# Coding:

**Question 1: Multilingual Caption Aligner & Validator**

**Prompt:**  
Write a Python script that accepts a dataset of image-caption pairs in English and automatically generates corresponding captions in a user-specified target language (e.g., Urdu, Spanish, Chinese) using the Hugging Face transformers translation pipeline.

Then, validate that the **translated captions preserve core semantic meaning** by comparing the cosine similarity between the sentence embeddings of the original and translated captions using a multilingual sentence encoder (e.g., LaBSE). Discard translations with similarity < 0.75.

**Deliverables:**

* Translation function
* Similarity scoring module
* Output: JSON or CSV with {image\_path, en\_caption, lang\_caption, similarity}

**Question 2: Dynamic Token-Region Mapping Visualizer**

**Prompt:**  
Given a pretrained Stable Diffusion model and a text prompt in English or another supported language, write a script that outputs the **text-to-image attention map** at each denoising step.

Then, implement a visual heatmap overlay on the generated image showing which words influenced which regions the most. The tool should support multilingual captions.

**Deliverables:**

* Code to hook into Stable Diffusion attention layers
* Multilingual prompt compatibility
* Heatmap visualizer

**Question 3: Cross-Lingual Prompt Cycle Evaluator**

**Prompt:**  
Build a function that accepts a prompt in any target language (e.g., Chinese), translates it to English, and then **back-translates** it into the original language using two separate translation models.

Compare the final back-translated sentence with the original using token-level edit distance and character-level BLEU score.

Highlight mismatches and compute an **explainability score** (0 to 1) indicating how well the meaning was preserved.

**Deliverables:**

* Translation pipeline
* Scoring module with token-level comparison
* CLI-based report output

**Question 4: Multilingual Token Embedding Collapser**

**Prompt:**  
Implement a function that takes any sentence in the target language, tokenizes it using a multilingual encoder (e.g., XLM-R), and **collapses subword tokens** (like "▁com", "puter") into full-word representations, then visualizes the average embedding per word using t-SNE or PCA.

Ensure compatibility across Urdu, Arabic, Chinese, and Spanish—each with different subword tokenization behavior.

**Deliverables:**

* Tokenizer abstraction layer
* Embedding aggregation logic
* Interactive 2D visualization plot

**Question 5: Prompt Paraphrase Discriminator**

**Prompt:**  
Design a binary classifier that takes **two prompts in the same or different languages** and predicts whether they describe the **same visual scene**.

Train on manually curated pairs (e.g., “a red bicycle under a tree” vs. “a red cycle parked beneath a tree” or their translations) using embeddings from LaBSE or mT5.

Ensure the system is robust to paraphrasing, synonyms, and language-specific expressions. Use at least one evaluation metric such as ROC-AUC.

**Deliverables:**

* Dataset creation strategy
* Classification model and training code
* Evaluation report

# Theoretical:

**1. Explain the differences between CLIP and traditional image-caption embedding models. How does CLIP handle cross-modal similarity?**

**Expected Concepts:**

* Contrastive learning with image-text pairs
* Joint embedding space
* Zero-shot generalization
* Differences from classification models (e.g., ResNet + classifier)

**2. What are the key challenges in multilingual text representation, and how do models like XLM-R or LaBSE address them?**

**Expected Concepts:**

* Vocabulary sharing vs. separate embeddings
* Sentence-level semantics across languages
* Pretraining with masked language modeling
* Language agnostic representations

**3. Why is “cross-lingual semantic preservation” important in text-to-image generation? How would you evaluate whether two multilingual prompts produce semantically consistent images?**

**Expected Concepts:**

* Language-independent visual grounding
* Cosine similarity in multilingual embedding space
* CLIPSim, visual alignment, or image-caption back-translation comparison

**4. How do diffusion models (like Stable Diffusion) generate images from text? Describe the forward and reverse processes briefly.**

**Expected Concepts:**

* Forward process: noise added to image
* Reverse process: image denoising conditioned on text
* Latent space usage for faster training (in LDM)
* Conditioning via cross-attention on text embeddings

**5. What is the role of attention in text-to-image synthesis? How would you adapt attention for multilingual prompts?**

**Expected Concepts:**

* Attention maps link text tokens to image regions
* Cross-attention in transformers/diffusion models
* Tokenization effects (e.g., wordpiece handling in Chinese vs. Urdu)
* Embedding normalization across languages

**6. Describe the purpose of FID and IS scores in evaluating generative models. What are their limitations in multilingual or fine-grained settings?**

**Expected Concepts:**

* FID: distributional similarity to real images
* IS: diversity and recognizability of generated images
* Limitations: not tied to prompt-text fidelity, not multilingual-aware

**7. How can you use contrastive loss for ensuring consistency across multilingual prompts in a text-to-image model?**

**Expected Concepts:**

* Pulling together embeddings from equivalent prompts in different languages
* Encouraging similar image outputs
* Triplet loss or dual-encoder mechanisms

**8. Why are standard tokenizers (like BPE or SentencePiece) problematic for cross-lingual tasks involving scripts like Arabic, Chinese, or Urdu? What are good practices to handle this?**

**Expected Concepts:**

* Language-specific tokenization issues
* Subword fragmentation
* Importance of consistent preprocessing
* Using pretrained multilingual tokenizers with normalization