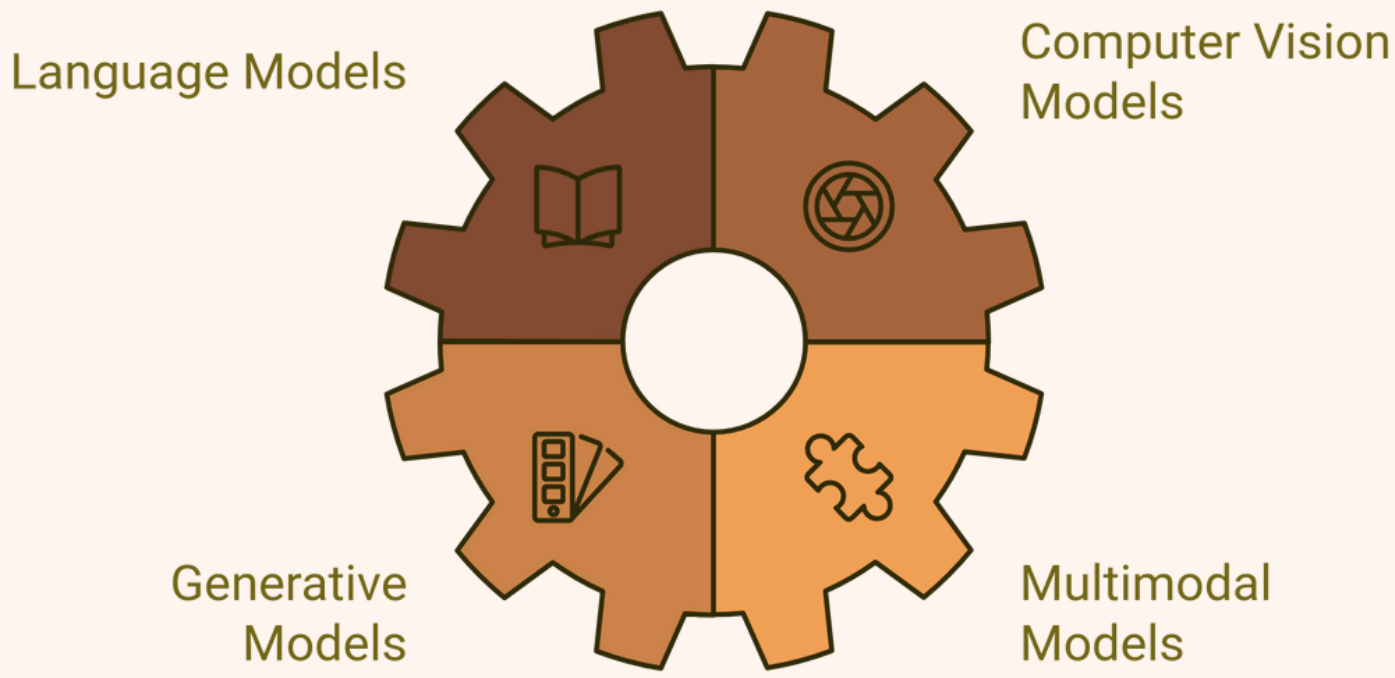
10 LLM INTERVIEW QUESTIONS

Bhavishya Pandit

Q1. What are different types of Foundation Models?

-Foundation models are large-scale AI models trained on vast amounts of unlabeled data using unsupervised methods. They are designed to learn general-purpose knowledge that can be applied to various tasks across domains. Common Types of Foundation Models-



