



# SNS COLLEGE OF TECHNOLOGY

Autonomous | NAAC A++ | NBA  
Accreditation



## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

### GENERATIVE AI

#### UNIT I- GENERATIVE MODELS FOR TEXT & IMAGE

##### PUZZLE

### 1. Language Model Families (GPT-series, LLaMA, Gemini)

Puzzle: The AI Consultant's Dilemma

Industry Context:

A multinational company wants to deploy an internal AI assistant for employees.

Given Systems:

1. GPT-series model via API
2. LLaMA deployed on private servers
3. Gemini integrated with images, audio, and text

Puzzle Task:

Match each model family with:

- Deployment style (Cloud-only / On-premise / Hybrid)
- Strength in multimodality (Low / Medium / High)
- Data privacy suitability
- Typical enterprise use case

Twist Question:

Why might an open-weight model outperform a larger proprietary model in some organizations?

Concepts Tested:

- LLM families
- Open vs. closed models
- Deployment trade-offs

---

### 2. Vision Foundation Models (Stable Diffusion, Midjourney, DALL·E)

Puzzle: The Creative Studio Showdown

Industry Context:

A design studio creates visuals for ads, games, and social media.

Given Models:

- Stable Diffusion
- Midjourney
- DALL·E

Puzzle Task:

Identify which model best fits the following requirements:

1. Full control and local customization
2. Highest artistic aesthetics with minimal prompts
3. Strong text-to-image alignment and safety filters

Challenge Question:

Why are diffusion models considered foundation models for vision tasks?

Concepts Tested:

- Vision foundation models
  - Diffusion-based generation
  - Model accessibility vs. control
- 

### **3. Core Preprocessing: Tokenization**

Puzzle: The Token Trap

Industry Context:

A chatbot gives inconsistent answers for similar-looking sentences.

Given Sentences:

- "AI-driven innovation"
- "AI driven innovation"
- "AIdriven innovation"

Puzzle Task:

1. Explain how subword tokenization handles each case
2. Predict which sentence results in the highest token count
3. Identify one risk of poor tokenization for multilingual models

Mini-Challenge:

Why do LLMs prefer subword tokenization over word-level tokenization?

Concepts Tested:

- Tokenization strategies

- BPE / WordPiece intuition
  - Language robustness
- 

#### **4. Core Preprocessing: Embeddings**

Puzzle: The Vector Map Mystery

Industry Context:

A search engine retrieves irrelevant results despite correct keywords.

Puzzle Task:

Answer the following:

1. What does each embedding vector represent?
2. Why are cosine similarity and dot product commonly used?
3. Which is closer in embedding space and why?
  - "doctor" and "hospital"
  - "doctor" and "engine"

Twist Question:

How do contextual embeddings differ from static embeddings like Word2Vec?

Concepts Tested:

- Semantic representations
  - Embedding spaces
  - Context awareness
- 

#### **5. Introduction to Fine-tuning**

Puzzle: Train or Adapt?

Industry Context:

A startup wants a domain-specific chatbot for legal documents.

Constraints:

- Small dataset
- Limited GPU budget
- High accuracy required

Puzzle Task:

Decide whether to:

- Train from scratch
- Fully fine-tune
- Use parameter-efficient fine-tuning

Explain your choice.

Concepts Tested:

- Fine-tuning strategies
  - Cost-performance trade-offs
- 

## 6. LoRA (Low-Rank Adaptation)

Puzzle: The Rank Reduction Case

Industry Context:

A 7B parameter LLM must be adapted for customer support FAQs.

Puzzle Task:

1. Identify which parameters LoRA modifies
2. Explain why low-rank matrices reduce training cost
3. Decide whether inference speed changes after applying LoRA

Challenge Question:

Why is LoRA ideal for multi-task adaptation?

Concepts Tested:

- Parameter-efficient learning
  - Matrix decomposition intuition
  - Practical deployment
- 

## 7. QLoRA

Puzzle: Precision vs. Performance

Industry Context:

A company wants to fine-tune a large LLM on a single GPU.

Given Clues:

- 4-bit quantization
- Frozen base model
- LoRA adapters on top

Puzzle Task:

1. Explain how QLoRA reduces memory usage
2. Identify one risk of aggressive quantization
3. State one reason accuracy is still preserved

Decision Puzzle:

When would full fine-tuning still be preferred over QLoRA?

Concepts Tested:

- Quantization
  - Efficient adaptation
  - Hardware-aware training
- 

## **8. Capstone Puzzle (Integrated)**

Puzzle: Deploy the Perfect AI Stack

Industry Context:

An ed-tech platform wants an AI tutor that explains concepts using text and images.

Constraints:

- Limited budget
- Need for multimodality
- Low hallucination tolerance
- Scalable to thousands of users

Puzzle Task:

Decide:

1. LLM family to choose
2. Vision model to integrate
3. Tokenization + embedding strategy
4. Adaptation method (LoRA / QLoRA / RAG)

Justify each decision briefly.

Concepts Tested:

- Integrated system design
- Foundation models
- Real-world constraints