1. What is Prompt Engineering?

Prompt engineering involves crafting and refining input prompts to guide language models (like GPT) toward generating accurate and relevant responses. It's a way of optimizing how we communicate with AI to get the best possible outputs.

2. What is Prompt Injection? What are the different types of prompt injection?

Prompt injection is a method of manipulating language models by inserting unauthorized input to influence their outputs. The two main types are:

- Direct prompt injection: Malicious prompts are inserted directly into the model's input.
- Indirect prompt injection: Malicious inputs are embedded within the data that the model processes.

3. What are the key advantages of using Retrieval Augmented Generation (RAG)?

- Current information: RAG can pull real-time data, ensuring responses are timely and relevant.
- Improved accuracy: By retrieving facts as needed, RAG can enhance the factual correctness of responses.
- Efficiency: RAG reduces the need for massive models by pulling information dynamically instead of relying solely on pre-trained knowledge.

4. What are the essential components of the ReAct prompting framework?

- Reasoning: The model is prompted to explain its thought process.
- Action: The model is guided to take actions based on its reasoning.
- Reflection: The model is asked to review and refine its previous decisions or actions.

5. What are the main advantages of Dense Retrieval over Sparse Retrieval in RAG systems?

- Greater relevance: Dense retrieval uses semantic matching to find more relevant content, while sparse retrieval focuses on keyword matching.
- Better handling of similar concepts: Dense retrieval can match meanings rather than just words, making it more flexible.
- Reduced noise: It filters out irrelevant results by focusing on conceptual rather than exact keyword matches.

6. Top-k Sampling (k=3):

For the probability distribution provided:

Word C (0.45), Word D (0.20), and Word A (0.15) would be in the candidate set with k=3 as they are the top three probabilities.

7. Top-p Sampling (p=0.7):

Given the probability distribution:

The smallest candidate set that meets the p=0.7 threshold includes Word A (0.45) and Word B (0.25), which together account for 70% of the probability.

8. Which factors could be influenced by the “temperature” setting when interacting with an LLM?

- Creativity level: Higher temperatures increase the diversity and randomness of the model's responses, while lower temperatures make them more focused and deterministic.
- Response variability: A high temperature can lead to more varied and unpredictable outputs.

9. What are the main advantages of COT (Chain of Thought) over Zero-Shot prompting?

- Improved reasoning: COT prompts break down complex tasks into logical steps, leading to more coherent results.
- Greater accuracy: By encouraging intermediate reasoning, COT often produces more precise and reliable answers.

10. What are the advantages of Auto-COT over COT?

- Automation of reasoning: Auto-COT automatically generates reasoning paths without the need for manual intervention.
- Increased efficiency: It streamlines the problem-solving process by providing multiple reasoning examples for better model performance.

11. What are the advantages of meta-prompting?

- Adaptable responses: Meta-prompting allows the model to adjust its behavior dynamically during a task.
- Versatility: It is useful for various tasks, enabling the model to generalize its responses effectively.

12. How should a research team exploring LLMs for creative writing control temperature and top-k parameters?

- Higher temperature: Encourages more creative, unexpected responses in creative writing.
- Lower top-k: Keeps the model's outputs focused on more relevant word choices while still allowing some degree of novelty.