1.Language Models -

Tasks: Machine translation, text summarization, question answering

Examples: BERT, GPT-3

2.Computer Vision Models -

Tasks: Image classification, object detection, image segmentation

Examples: ResNet, VGGNet

3.Generative Models -

Tasks: Creative writing, image generation, music composition

Examples: DALL-E, Imagen

4.Multimodal Models -

Tasks: Image captioning, visual question answering

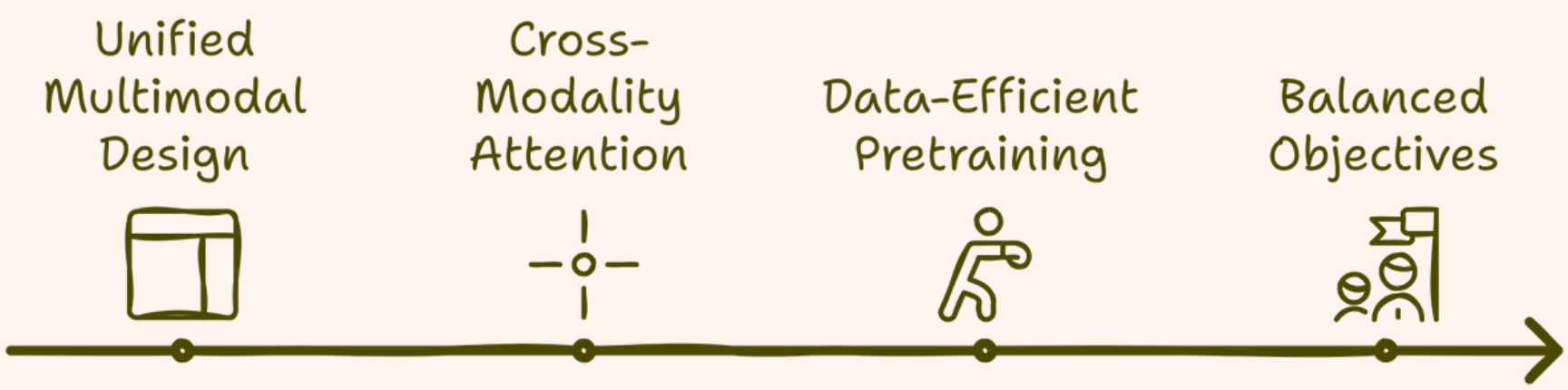
Examples: PaLM, LaMDA

Q2. In what ways does Gemini's architecture optimize training efficiency and stability compared to other multimodal LLMs like GPT-4?

-Gemini's architecture optimizes training efficiency and stability compared to multimodal models like GPT-4 in several ways:

1.**Unified Multimodal Design:** Gemini integrates text and image processing in a single model, improving parameter sharing and reducing complexity.

2.Cross-Modality Attention: Enhanced interactions between text and images lead to better learning and stability during training.



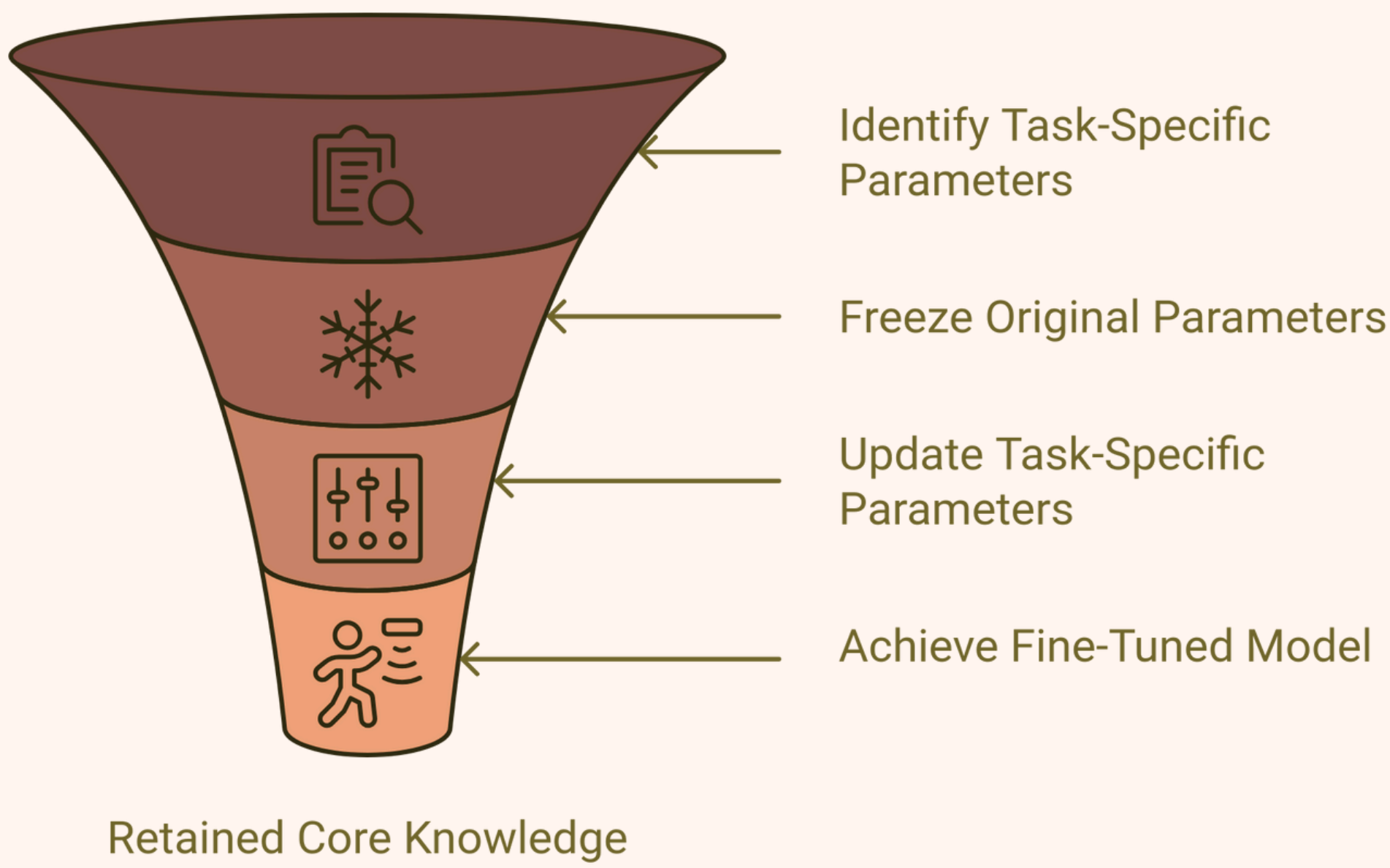
3.Data-Efficient Pretraining: Self-supervised and contrastive learning allow Gemini to train with less labeled data, boosting efficiency.

