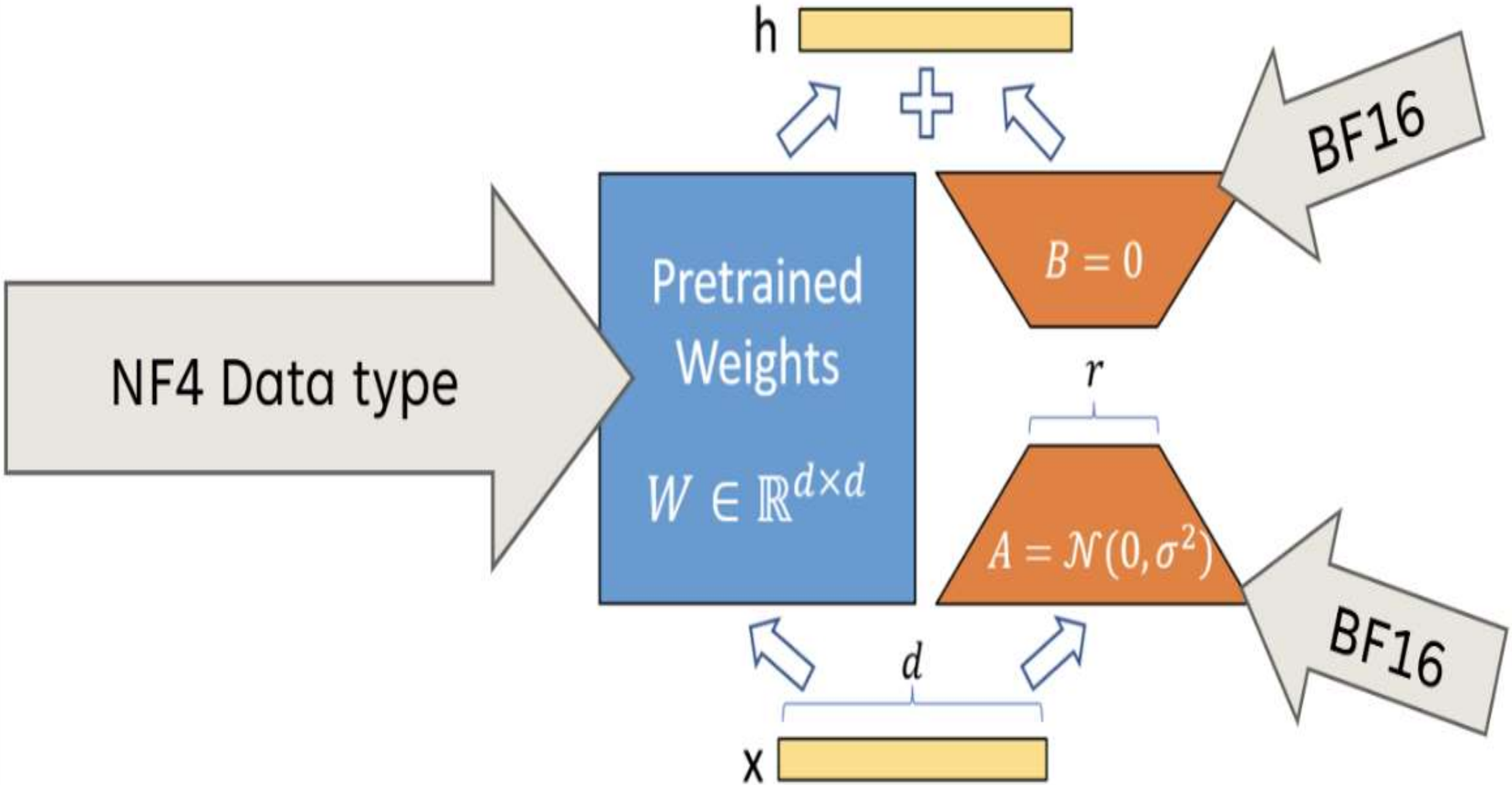
DORA WEIGHTS DECOMPOSITION

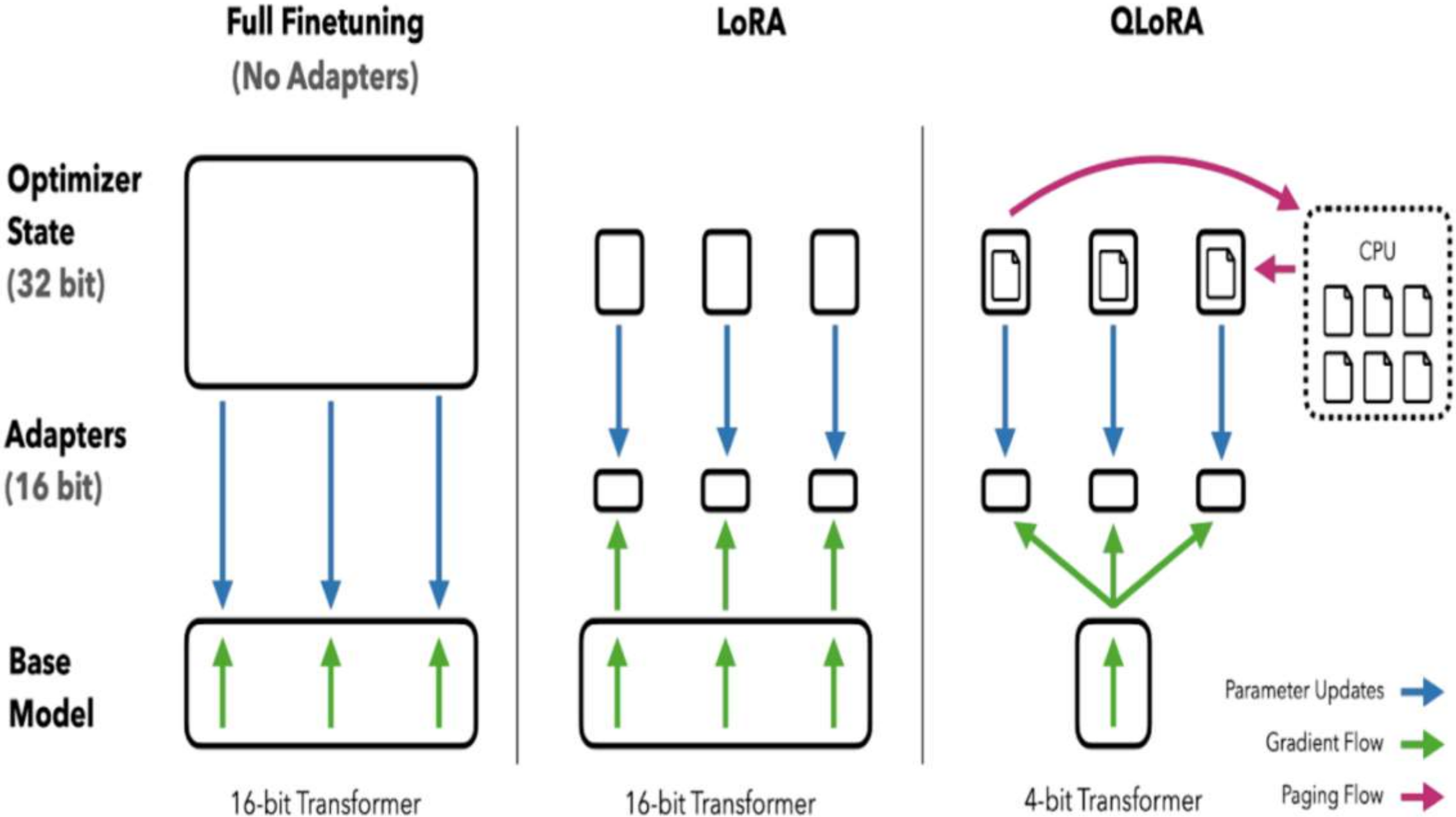


$$h = \text{Magnitude} \cdot \text{Direction} \left(W_0 x + W_B W_A x \right)$$

DORA'S MATHEMATICAL REPRESENTATION




h
NF4 Data type
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
BF16
r
$A = \mathcal{N}(0, \sigma^2)$
BF16
d
x


Full Finetuning
(No Adapters)
LoRA
QLoRA
Optimizer
State
(32 bit)
Adapters
(16 bit)
Base
Model
CPU
16-bit Transformer
16-bit Transformer
4-bit Transformer
Parameter Updates
Gradient Flow
Paging Flow

STRENGTHS & LIMITATIONS

LoRA:

- **Strengths:** Easy to implement, computationally efficient.
- **Limitations:** Less adaptable to varying data complexities.

DORA:

- **Strengths:** Flexible and adaptive, better suited for varying tasks.
- **Limitations:** Higher computational overhead compared to LoRA.

QLoRA:

- **Strengths:** Optimal for memory-limited environments, significant storage savings.
- **Limitations:** Quantization may introduce small losses in accuracy.

SCENARIO: LEGAL CASE PREDICTION - PERFORMANCE ANALYSIS

Objective: Predicting legal case outcomes using different fine-tuning methods.

Dataset: A legal dataset with case details divided into training and validation sets.

COMPARISON BASED ON ACCURACY:

Method	Epoch	Accuracy	Notes
LoRA	5	68.2%	Best performance; strikes a balance between parameter efficiency and accuracy.
QLoRA	5	67.9%	Close to LoRA in accuracy, with the advantage of lower memory requirements due to quantization.
DORA	3	67.58%	Achieved peak accuracy faster due to dynamic adaptation but did not maintain higher accuracy over more epochs.

INSIGHTS FOR LEGAL PREDICTION

LoRA: Suited for achieving higher accuracy in legal case prediction tasks where computational resources are not a constraint.

QLoRA: Ideal for deploying models in resource-limited environments while maintaining close accuracy to LoRA.

DORA: Good for scenarios requiring rapid adaptation but might need more computational power for sustained accuracy.

CONCLUSION

Choosing the Right Method: Depends on the task requirements, resource availability, and desired balance between accuracy and efficiency.

Future Prospects: Continued research in parameter-efficient fine-tuning could make large models more accessible.

Combining these techniques may yield even better results for specialized applications.