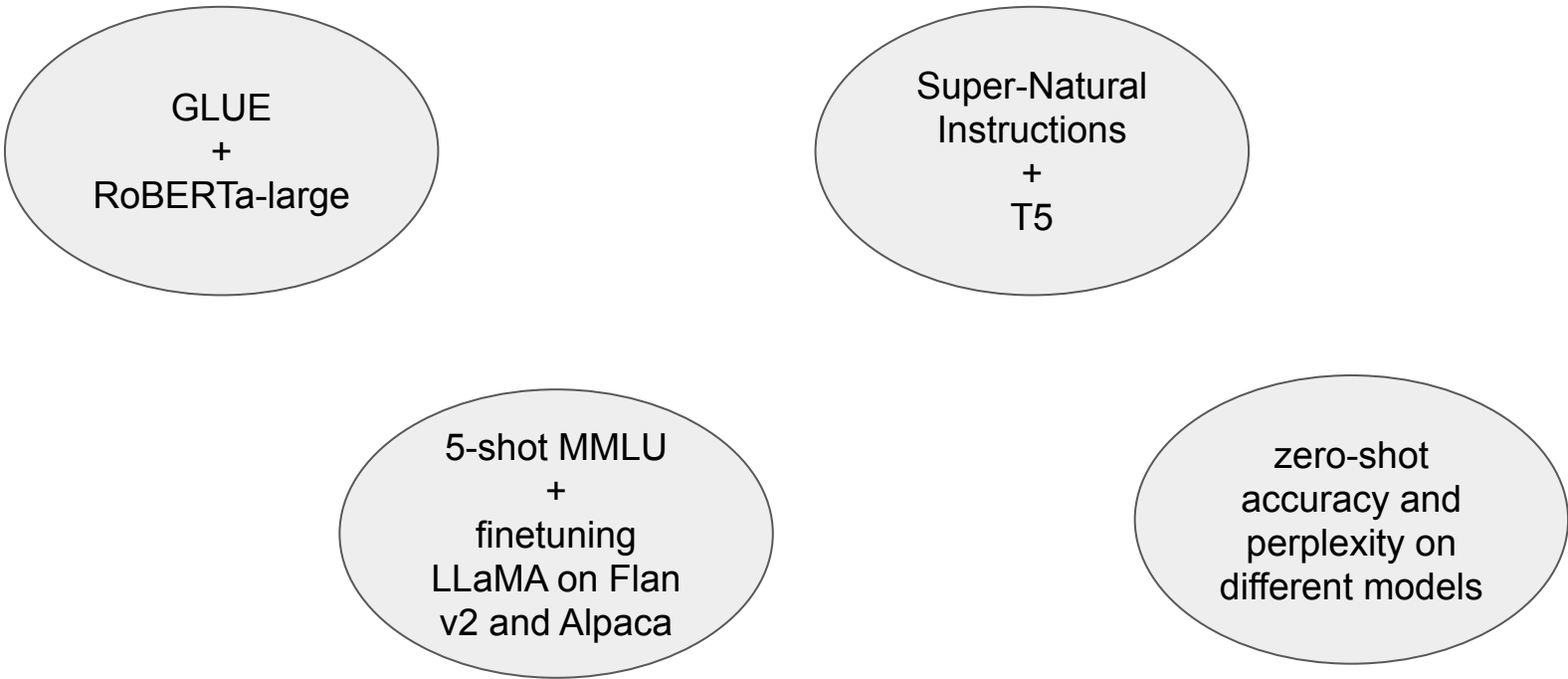


Experimental setup



GLUE
+
RoBERTa-large

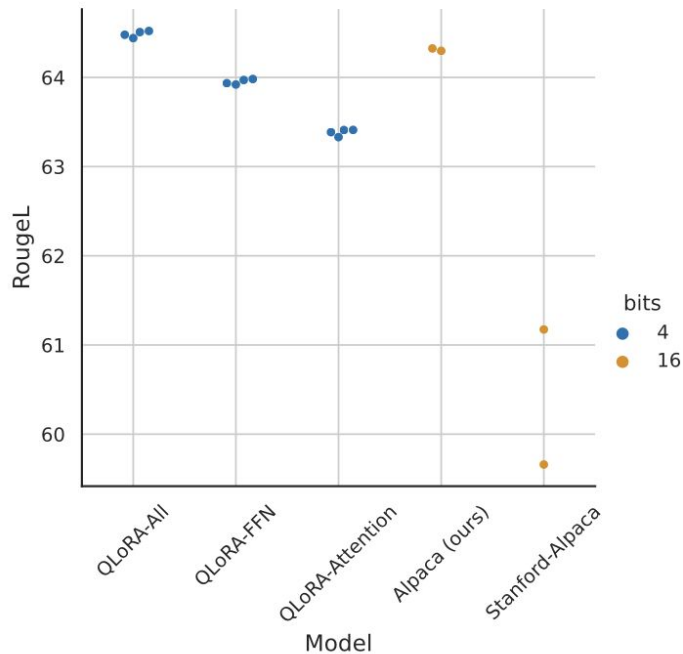
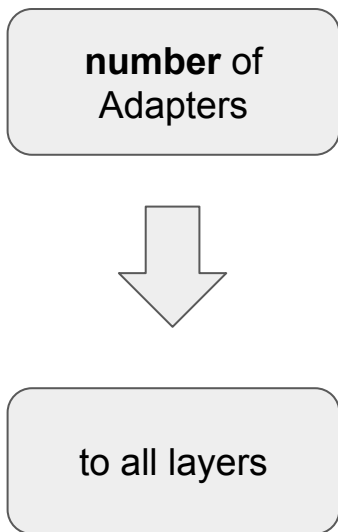
Super-Natural
Instructions
+
T5