4.Balanced Objectives: Better synchronization of text and image losses ensures stable training and smoother convergence.

Q3. How does Parameter-Efficient Fine-Tuning (PEFT) prevent catastrophic forgetting in LLMs?

-

Parameter-Efficient Fine-Tuning (PEFT) helps prevent catastrophic forgetting in LLMs by updating only a small set of task-specific parameters, while keeping most of the model's original parameters frozen. This approach allows the model to adapt to new tasks without overwriting previously learned knowledge, ensuring it retains core capabilities while learning new information efficiently.

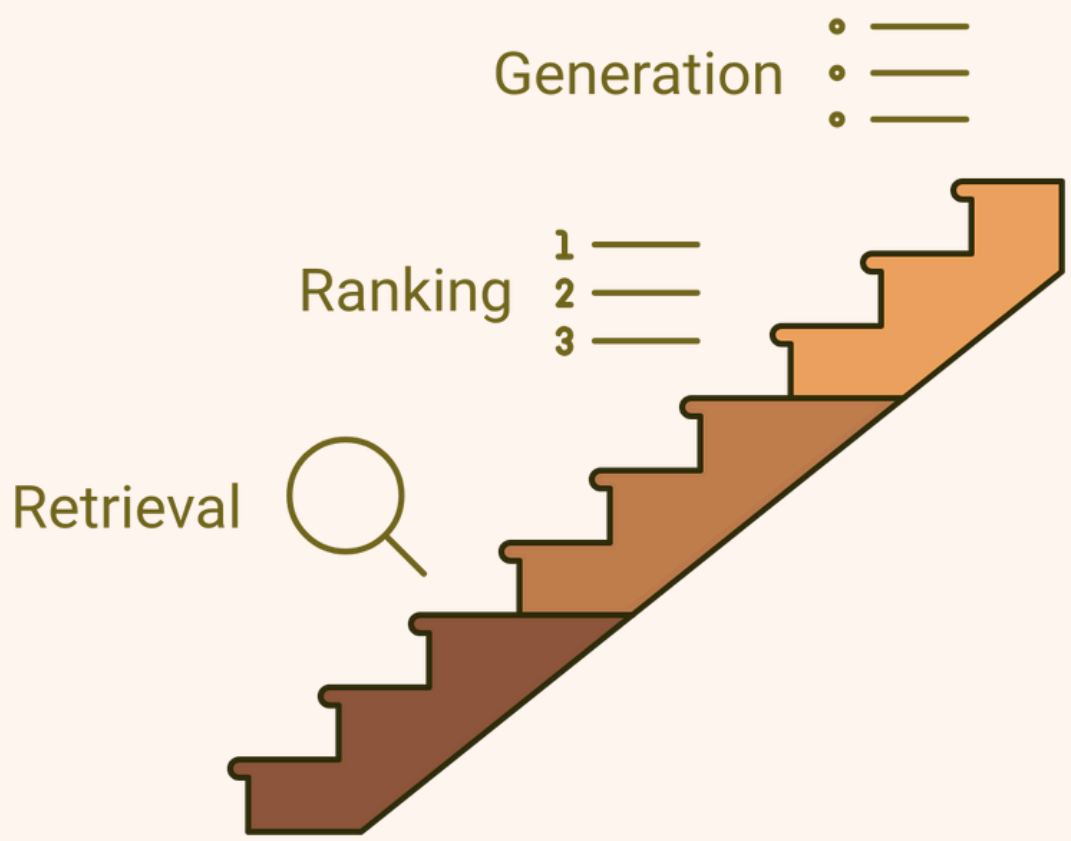


Q4. What are the key steps involved in the Retrieval-Augmented Generation (RAG) pipeline?

Key steps in the Retrieval-Augmented Generation (RAG) pipeline are:

1.Retrieval: The query is encoded and compared with precomputed document embeddings to retrieve relevant documents.

