Assignment 12 - [GANs, VAEs, LLMs]

Q1)
ANS:-
Better domain generalization in Large Language Models (LLMs) is influenced by several key factors:

1. Model Architecture: Transformer-based architectures with deep layers and large parameters improve the model's ability to capture diverse linguistic patterns, enhancing generalization.

2. Training Data: Diverse and extensive training datasets expose the model to various linguistic contexts, improving its capacity to generalize to new domains.

3. Finetuning Techniques: Techniques like domain-adaptive pretraining and task-specific finetuning refine the model for specific domains, improving its ability to transfer knowledge to unseen domains.

Q2)
ANS:-

Mode collapse in Generative Adversarial Networks (GANs) occurs when the generator produces limited or repetitive outputs, failing to capture the full diversity of the target data distribution.

Symptoms:
- The generator repeatedly generates very similar samples, even when given different inputs.
- Diversity in generated outputs is significantly reduced.

Impact:
- Quality: Individual samples may look realistic, but the overall variety of outputs suffers.
- Diversity: The generator fails to represent the full range of possibilities in the data, leading to poor coverage of the data distribution.

Mode collapse limits the effectiveness of GANs in generating varied and high-quality outputs.

Q3)
ANS:-

Aspect	Standard Autoencoders	Variational Autoencoders (VAEs)
Latent Space	Deterministic, fixed latent representation	Probabilistic, continuous latent space (e.g., Gaussian)

Latent Space Structure	Unstructured, no regularization	Structured and regularized for smooth transitions
Generative Capability	Primarily reconstructs inputs, not designed for generation	Can generate new data points by sampling from the latent distribution
Reconstruction Focus	Focus on minimizing reconstruction error	Balances reconstruction accuracy and latent space regularization
Sampling	Direct encoding of inputs, no randomness	Samples from latent distributions to create diverse outputs
Key Strength	Good for dimensionality reduction and data compression	Ideal for generative tasks and producing diverse, realistic data
Output Diversity	Limited to reconstructing input data	Generates new, varied outputs by sampling from the latent space

Q4)

ANS:-

Self-Attention Mechanism Complexity:
- In a transformer, the self-attention mechanism computes attention scores between every pair of tokens in the input sequence. This results in a complexity of $O(n^2)$, where n is the sequence length & requires nxn operations in process.

Overall Transformer Complexity:
- The total complexity for a transformer layer, considering both the self-attention mechanism and feed-forward layers, is $O(n^2 \cdot d)$, where d is the dimensionality of the model (hidden size).
Implications for Long Sequences:

1. Memory Usage: As sequence length increases, both computational and memory requirements grow quadratically. This makes it challenging to process very long sequences efficiently, especially on limited hardware like GPUs.

2. Time Consumption: Processing long sequences becomes slower due to the quadratic growth of operations with respect to sequence length. This can significantly impact real-time applications or large-scale tasks.

3. Efficiency Trade-offs: To handle longer sequences, transformer variants such as Sparse Transformers, Linformers, and Longformers have been developed, which aim to reduce the complexity of self-attention from $O(n^2)$ to more manageable levels (e.g., $O(n \log n)$ or even $O(n)$), enabling the efficient processing of longer sequences without a massive increase in computational cost.