간단 논세

QLoRA: Efficient Finetuning of Quantized LLMs

Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." *Advances in Neural Information Processing Systems* 36 (2024).

2024.10.04 유하영

previous research

대규모 언어 모델(LLMs)을 파인튜닝(finetuning) 하는 것은 모델의 성능을 향상시키고 원하는 행동을 추가하거나 원치 않는 행동을 제거하는 데 매우 효과적
그러나 **매우 큰 모델을 파인튜닝하는 것은 비용이 매우 많이 든다.**

LoRA
:Low-Rank Adaptation

사전 학습된 가중치를 고정된 상태로 유지하면서
추가적인 Low-Rank matrix를 학습하여
신경망의 일부 레이어를 간접적으로 학습

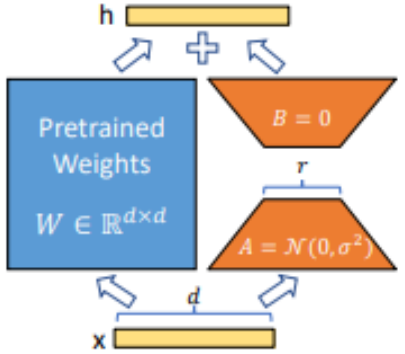


Figure 1: Our reparametrization. We only train A and B.

QLoRA
성능 저하 없이 4비트로 양자화된 모델을 파인튜닝
사전 학습된 모델을 4비트로 양자화한 후, 작은 세트의 학습 가능한 저차원 어댑터 가중치(Low-rank Adapter weights)를 추가하여 양자화된 가중치를 통해 **gradient 역전파**