2. Ranking: The retrieved documents are ranked based on their relevance to the query.

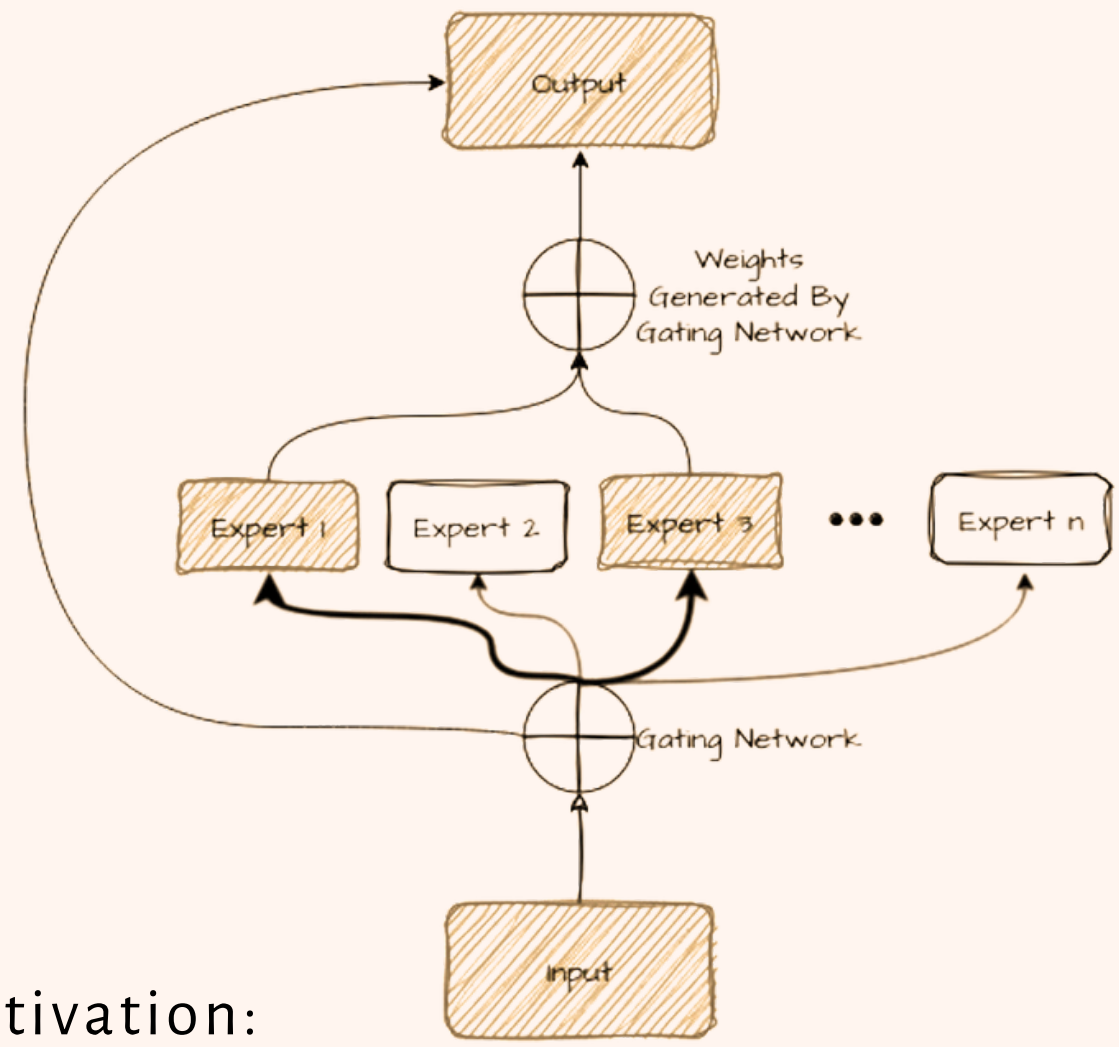


3. Generation: The top-ranked documents are used as context by the LLM to generate more informed and accurate responses.

This hybrid approach enhances the model's ability to produce context-aware outputs by incorporating external knowledge during generation.

Q5. How does the Mixture of Experts (MoE) technique improve LLM scalability?

Mixture of Experts (MoE) improves LLM scalability by using a gating function to activate only a subset of expert models (sub-networks) for each input, rather than the entire model.



This selective activation:

