Introduction to Large Language Models
Assignment- 7

Number of questions: 8 **Total mark: 6 X 1 + 2 X 2 = 10**

QUESTION 1: [1 mark]

Which of the following best describes how ELMo's architecture captures different linguistic properties?

a) The model explicitly assigns specific linguistic functions to each layer.
b) The lower layers capture syntactic information, while higher layers capture semantic information.
c) All layers capture the similar properties.
d) ELMo uses a fixed, non-trainable weighting scheme for combining layer-wise representations.

Correct Answer: b

Solution: ELMo uses a multi-layer bidirectional LSTM architecture, where different layers capture different aspects of language. Empirical evidence shows that lower layers focus more on syntactic information while higher layers capture more semantic nuances.

QUESTION 2: [1 mark]

BERT and BART models differ in their architectures. While BERT is ___(i)___ model, BART is___(ii)___ one. Select the correct choices for (i) and (ii).

a) i: Decoder-only ; ii: Encoder-only
b) i: Encoder-decoder ; ii: Encoder-only
c) i: Encoder-only ; ii: Encoder-decoder
d) i: Decoder-only ; ii: Encoder-decoder

Correct Answer: c

Solution: BERT is an encoder-only transformer model, while BART is an encoder-decoder model.

QUESTION 3: [1 mark]

The pre-training objective for the T5 model is based on:

a) Next sentence prediction
b) Masked language modelling
c) Span corruption and reconstruction

d) Predicting the next token

Correct Answer: c

Solution: T5 is trained using a span corruption objective, which requires the model to reconstruct masked spans of text.

QUESTION 4: [1 mark]

Which of the following datasets was used to pretrain the T5 model?

a) Wikipedia
b) BookCorpus
c) Common Crawl
d) C4

Correct Answer: d

Solution: T5 was pretrained on the “C4” (Colossal Clean Crawled Corpus) dataset.

QUESTION 5: [1 mark]

Which of the following special tokens are introduced in BERT to handle sentence pairs?

a) [MASK] and [CLS]
b) [SEP] and [CLS]
c) [CLS] and [NEXT]
d) [SEP] and [MASK]

Correct Answer: b

Solution: BERT introduces the [CLS] token at the start for classification or overall sequence representation and the [SEP] token to separate sentences. Thus, the special tokens are “[SEP]” and “[CLS]”.

QUESTION 6: [2 marks]

ELMo and BERT represent two different pre-training strategies for language models. Which of the following statement(s) about these approaches is/are true?

a) ELMo uses a bi-directional LSTM to pre-train word representations, while BERT uses a transformer encoder with masked language modeling.
b) ELMo provides context-independent word representations, whereas BERT provides context-dependent representations.
c) Pre-training of both ELMo and BERT involve next token prediction.

d) Both ELMo and BERT produce word embeddings that can be fine-tuned for downstream tasks.

Correct Answer: a, d

Solution: ELMo uses bidirectional LSTMs with a language modeling objective, while BERT uses a transformer encoder and masked language modelling. Both can produce embeddings that are fine-tuned for downstream tasks. Hence, the correct answers are (a) and (d).

QUESTION 7: [1 mark]

Decoder-only models are essentially trained based on probabilistic language modelling. Which of the following correctly represents the training objective of GPT-style models?

a) P(y | x) where x is the input sequence and y is the gold output sequence
b) P(x | y) where x is the input sequence and y is the gold output sequence
c) $P(w_t \mid w_{1:t-1})$, where w_t represents the token at position t, and $w_{1:t-1}$ is the sequence of tokens from position 1 to t-1
d) $P(w_t \mid w_{1:t+1})$, where w_t represents the token at position t, and $w_{1:t+1}$ is the sequence of tokens from position 1 to t+1

Correct Answer: c

Solution: Decoder-only (GPT-style) models are trained using left-to-right language modeling, predicting each token given all previous tokens. Thus, the objective is $P(w_t \mid w_{1:t-1})$.

QUESTION 8: (Numerical Question) [2 marks]

In the previous week, we saw the usage of **einsum** function in numpy as a generalized operation for performing tensor multiplications. Now, consider two matrices: $A = \begin{bmatrix} 1 & 5 \\ 3 & 7 \end{bmatrix}$ and $A = \begin{bmatrix} 2 & -1 \\ 4 & 2 \end{bmatrix}$. Then, what is the output of the following numpy operation?

```
numpy.einsum('ij,ij->', A, B)
```

Correct Answer: 23

Solution: The operation `numpy.einsum('ij,ij->', A, B)` computes the elementwise product of A and B, then sums all those products.

Thus, output = 2*1 + (-1)*5 + 4*3 + 2*7 = 2 – 5 + 12 + 14 = 23