QLoRA

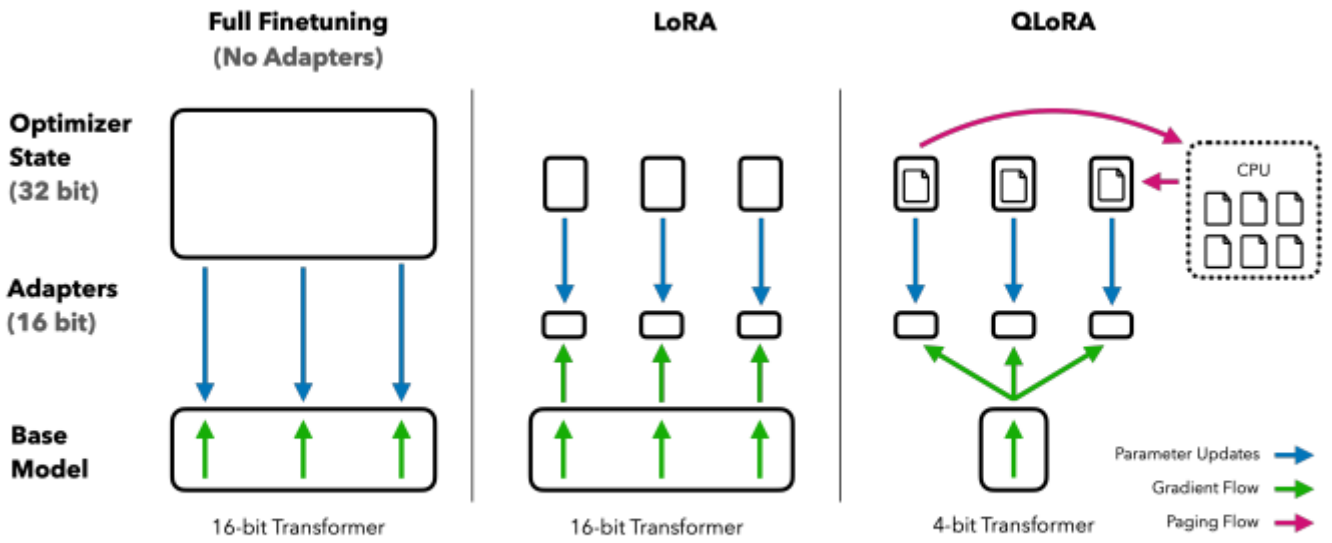
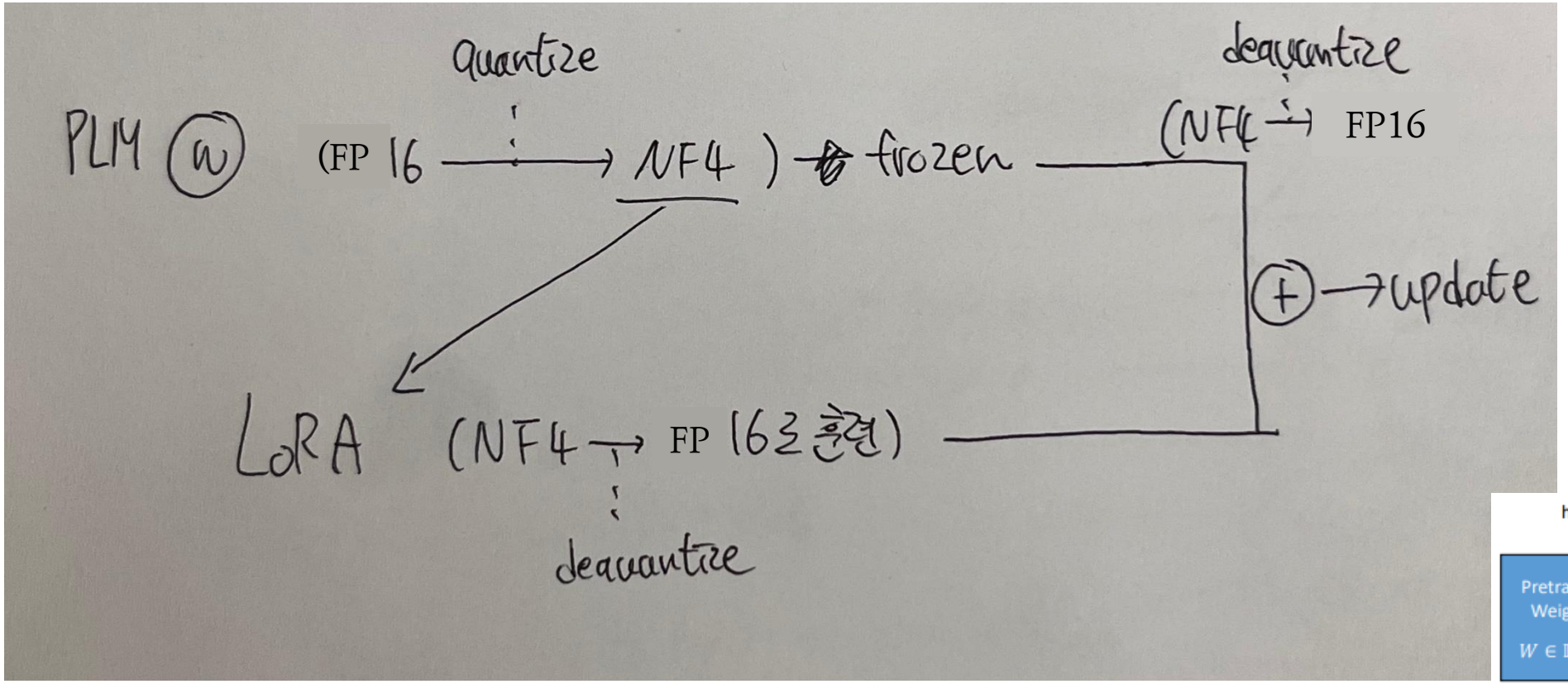


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

1. 4-bit NormalFloat (NF4)
2. Double Quantization
3. Paged Optimizers

GPU가 사용하는 VRAM 페이지를 CPU의 RAM에도 일부 저장할 수 있게 할당해주는 기술

4-bit NormalFloat (NF4)



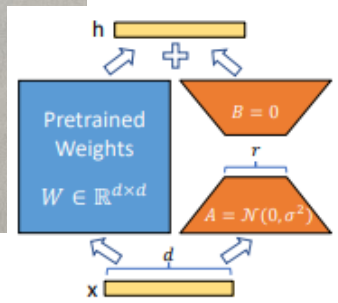


Figure 1: Our reparametrization. We only train A and B.