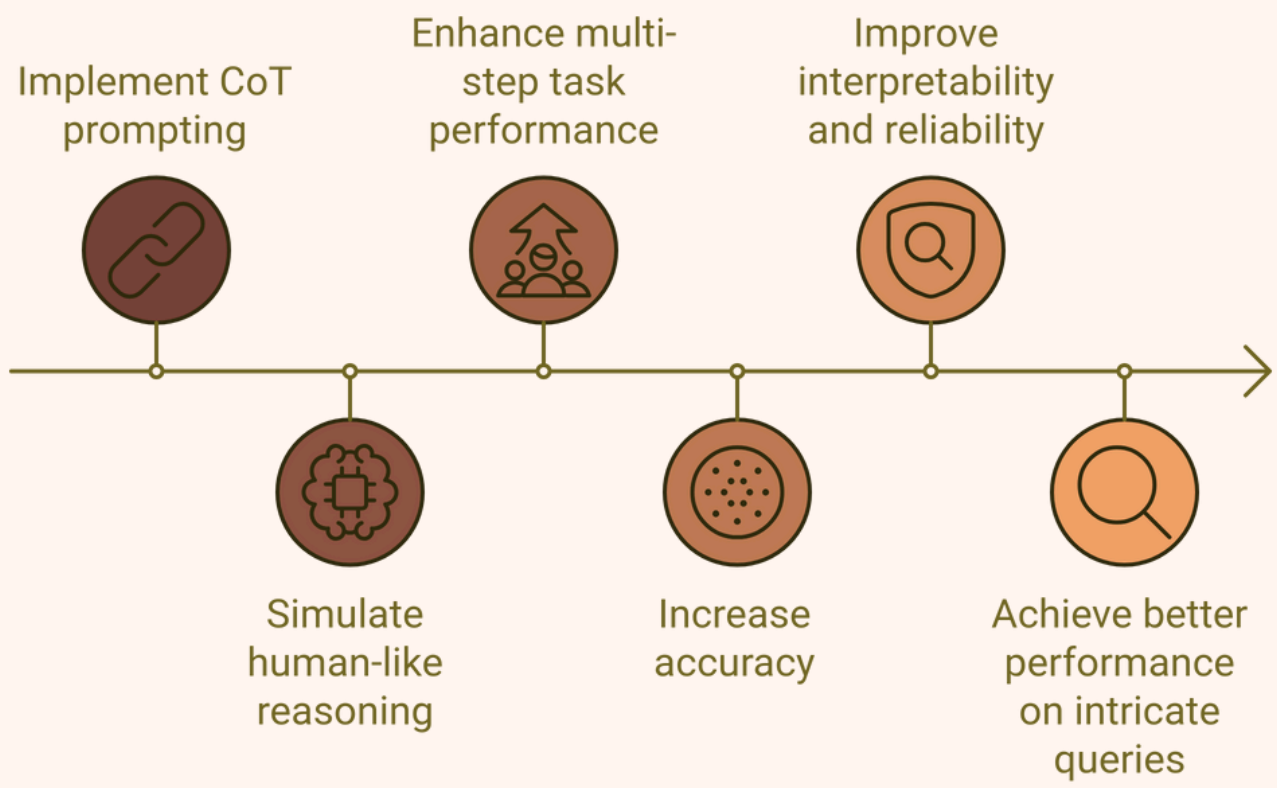
- Reduces computational load: Only a few experts are active per query, minimizing resource usage.
- Maintains high performance: The model dynamically selects the most relevant experts for each input, ensuring task complexity is handled effectively.

MoE enables efficient scaling of LLMs, allowing larger models with billions of parameters while controlling computational costs.

Bhavishya Pandit

Q6. What is Chain-of-Thought (CoT) prompting, and how does it improve complex reasoning in LLMs?

-Chain-of-Thought (CoT) prompting helps LLMs handle complex reasoning by encouraging them to break down tasks into smaller, sequential steps. This improves their performance by:



- Simulating human-like reasoning: CoT prompts the model to approach problems step-by-step, similar to how humans solve complex issues.

- Enhancing multi-step task performance: It's particularly effective for tasks involving logical reasoning or multi-step calculations.

- Increasing accuracy: By guiding the model through a structured thought process, CoT reduces errors and improves performance on intricate queries.

