Week 10 : Assignment 10

The due date for submitting this assignment has passed.

Due on 2025-04-02, 23:59 IST.

Assignment submitted on 2025-03-31, 16:20 IST

1) How do Prefix Tuning and Adapters differ in terms of where they inject new task-specific parameters in the Transformer architecture? ***1 point***

- Prefix Tuning adds new feed-forward networks after every attention block, while Adapters prepend tokens.
- Both approaches modify only the final output layer but in different ways.
- Prefix Tuning learns trainable “prefix” hidden states at each layer’s input, whereas Adapters insert small bottleneck modules inside the Transformer blocks.
- Both approaches rely entirely on attention masks to inject new task-specific knowledge.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Prefix Tuning learns trainable “prefix” hidden states at each layer’s input, whereas Adapters insert small bottleneck modules inside the Transformer blocks.

2) The Structure-Aware Intrinsic Dimension (SAID) improves over earlier low-rank adaptation approaches by: ***1 point***

- Ignoring the network structure entirely
- Learning one scalar per layer for layer-wise scaling
- Sharing the same random matrix across all layers
- Using adapters within self-attention layers

No, the answer is incorrect.
Score: 0
Accepted Answers:
Learning one scalar per layer for layer-wise scaling

3) Which of the following are correct about the extensions of LoRA? ***1 point***

- [x] LongLoRA supports inference on longer sequences using global attention

- [x] QLoRA supports low-rank adaptation on 4-bit quantized models
- [x] DyLoRA automatically selects the optimal rank during training
- [x] LoRA+ introduces gradient clipping to stabilize training

No, the answer is incorrect.
Score: 0
Accepted Answers:
QLoRA supports low-rank adaptation on 4-bit quantized models
DyLoRA automatically selects the optimal rank during training

4) Which pruning technique specifically removes weights with the smallest absolute values first, potentially followed by retraining to recover accuracy? ***1 point***

- (•) Magnitude Pruning
- ( ) Structured Pruning
- ( ) Random Pruning
- ( ) Knowledge Distillation

Yes, the answer is correct.
Score: 1
Accepted Answers:
Magnitude Pruning

5) In Post-Training Quantization (PTQ) for LLMs, why is a calibration dataset used? ***1 point***

- ( ) To precompute the entire attention matrix for all tokens.
- ( ) To remove outlier dimensions before applying magnitude-based pruning.
- ( ) To fine-tune the entire model on a small dataset and store the new weights.
- (•) To estimate scale factors for quantizing weights and activations under representative data conditions.

Yes, the answer is correct.
Score: 1
Accepted Answers:
To estimate scale factors for quantizing weights and activations under representative data conditions.

6) Which best summarizes the function of the unembedding matrix W_U? ***1 point***

- ( ) It merges the queries and keys for each token before final classification.
- (•) It converts the final residual vector into vocabulary logits for next-token prediction.
- ( ) It is used for normalizing the QK and OV circuits so that their norms match.
- ( ) It acts as a second attention layer that aggregates multiple heads.

Yes, the answer is correct.
Score: 1
Accepted Answers:
It converts the final residual vector into vocabulary logits for next-token prediction.

7) Which definition best matches an induction head as discovered in certain Transformer circuits? ***1 point***

- ( ) A head that specifically attends to punctuation tokens to determine sentence boundaries
- ( ) A feed-forward sub-layer specialized for outputting next-token probabilities for out-of-distribution tokens
- (•) A head that looks for previous occurrences of a token A, retrieves the token B that followed it last time, and then predicts B again
- ( ) A masking head that prevents the model from looking ahead at future tokens

Yes, the answer is correct.

Score: 1

Accepted Answers:

A head that looks for previous occurrences of a token A, retrieves the token B that followed it last time, and then predicts B again

8) In mechanistic interpretability, how can we define ‘circuit’? ***1 point***

- ○ A data pipeline for collecting training examples in an autoregressive model
- ○ A small LSTM module inserted into a Transformer for additional memory
- ○ A device external to the neural network used to fine-tune certain parameters after training
- ◉ A subgraph of the neural network hypothesized to implement a specific function or behaviour

Yes, the answer is correct.
Score: 1

Accepted Answers:

A subgraph of the neural network hypothesized to implement a specific function or behaviour

9) Which best describes the role of Double Quantization in QLoRA? ***1 point***

- ○ It quantizes the attention weights twice to achieve 1-bit representations.
- ○ It reinitializes parts of the model with random bit patterns for improved regularization.
- ◉ It quantizes the quantization constants themselves for additional memory savings.
- ○ It systematically reverts partial quantized weights back to FP16 whenever performance degrades.

Yes, the answer is correct.
Score: 1

Accepted Answers:

It quantizes the quantization constants themselves for additional memory savings.

10) Which of the following are true about sequence-level distillation for LLMs? ***1 point***

- ☑ It trains a student model by matching the teacher’s sequence outputs (e.g., predicted token sequences) rather than just individual token distributions.
- ☐ It requires storing only the top-1 predictions from the teacher model for each token.
- ☑ It can be combined with word-level distillation to transfer both local and global knowledge.
- ☐ It forces the teacher to produce a chain-of-thought explanation for each example.

Yes, the answer is correct.
Score: 1

Accepted Answers:

It trains a student model by matching the teacher’s sequence outputs (e.g., predicted token sequences) rather than just individual token distributions.
It can be combined with word-level distillation to transfer both local and global knowledge.