Question 1: What are the three different levels of using large language models? Write a short paragraph to detail each of the steps.

Answer: The three levels of using LLMs are as below: -

1. **Level 1 - Prompt Engineering**: It involves prompting or instructing the model out-of-the-box, what to do in natural language without tweaking any of its parameters. The focus is on identifying the best ways to craft a prompt such that we get relevant and reliable response from the model, for the task on hand. General rule is to be as specific and descriptive as possible. Some proven approaches are In-Context/Few Shot prompting, Chain-Of-Thought prompting etc.

2. **Level 2 – Task specific supervised fine tuning**: Generic pre-trained LLMs are generally not equipped to perform specialized/domain specific tasks. They need to be trained further on smaller but adequate supervised data to follow task specific instructions. This is known as task specific supervised fine tuning. It involves updating a few of the model parameters. Where there is no access to sizeable data/compute for performing finetuning, PEFT techniques like LoRA can be used where the original LLM parameters are frozen. Approaches like RLHF + PPO or DPO are used to make the model responses more aligned to human preference.

3. **Level 3 – Building LLM from Scratch**: This involves building and training an LLM from scratch. It is relevant for situations where we cannot use open source LLMs with proprietary data. It involves updating all model parameters, needs huge compute resources and of course is very costly.

Question 2: What are some common limitations of large language models, including ChatGPT?

Answer: Some common limitations of LLMs, including ChatGPT are as below:-

- Lack of factuality/reliability of results: LLM outputs can often be fluid but wrong/toxic/undesirable. The outputs could also be biased at times (when the training data itself is biased)
- Lack of robustness: LLMs can sometimes suffer from poor domain generalization. Responses can be poor/less effective when tried on new domains that the LLM was not exposed to during pre-training.
- Contextual limits: LLMs struggle with maintaining context across long, complex interactions.
- Lack of understanding: LLMs just recognize patterns but do not have consciousness and comprehension of the meaning behind the text.
- Lack of Real-Time-Knowledge: Their knowledge is limited to their last training data update and cannot access real-time information on their own.

- LLMs could cause social disruption like job loss/re-alignment, increase in targeted phishing/fraud/manipulation, increased use of algorithmic decisions creating disparities etc.

Question 3: How can Retrieval Augmented Generation (RAG) boost the performance of LLMs? In other words, what are the advantages of using RAG?

Answer: Following are the key aspects on how RAG can boost the performance of LLMs:

- Adaptability and Interpretability: Helps LLMs handle evolving facts and information with increased interpretability due to the augmented context.
- Efficiency: Allows augmentation of external information with LLMs' pretrained knowledge, eliminating the need to retrain LLM for new knowledge.
- Reliability: Improves the factual consistency and accuracy of the generated content, ie. reducing hallucinations.