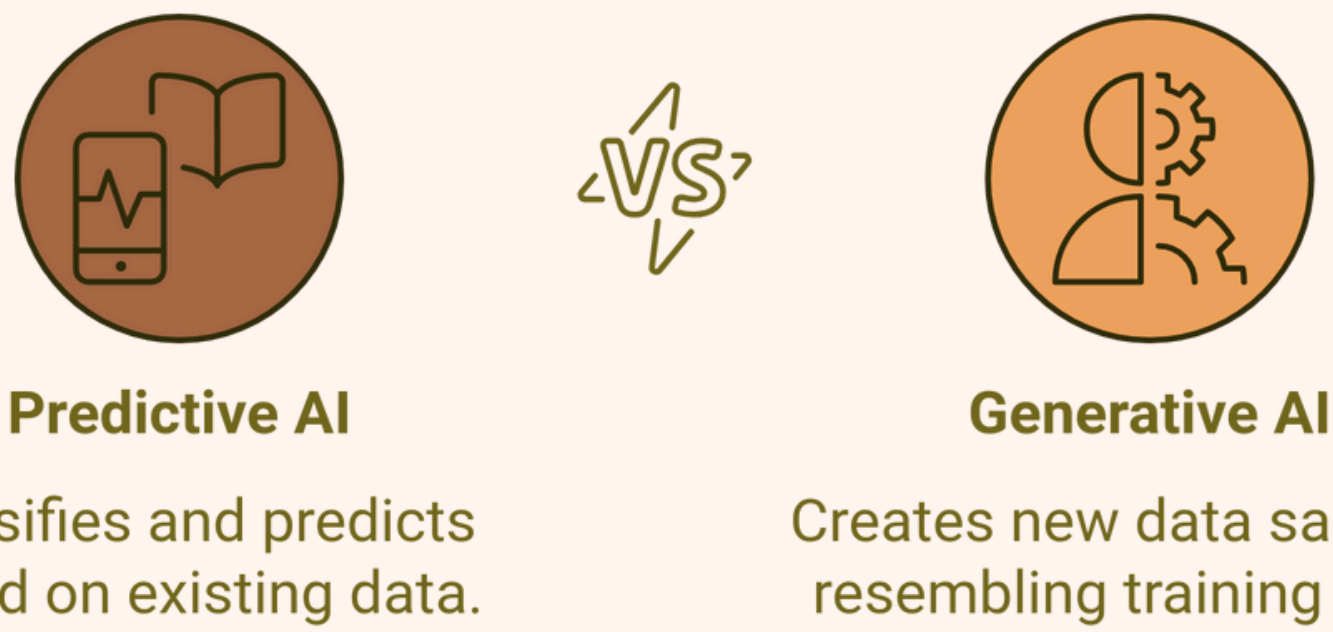
CoT improves LLMs' interpretability and reliability in tasks that require deeper reasoning and decision-making.

Q7. What is the difference between discriminative AI and Generative AI?

-Predictive/Discriminative AI:

- Focuses on predicting or classifying data based on existing data. It models the conditional probability P(y|x)P(y|x)P(y|x), where y is the target variable and x represents the input features.

- Examples include tasks like classification (e.g., image recognition), regression (e.g., predicting stock prices), and applications such as spam detection and disease diagnosis.



Generative AI:

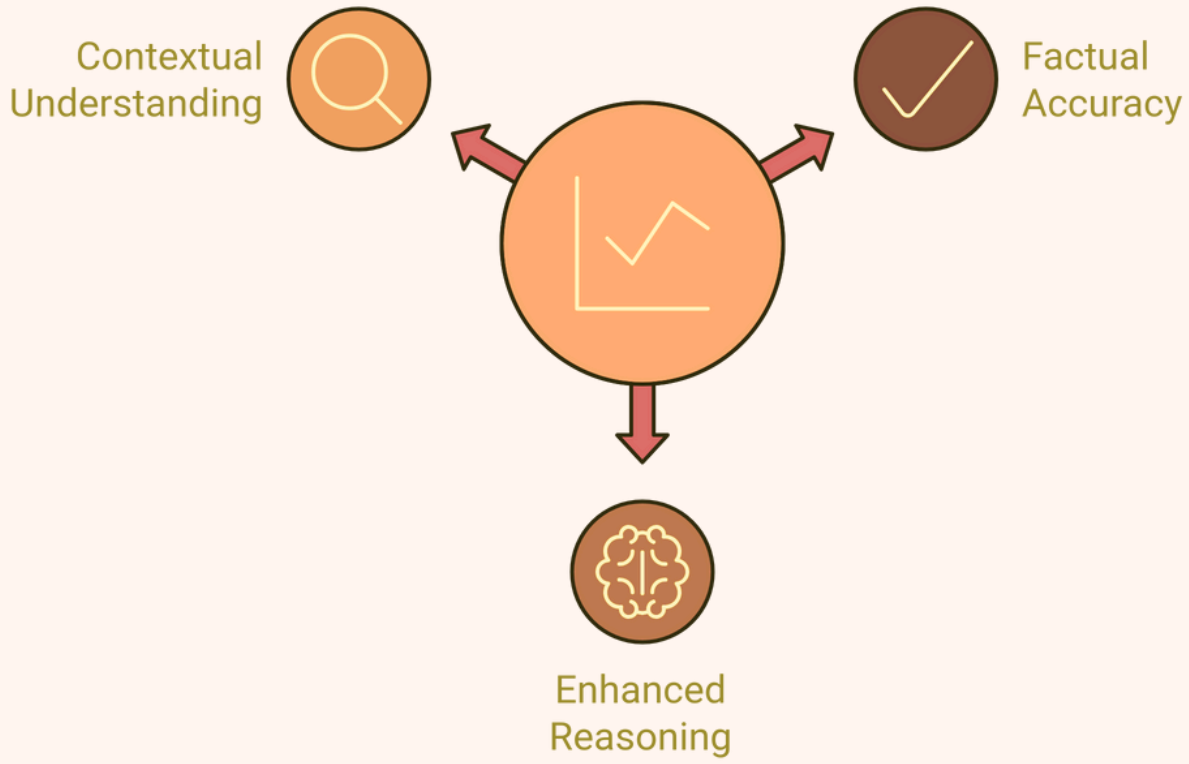
- Focuses on generating new data samples that resemble the training data. It models the joint probability P(x,y)P(x, y)P(x,y), allowing it to create new instances of data

- Examples include generating text, images, music, and other content. Techniques used are Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and large language models like GPT.

