Assignment 13 - [ChatGPT LLM]

Q1)
ANS:-

There are three main levels of using large language models (LLMs):

1. Using Pre-trained Models: At this level, users utilize LLMs as they are, without any customization. These models are pre-trained on vast amounts of data and can be used directly via APIs to perform tasks like answering questions, summarizing text, or generating content.

2. Fine-tuning Models: In this step, users fine-tune pre-trained LLMs on specific, domain-relevant datasets to improve their performance for specialized tasks. This customization allows the model to learn from a focused dataset, making it more accurate for certain applications like customer support or medical diagnostics.

3. Custom Training from Scratch: This advanced level involves training an LLM from scratch using specific data and architecture choices. It requires large datasets, high computing power, and significant expertise. This approach is used when highly specialized or proprietary models are needed for unique tasks that aren't well-served by existing models.

Q2)
ANS:-
Large language models (LLMs) like ChatGPT have several limitations:

1. Lack of Real-Time Knowledge: LLMs are typically trained on data up until a certain point (in the case of ChatGPT, until 2021 or 2023 depending on the version). They cannot access or update real-time information unless integrated with external databases or browsing capabilities.

2. Inconsistent Accuracy: While LLMs can generate detailed, coherent responses, they may sometimes produce factually incorrect or misleading information due to their reliance on probabilistic language generation rather than true understanding.

3. Limited Understanding of Context: Although LLMs handle context well in short exchanges, they can lose track of complex or prolonged contexts, leading to irrelevant or contradictory responses in long conversations.

4. Lack of Common Sense and Reasoning: LLMs are not truly intelligent; they can struggle with tasks requiring deep reasoning, real-world common sense, or understanding of nuanced situations, which can lead to illogical conclusions or responses.

Q3)
ANS:-
Retrieval Augmented Generation (RAG) can significantly boost the performance of the large language models (LLMs) by combining the strengths of information retrieval systems with the generative capabilities of LLMs. Here are the key advantages of using RAG:

1. Access to Up-to-Date Information: RAG allows LLMs to query external sources of information (e.g., databases, search engines) in real-time. This helps overcome the limitation of outdated knowledge in pre-trained models, enabling responses that reflect the most current facts and events.

2. Improved Accuracy and Relevance: By retrieving relevant documents or pieces of information from external databases, RAG enhances the accuracy of LLM outputs. It grounds the generated content in specific, factual data rather than relying solely on probabilistic generation, reducing the risk of hallucinations or inaccurate responses.

3. Handling Specialized Queries: For highly specialized or technical topics, RAG improves LLM performance by fetching domain-specific knowledge from relevant sources.

4. Reduced Memory Burden: LLMs often struggle to maintain context over long conversations. RAG can offload some of this burden by dynamically retrieving relevant information as needed, reducing the need to store or process large amounts of conversational history within the model.