Question 1 Consider the equations for a standard LSTM cell:

$$
\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \\
f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \\
o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \\
\tilde{c}_t &= \phi(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \phi(c_t)
\end{aligned}
$$

In the equations above, which term explicitly represents the memory component that enables the LSTM to retain **long-term information** across timesteps?

- ☐ Output gate o_t
- ☐ Hidden state h_t
- ☐ Cell state c_t
- ☐ Candidate cell state $\tilde{c}_t$
- ☐ Input gate i_t

Solution: The cell state c_t retains long-term information while the hidden state h_t acts as a short-term memory.

Question 2 BERT introduces a special token, [CLS], at the beginning of every input sequence. Which of the following statements best describes the purpose of the [CLS] token?

- ☐ It serves as a placeholder whose final hidden representation acts as a holistic sequence-level embedding, typically used for classification or next-sentence prediction tasks.
- ☐ It serves primarily to separate multiple sentences within the same input (the same role as [SEP] does).
- ☐ It simply marks sentence boundaries and carries no trainable embeddings of its own.
- ☐ It marks the exact midpoint of the input sequence to ensure balanced bidirectional attention.
- ☐ It is used only during masked language modeling and is dropped for downstream tasks.

Solution: The [CLS] special token is introduced to aggregate information about the entire sequence in its embedding and is used as input to a classification model.

Question 3 From the following set of models: {`ELMo`, `BERT`, `GPT`, `BART`, `T5`}, which group can each be directly used for both classification and generation tasks (without any modifications)?

- ☐ ELMo, BERT
- ☐ BERT, GPT
- ☐ BART, T5
- ☐ BERT, GPT, T5
- ☐ ELMo, BART, GPT

Solution: BART and T5 are encoder-decoder models capable of both classification and text generation. GPT also supports both tasks; however, in this question, it is always paired with bidirectional models like BERT and ELMo, which are not suitable for generation.