Q8. How does knowledge graph integration enhance LLMs?

Integrating knowledge graphs with LLMs enhances performance by adding structured, factual knowledge. Key benefits include:

- **Factual accuracy:** The model can cross-check information against the knowledge graph, reducing hallucinations and improving correctness



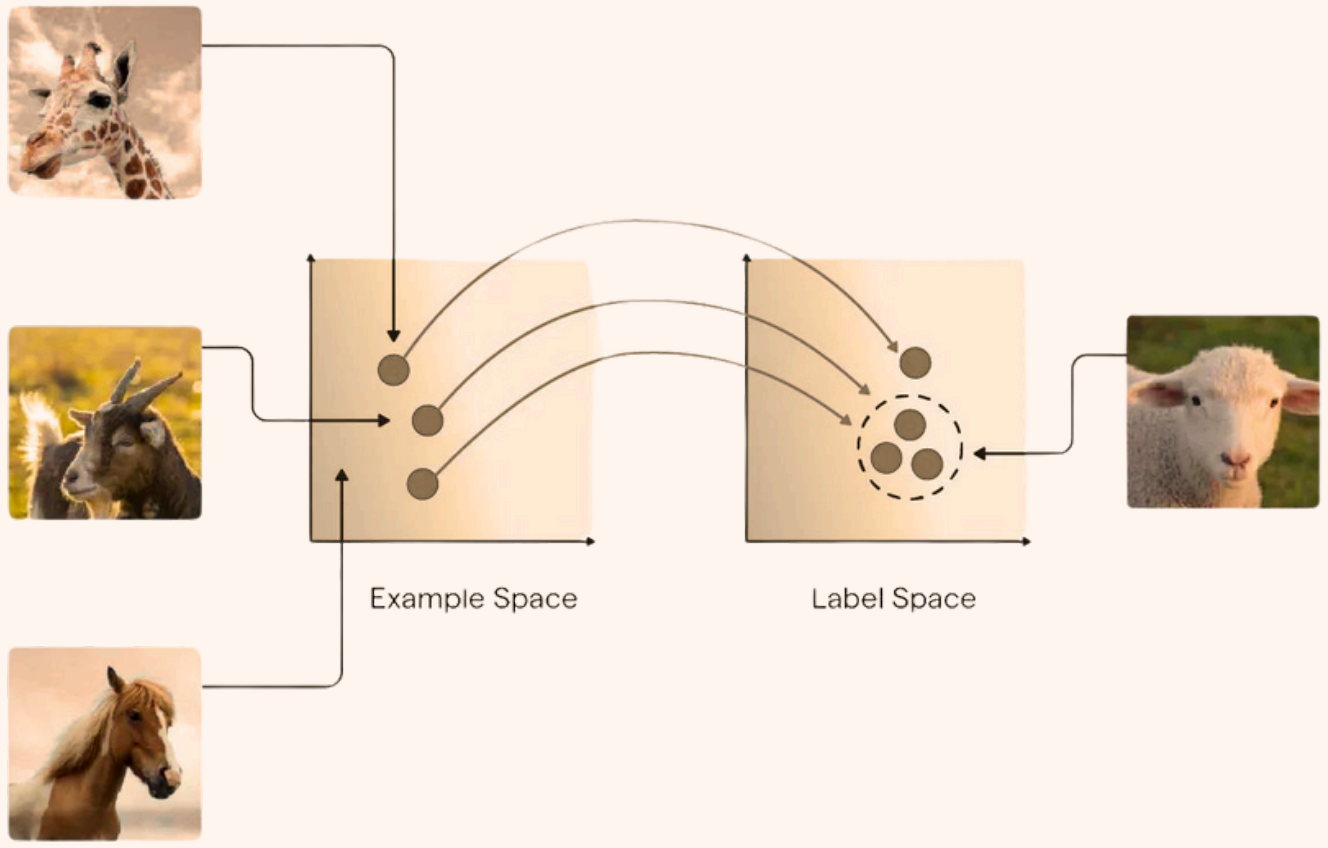
- **Enhanced reasoning:** Knowledge graphs support logical reasoning by leveraging relationships between entities, enabling better handling of complex queries.

- **Contextual understanding:** Structured data helps the model understand context and relationships, improving response quality.

This integration is particularly valuable in tasks like question answering, entity recognition, and recommendation systems, where structured knowledge plays a critical role.