5-shot MMLU
+
finetuning
LLaMA on Flan
v2 and Alpaca

zero-shot
accuracy and
perplexity on
different models

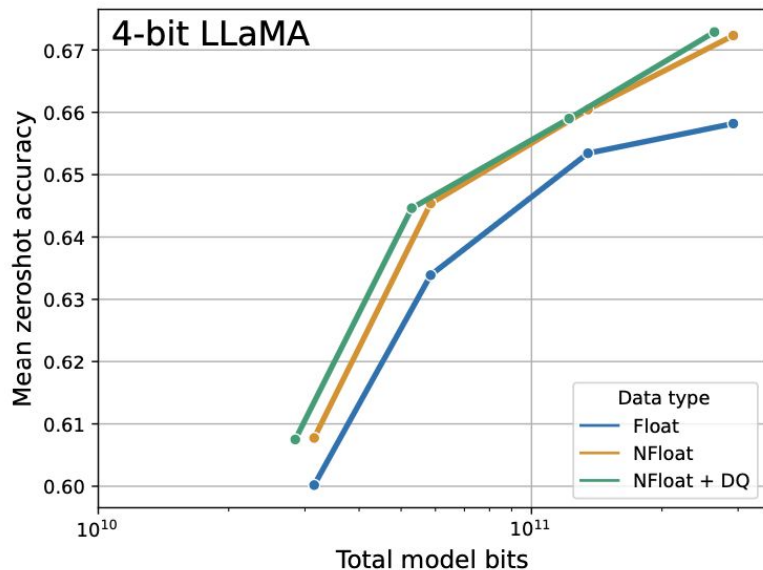
LoRA hyperparameters

Default LoRA hyperparameters do not match 16-bit performance



<Figure 2>

4-bit NF > 4-bit FP



<Figure 3>

Table 2: Pile Common Crawl mean perplexity for different data types for 125M to 13B OPT, BLOOM, LLaMA, and Pythia models.

Data type	Mean PPL
Int4	34.34
Float4 (E2M1)	31.07
Float4 (E3M0)	29.48
NFloat4 + DQ	27.41

k-bit QLoRA vs 16-bit full finetuning, 16-bit LoRA performance

Table 3: Experiments comparing 16-bit BrainFloat (BF16), 8-bit Integer (Int8), 4-bit Float (FP4), and 4-bit NormalFloat (NF4) on GLUE and Super-NaturalInstructions. QLoRA replicates 16-bit LoRA and full finetuning.

Dataset Model	GLUE (Acc.)	Super-NaturalInstructions (RougeL)				
	RoBERTa-large	T5-80M	T5-250M	T5-780M	T5-3B	T5-11B
BF16	88.6	40.1	42.1	48.0	54.3	62.0
BF16 replication	88.6	40.0	42.2	47.3	54.9	-
LoRA BF16	88.8	40.5	42.6	47.1	55.4	60.7
QLoRA Int8	88.8	40.4	42.9	45.4	56.5	60.7
QLoRA FP4	88.6	40.3	42.4	47.5	55.6	60.9
QLoRA NF4 + DQ	-	40.4	42.7	47.7	55.3	60.9

: the lost performance caused by quantization can be fully recovered through adapter finetuning

Table 4: Mean 5-shot MMLU test accuracy for LLaMA 7-65B models finetuned with adapters on Alpaca and FLAN v2 for different data types. Overall, NF4 with double quantization (DQ) matches BFloat16 performance, while FP4 is consistently one percentage point behind both.

LLaMA Size Dataset	Mean 5-shot MMLU Accuracy								Mean
	7B		13B		33B		65B		
	Alpaca	FLAN v2	Alpaca	FLAN v2	Alpaca	FLAN v2	Alpaca	FLAN v2	
BFloat16	38.4	45.6	47.2	50.6	57.7	60.5	61.8	62.5	53.0
Float4	37.2	44.0	47.3	50.0	55.9	58.5	61.3	63.3	52.2
NFloat4 + DQ	39.0	44.5	47.5	50.7	57.3	59.2	61.8	63.9	53.1

- 1) QLoRA with NF4 replicates both 16-bit full finetuning and 16-bit LoRA finetuning performance.
- 2) NF4 is superior to FP4 in terms of quantization precision

결과 정리

- LoRA 하이퍼파라미터에서 중요한 것은 **Adapter** 사용 개수와 그것이 모든 레이어에 적용되어야 한다는 것
- Floating point 보다는 Normal Float이 더 뛰어난 성능을 냄
- 4-bit QLoRA with NF4 data type matches 16-bit full finetuning and 16-bit LoRA finetuning performance.
- QLoRA 짱!

감사합니
다!