Quantile Quantization

Quantile Quantization(분위 양자화)
데이터의 분포를 작은 차원으로 바꾸기 위해 사용

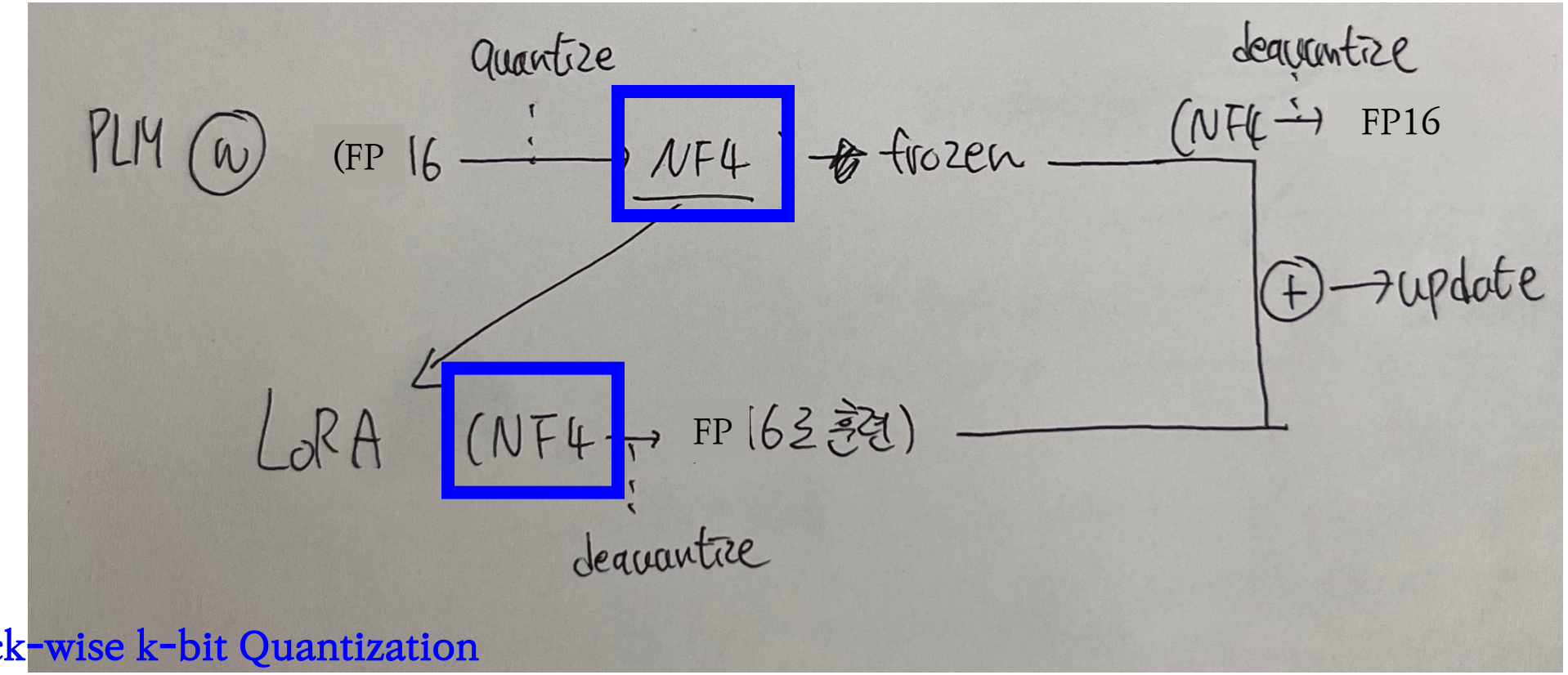
데이터를 순서대로 정렬한 후, 데이터를 특정 비율로 나누는 지점(사분위수)을 찾는다. 각 **양자화 구간에 할당되는 데이터의 개수가 동일하도록 보장하는 양자화 방식**

하지만, **비싼 quantile 추정비용**
-〉 이미 정해진 양자화 구간에 데이터 포인트들을 맵핑 **"단일 고정 분포"**

1.정규 분포에 대한 k-bit quantile quantization 데이터 유형을 얻기 위해 N(0,1) 분포의 $2k+1$개 구간개수를 추정한다.
2.이 데이터 유형을 가지고, 값을 [-1,1] 범위로 정규화한다.
3.Input 가중치 tensors를 absolute maximum rescaling을 통해 [-1,1] 범위로 정규화함으로써 양자화한다.

Double Quantization

:추가적인 메모리 절감을 위해 **가중치를 양자화할 때 추가적으로 발생하는 값인 '양자화 상수 (quantization constant)'를 양자화**



