Part 3 – Short Concept Questions

1. What is a RAG system, and why is it better than a plain chatbot?

   A Retrieval-Augmented Generation (RAG) system first retrieves pertinent documents from an external knowledge base in order to integrate language modelling and information retrieval. The model then produces responses based on reducing hallucinations, and enables the system to use private or updated data beyond the information it has retrieved. Thus, it outperforms a plain chatbot by enhancing factual accuracy, reducing hallucinations, and enabling the system to use private or updated data beyond its training set.

2. What are embeddings, and why are they used in vector search?

   Embeddings are the semantic meaning of text, images, or other data is represented by dense numerical vectors. In a vector space, similar items have close embeddings. Semantic search is made possible by vector search, which employs these embeddings to return contextually relevant results rather than precise keyword matching.

3. How does YOLO differ from traditional object detection methods?

   YOLO predicts bounding boxes and class probabilities in a single network pass, treating object detection as a single regression problem. Conventional techniques, such as R-CNN, employ several steps, including classification and region proposal. YOLO is therefore far quicker and better suited for real-time applications.

4. One challenge in real-world posture detection and how you'd handle it.

   A significant problem can be occlusion, in which body parts are obscured by objects or other people. By applying temporal consistency across video frames and utilizing pose estimation models trained on a variety of data, this can be resolved. Setups with multiple cameras can increase accuracy even more.