



JK Lakshmipat University, Jaipur
Department of Computer Science and Engineering
Large Language Models (CS1230)
End-Term, Date: April 22, 2024

Time: 2:30 hours

B.Tech. Semester VI

Max mark: 35

Please read the following instructions carefully.

- There are 8 questions.
- **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. Discuss self-attention mechanism in three kinds of transformer architectures in detail. Draw diagrams extensively. (3+1)
2. We have seen an exponential increase in the number of parameters in GPT models from OpenAI. How such an exponential increase is justified? How is such an exponential increase achieved? (2+2)
3. The three steps for RLHF are as provided below. Find out the defect(s) in the following, correct it and justify your correction. Briefly explain each step. (2+3)
 - (a) Language model pretraining
 - (b) Fine tuning with RL using KL penalty
 - (c) Reward model training
4. What are the tasks that BERT, T5 and GPT are trained on? Compare the difficulty of the tasks mentioned. (2+2+1+2)
5. Describe transformer architecture as developed in the research paper *Attention is all you need*. (3)
6. Why RNNs, including LSTMs, couldn't achieve what transformer based architectures could? Basically compare the two architectures for sequence to sequence problems. (2+2)
7. Differentiate how is tokenization in the modern LLMs different from `fasttext`. (4)
8. Write down the cost function for word2vec skipgram model without negative sampling. Draw the diagram of the corresponding neural network that may be used to learn word embeddings. (2+2)