Block-wise k-bit Quantization

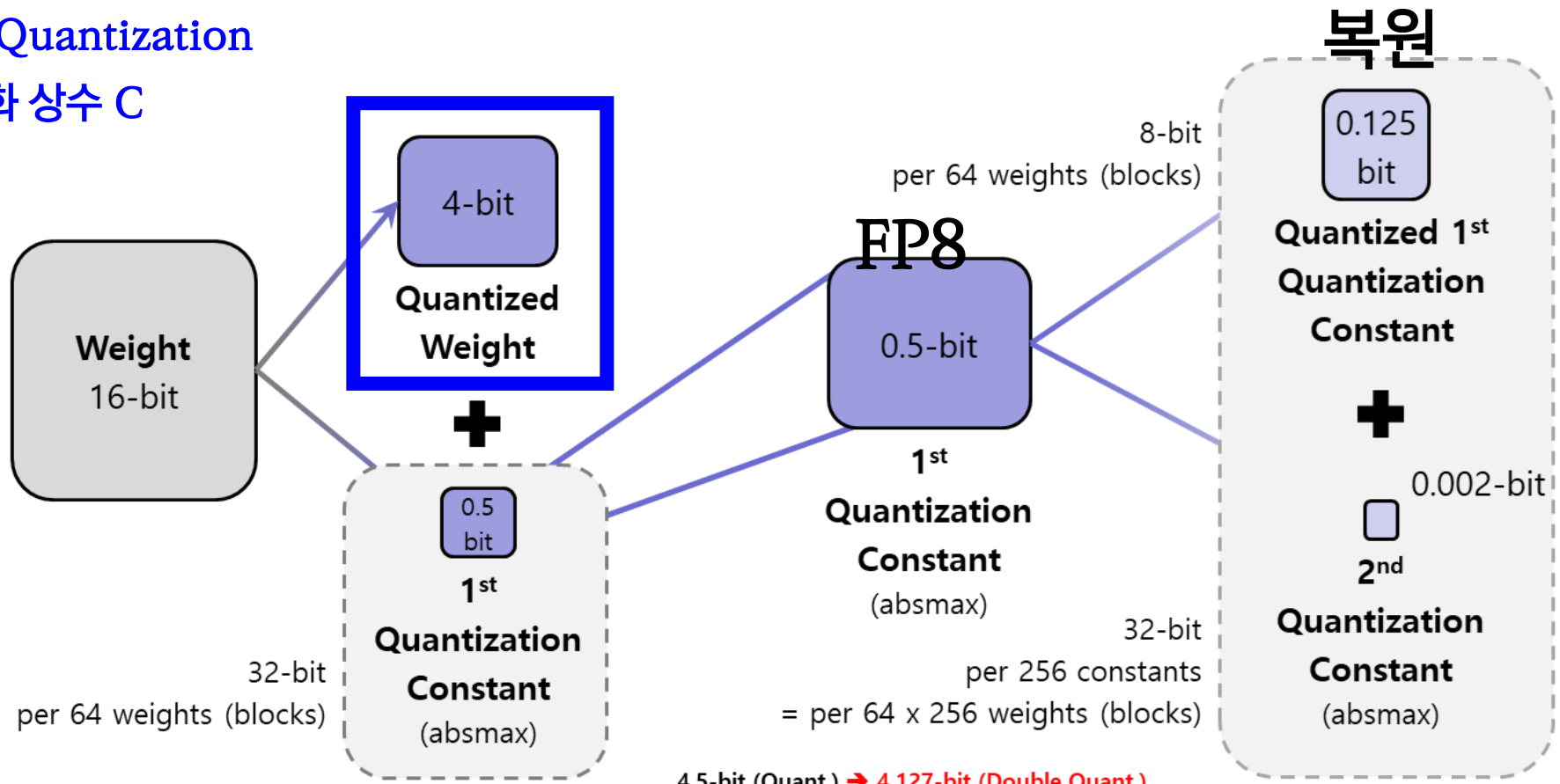
입력 텐서를 작은 블록으로 나누고, 자신의 양자화 상수 c를 갖는 독립적으로 양자화

Double Quantization

:추가적인 메모리 절감을 위해 **가중치를 양자화할 때 추가적으로 발생하는 값인 '양자화 상수 (quantization constant)'를 양자화**

