# ELL881/AIL821: LLMs - Introduction and Recent Advances

**End-Semester Examination** (Semester I, 2024-25), **IIT Delhi**

**Total Marks: 50** Time: 2 hours

---

**Question 0: For each statement below, identify whether it is True or False. Give proper justification for your answer. No marks for no justification.** ($2 \times 5 = 10$ marks)

(a) Attention with Linear Biases (ALiBi) adds positional embeddings to token embeddings.

(b) In WordPiece tokenization, the algorithm stores and applies merge rules in sequence during inference, similar to BPE's approach.

(c) Rotary Positional Encoding (RoPE) represents positions as complex numbers and tokens as pure rotations.

(d) In temperature sampling, low temperature flattens the output probability distribution of the tokens.

(e) Tool Augmented Language Models (TALMs) require extensive pre-training on tool-specific datasets to effectively utilize external tools like calculators and search engines.

**Question 1: Parameter-Efficient Fine-Tuning & Model Compression** (10 marks)

(a) What are the optimization strategies incorporated by QLoRA on top of LoRA? Explain each strategy in brief. **(4 marks)**

(b) Consider that we want to distill knowledge from a teacher model, parameterized by $\theta_T$, to a student model, parameterized by $\theta_S$. Also, assume that the probability distribution generated by the teacher model as output is $p_T(.|x;\theta_T)$ and that from the student model is $p_S(.|x;\theta_S)$, for an input $x$.

(i) Write the loss function for performing knowledge distillation from teacher to student model at the **word-level**, for a sentence **s**. Assume the length of the sentence to be $L$, and $\mathcal{V}$ to be the target vocabulary set. **(2 marks)**

(ii) What changes need to be incorporated into the loss function when we want to do distillation at a sequence level? **(2 marks)**

(c) How does prompt tuning differ from prefix tuning in its implementation? **(2 marks)**

**Question 2: Alignment of Language Models** (10 marks)

Consider a class of loss functions $f$ that are linear in $v_f(r_\theta(x,y) - \mathbb{E}_Q[r_\theta(x',y')])$, where: $r_\theta(x,y) = l(x,y)\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$ is the implied reward, $Q(X',Y'|x,y)$ is an input-conditioned reference point distribution and $v_f : \mathbb{R} \to \mathbb{R}$ is a value function that is non-decreasing everywhere and concave in $(0,\infty)$

(a) Prove that the **PPO-Clip loss** and the **Direct Preference Optimization (DPO)** loss functions belong to this class. For each loss function, provide the specific constructions of $l(y)$, $r_\theta$, $v_f(\cdot)$, and $Q$ that satisfy the given criteria. Additionally, explain how these constructions relate to human decision-making processes in the context of language model alignment. **(5 marks)**

(b) Assuming the value function is logistic, for a reward function $r_a^*$ that maximizes RLHF loss function, prove the following:

$\exists r_b^* : r_b^*(x,y) = r_a^*(x,y) + h(x), \exists h(x) : \pi^*_{r_b^*} = \pi^*_{r_a^*}$ **(3 marks)**

(c) Consider a scenario where you are given a large language model, such as Llama-3.1-base, and you perform direct preference tuning (i.e. RLHF or DPO) on it without any prior supervised fine-tuning. **(2 marks)**

(i) Analyze the potential consequences of this approach on the model's performance, capabilities, and behavior.

(ii) Explicitly mention which term of the loss function will create an obstacle to do so.

**Question 3: Knowledge Editing in Language Models** (6 marks)

(a) Consider a language model $f(x;\theta)$ with parameters $\theta$. Explain how the KnowledgeEditor method represents and computes parameter updates $\Delta\theta$ to edit a specific fact. What information does it use as input, and how does it leverage gradient information? **(3 marks)**

(b) Describe the constrained optimization approach for training the KnowledgeEditor. **(3 marks)**

(i) Write out the mathematical formulation of the optimization objective.

(ii) Explain the purpose of the constraint.

(iii) Why is KL divergence used for the constraint instead of a parameter space norm like $L_2$?

**Question 4: Multimodal Models** (6 marks)

(a) The standard Transformer receives as input a 1D sequence of token embeddings. However, when dealing with images, a challenge is posed – each image is a matrix of dimension $H \times W \times C$, where $H$, $W$ and $C$ are respectively the height, width and number of channels of an image.

How does a Vision Transformer (ViT) handle images as input? Specifically state the input format (with dimensions) which goes into the ViT, and whether any architectural changes need to be made to ViTs compared to standard Transformer Encoders. **(3 marks)**

(b) As discusssed in class, *VideoCLIP* is an approach to pre-train a unified model for zero-shot video and text understanding.

Assuming $z_v$ and $z_t$ to be the hidden states corresponding to video and text clips respectively, write the loss function used for pre-training using the *VideoCLIP* approach. **(3 marks)**

**Question 5: Interpreting the Inner Workings of LLMs** (8 marks)

(a) What are *control tasks* in the context of probing classifers? Why are they needed? **(1.5 marks)**

(b) Explain what are the $QK$ and $OV$ circuits? Briefly describe their role in the interpretability of the working of LLMs. **(2 marks)**

(c) Consider the *Patchscopes* framework, as discussed in class.

Assume, given an input sequence of $n$ tokens $S = \langle s_1, ..., s_n \rangle$ and a model $\mathcal{M}$ with $L$ layers, $h_i^l$ denotes the hidden representation obtained at layer $l$ and position $i$, when running $\mathcal{M}$ on $S$. To inspect $h_i^l$, a separate inference pass of a model $\mathcal{M}^*$ with $L^*$ layers is considered on a target sequence $T = \langle t_1, ..., t_m \rangle$ of $m$ tokens. A mapping function $f(h;\theta) : \mathbf{R}^d \to \mathbf{R}^{d^*}$ parameterized by $\theta$ is defined that operates on hidden representations of $\mathcal{M}$, where $d$ and $d^*$ denote the hidden dimension of representations in $\mathcal{M}$ and $\mathcal{M}^*$, respectively. $f(h_i^l)$ is patched in place of $\overline{h}_j^{l^*}$ while doing a forward pass in $\mathcal{M}^*$.

(i) Write the conditions on $\mathcal{M}$, $\mathcal{M}^*$, $l$, $l^*$ and $f$, for which *Patchscopes* becomes equivalent to *Logit Lens*. **(2 marks)**

(ii) How can we use the *Patchscopes* framework to check if certain attributes (like, entity information) are encoded in the hidden representation of a layer in a model? How is using *Patchscopes* for this purpose beneficial over the approach using probing classifiers? **(2.5 marks)**