POKHARA ENGINEERING COLLEGE
Internal Assessment Examination

Level: Bachelor | Semester – Spring | Year : 2025
Programme: Computer | Full Marks: 100
Course: Generative AI | Pass Marks: 45
Time : 3hrs.

Candidates are required to give their answers in their own words as far as practicable.

The figures in the margin indicate full marks.

Attempt all the questions.

1. Explain how AI has evolved from rule-based systems to today's generative models like ChatGPT. Emphasize how generative AI differs from traditional AI systems in terms of learning, reasoning, and outputs. Provide examples. 10

2. Python has become one of the most widely used programming languages, especially in the field of AI and data science. 5+5+3+7

   a. In your own words, explain why Python is so popular in today's tech landscape, particularly in AI and machine learning. Support your answer with at least three solid reasons, such as its ecosystem, syntax, or community support.
   b. What is the difference between a Python list and a NumPy array? Why would you use NumPy in AI projects?
   c. Write a simple function in Python that takes a list of numbers and returns a new list with each number squared.
   d. Explain the role of FastAPI in building AI-powered applications.How does it help in deploying machine learning or generative AI models effectively? Mention at least two key features of FastAPI that make it suitable for such applications.

3. In your own words, explain how transformer models work and why they are more effective for generative tasks like summarization, translation, or text generation. You may use a real-world analogy (e.g., teamwork, conversation, reading a book) to explain how attention and parallel processing help transformers perform better than older models. Explain with relevant diagrams. 10

4. You're working on a university assistant system that helps students ask questions like "How do I apply for a semester break?" based on student handbooks and policy documents. 5+5+5

   To build this system, you need to convert the documents into a format that can be searched using vector similarity.

   In your own words, explain the following concepts and how they help in building this system:

   a. What is chunking and why is it important before generating embeddings?
   b. What are vector embeddings? Why is it important?
   c. What search strategies can be used (e.g., similarity search, hybrid search), and how do they differ?

5. You are designing a support chatbot that needs to handle both billing-related and technical queries using a single large language model (LLM) API. The chatbot often gives incorrect or irrelevant responses. 5+5+5+5

   To improve the chatbot's reliability using prompt engineering only, explain the following prompting techniques you've learned:

   a. Zero-shot prompting
   b. Few-shot prompting
   c. Chain-of-thought prompting
   d. Role based prompting

For each technique, describe:

a. What it is.
b. When and why you would use it
c. An example of how it could be applied in the context of this chatbot.

6. You are assigned to develop a context-aware chatbot for a university that can answer questions like "What is the process to defer a semester?" based on uploaded documents (e.g., academic policies, handbooks). 5+5+5 +5+5

Explain the end-to-end process of building such a chatbot. Your answer should include:

a. What components are needed in the frontend and backend to support uploading documents and chatting with the bot.
b. What happens after a document is uploaded — how is it processed and stored to make it searchable?
c. What is vectorization, and why is it important for this system?
d. What kind of database is used to store and retrieve embeddings?
e. How is a user's question handled — what steps are involved in retrieving relevant information and generating the final response?