



JK Lakshmipat University, Jaipur
Department of Computer Science and Engineering
Large Language Models (CS1230)
End-Term, Date: December 01, 2025

Time: 2:30 hours

B.Tech. Semester VII CSE, ECE

Max mark: 42

Please read the following instructions carefully.

- There are 5 questions printed on THREE pages.
- **No marks** for providing just expressions/answers unless accompanied with correct justification and/or derivation.
- In case of any doubt, make your assumption, write it clearly and continue. DO NOT call your instructor.

1. (2+2+2 points) Consider a decoder-only large language model (LLM) with the following hyperparameters:

- Number of decoder layers: l
- Hidden dimension: d
- Number of attention heads: h
- Sequence length (number of previously generated tokens to cache): T

The model can use three different attention mechanisms: Multi-Head Attention (MHA), Multi-Query Attention (MQA) and Grouped-Query Attention (GQA). Also assume that all other hyperparameters remain identical, and each head has equal dimension $\frac{d}{h}$.

Compute the total KV cache size (number of stored key and value vectors) for a sequence of length T for each attention mechanism: (CO1, CO3)

2. (3*1+2+3 points) Mention the training tasks for BERT, T5 and GPT-3. Compare the difficulty of the tasks mentioned with reasons. Show attention diagrams for mechanisms in the three models. (CO1, CO2)

3. (2+2+(1+1) points) AI regulations. (CO5)

- Draw and label the risk pyramid of the EU AI Act.
- What is your opinion on keeping 10^{25} floating point operations (FLOPs) as a threshold for the classification of general purpose AI models with systemic risk?
- Briefly state
 - one application of AI that is classified under Article 5; and
 - one application that is classified under Article 6 (enumerated in Annex III) of the EU AI Act.

4. (2+(1+1)+2+2+2 points) In the alignment of LLMs with human values, RLHF is used post training. Use of policy gradient methods is common in the LLMs today. Some of the objective functions used in post training phase are provided below. The meaning of the variables are same as in the class. (CO3, CO4)

$$\underset{\theta}{\text{maximize}} \mathbb{E}_{s_t \sim d^{\pi_\theta}, a_t \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(a_t | s_t) A^\pi(s_t, a_t)] \quad (1)$$

$$\underset{\theta}{\text{maximize}} \mathbb{E}_t \left[\frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} A_t \right] \quad (2)$$

$$\text{subject to } \mathbb{E}_t [D_{\text{KL}}(\pi_{\theta_{\text{old}}}(\cdot | s_t) \| \pi_\theta(\cdot | s_t))] \leq \delta \quad (3)$$

$$\underset{\theta}{\text{maximize}} \mathbb{E}_t [\min(r_t(\theta) A_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) A_t)] \quad (4)$$

$$\underset{\theta}{\text{maximize}} \mathbb{E}_{q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | q)} \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left[\min \left(\frac{\pi_\theta(o_{i,t} | q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} | q, o_{i,<t})} \hat{A}_{i,t}, \right. \right. \\ \left. \left. \text{clip} \left(\frac{\pi_\theta(o_{i,t} | q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} | q, o_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right) - \beta \hat{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right] \quad (5)$$

$$\underset{\theta}{\text{maximize}} \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \log \sigma \left(\beta \left[\log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right] \right) \quad (6)$$

- Why is working with the *log* of policy gradients justified instead of the policy gradients?
- Equation number 2 and 3 correspond to TRPO. Can we change the order of the parameters in the KL divergence term? Reason. Why is TRPO not used in practice?
- Breidley Terry model is used in defining of DPO objective. How is an LLM trained using DPO?
- How does the PPO objective simplify TRPO?
- How does PPO objective improved in GRPO?

5. ((3+2)+3+3+1 points) As we know DeepSeek has improved the architecture of LLMs. Refer to the figure provided below, Attention and FFNN layers in the transformer blocks have been modified as provided. (CO3, CO4)

- Write the equations corresponding to Multihead Latent Attention with the dimensions of the parameter matrices used in the process. Also provide the size of the KV cache in this case.
- If FFNN in normal transformer block looks like $512 \times 2048 \times 512$, you want to keep experts of size $32 \times 128 \times 32$ how many experts do you need and how will you use them to achieve the parameter count for this layer to be the **same** at the inference time?
- If you want to reduce the number of parameters used at the inference by 32 times. How many experts should be used? Take into account the values of N_s and K_r (as provided in the figure below and also understood from the use of these variables in the class).
- Does d_{model} have to be equal to $n_h d_h$ in all the cases. Reason for your answer.