Q9. What is zero-shot learning, and how does it apply to LLMs?

Zero-shot learning enables LLMs to perform tasks they haven't been explicitly trained for by leveraging their broad understanding of language and general concepts. Instead of needing task-specific fine-tuning, the model can generate relevant outputs based on the instructions provided in the prompt.



For example:

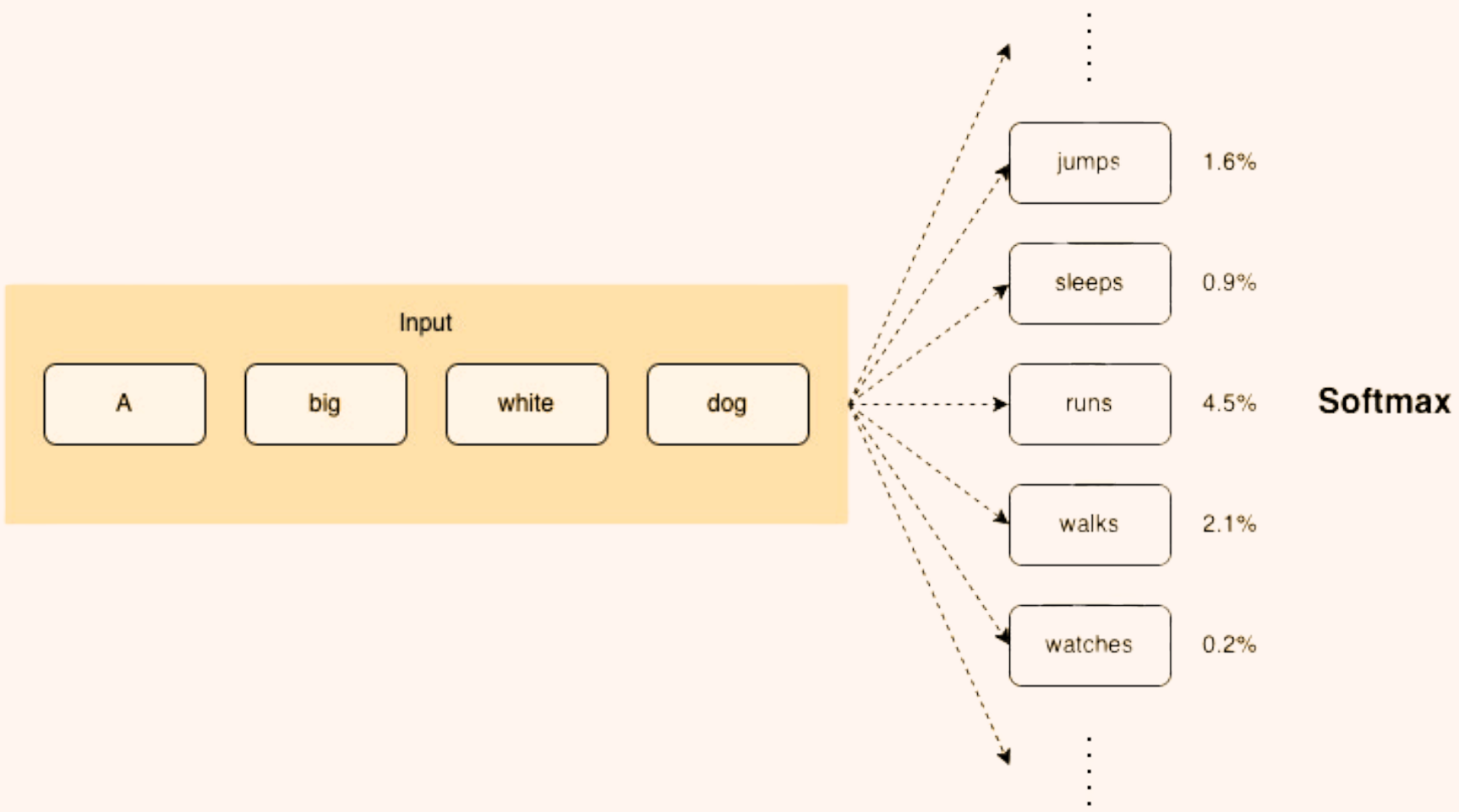
- **Text classification:** The model can categorize text without specific training, simply by understanding the prompt's context.

- **Translation or summarization:** LLMs can translate or summarize text using provided instructions, even without task-specific fine-tuning.

This shows the LLMs' ability to generalize across tasks, making them versatile for various applications.

Q10. How does Adaptive Softmax speed up large language models?

Adaptive Softmax accelerates LLMs by categorizing words into frequency groups, allowing for fewer computations for infrequent words. This approach lowers overall computational costs while preserving accuracy, making it effective for efficiently managing large vocabularies.



Follow for more AI/ML posts

SAVE

LIKE

SHARE

Bhavishya Pandit