Block-wise k-bit Quantization

양자화 상수 C



32-bit / (1 block*64weights) = 0.5

0.5-bit

0.127-bit

FP32

32-bit / (64-block*256weights) = 0. 002

QLoRA

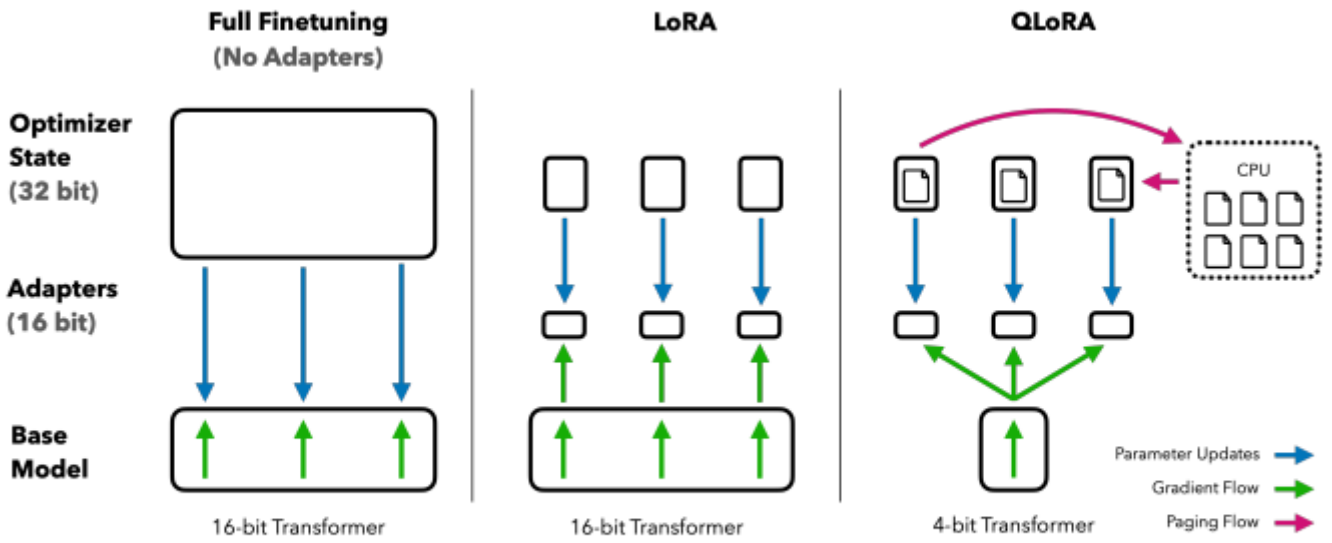


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

1. 4-bit NormalFloat (NF4)
2. Double Quantization
3. Paged Optimizers

GPU가 사용하는 VRAM 페이지를 CPU의 RAM에도 일부 저장할 수 있게 할당해주는 기술

Result

Model / Dataset	Params	Model bits	Memory	ChatGPT vs Sys	Sys vs ChatGPT	Mean	95% CI
GPT-4	-	-	-	119.4%	110.1%	**114.5%**	2.6%
Bard	-	-	-	93.2%	96.4%	94.8%	4.1%
Guanaco	65B	4-bit	41 GB	96.7%	101.9%	**99.3%**	4.4%
Alpaca	65B	4-bit	41 GB	63.0%	77.9%	70.7%	4.3%
FLAN v2	65B	4-bit	41 GB	37.0%	59.6%	48.4%	4.6%
Guanaco	33B	4-bit	21 GB	96.5%	99.2%	**97.8%**	4.4%
Open Assistant	33B	16-bit	66 GB	91.2%	98.7%	94.9%	4.5%
Alpaca	33B	4-bit	21 GB	67.2%	79.7%	73.6%	4.2%
FLAN v2	33B	4-bit	21 GB	26.3%	49.7%	38.0%	3.9%
Vicuna	13B	16-bit	26 GB	91.2%	98.7%	**94.9%**	4.5%
Guanaco	13B	4-bit	10 GB	87.3%	93.4%	90.4%	5.2%
Alpaca	13B	4-bit	10 GB	63.8%	76.7%	69.4%	4.2%
HH-RLHF	13B	4-bit	10 GB	55.5%	69.1%	62.5%	4.7%
Unnatural Instr.	13B	4-bit	10 GB	50.6%	69.8%	60.5%	4.2%
Chip2	13B	4-bit	10 GB	49.2%	69.3%	59.5%	4.7%
Longform	13B	4-bit	10 GB	44.9%	62.0%	53.6%	5.2%
Self-Instruct	13B	4-bit	10 GB	38.0%	60.5%	49.1%	4.6%
FLAN v2	13B	4-bit	10 GB	32.4%	61.2%	47.0%	3.6%
Guanaco	7B	4-bit	5 GB	84.1%	89.8%	**87.0%**	5.4%
Alpaca	7B	4-bit	5 GB	57.3%	71.2%	64.4%	5.0%
FLAN v2	7B	4-bit	5 GB	33.3%	56.1%	44.8%	4.0%

ChatGPT와의 유사성
:유사하게 답변을생성하는지