



# Introduction to Large Language Models

From Foundation to Real World Applications

Pritam Prakash Shete  
Scientific Officer G  
Computer Division, BARC





# Agenda

Introduction

Large Language Models

Prompt Engineering

In Context Learning

Training LLMs

LLM Applications


Deploying LLMs

Conclusions





# Introduction

- OpenAI
    - GPT 3.5, GPT 4
    - GPT 4o, GPT o1
  - Access – API
  - Multimodal
    - Text, Image
- 
- Google
    - Gemini 1
    - Gemini 1.5
  - Access – API
  - Multimodal
    - Text, Image



# Introduction

- Meta

- Llama 2, Llama 3
- Llama 3.1, Llama 3.2

- Access – Open Source

- Multimodal

- Text, Image


- Google

- Gemma

- Gemma 2

- Access – Open Source

- Multimodal

- Text, Image
- 



# Large Language Models

- Next word or token prediction

I love eating ice

I love eating ice cream

I love eating ice cream with

I love eating ice cream with my

I love eating ice cream with my friends

I love eating ice cream with my friends at

I love eating ice cream with my friends at beach.

# Large Language Models

- Natural Language Processing (NLP)
- Understand, Interpret, and Generate human language
- Sentiment analysis
- Text summarization
- Machine translation

# Large Language Models


- Large number of parameters
  - ChatGPT-4 - ~1.8 Trillion parameters
- Trained using large amount of dataset
  - ChatGPT-4 - 570 GB or 300 Billion words

\*OpenAI ChatGPT 4





# Large Language Models



|                        |                         |
|------------------------|-------------------------|
| • Model name           | Llama 3.1 405B Instruct |
| • Company              | Meta                    |
| • Access               | Open source             |
| • Number of parameters | 405 Billion parameters  |
| • Context size         | 128,000 (128K) tokens   |
| • LLM type             | Instruction fine-tuned  |

\*Meta Llama 3.1 405B Instruct

# Sentiment Analysis

Identify sentiment of following sentence as either positive or negative.

I like "Harry Potter and the Goblet of Fire" book very much.

The sentiment of the sentence "I like 'Harry Potter and the Goblet of Fire' book very much" is **Positive**.

The word "like" and the phrase "very much" both indicate a **strong positive sentiment**, suggesting that the speaker has a very positive opinion of the book.

# Summarization

Summarize the following paragraph.

A chemical reaction is a process that leads to the chemical transformation of one set of chemical substances to another. When chemical reactions occur, the atoms are rearranged and the reaction is accompanied by an energy change as new products are generated. Classically, chemical reactions encompass changes that only involve the positions of electrons in the forming and breaking of chemical bonds between atoms, with no change to the nuclei (no change to the elements present), and can often be described by a chemical equation.

A chemical reaction transforms one set of substances into another by rearranging atoms and changing energy levels. This process typically involves changes in electron positions and chemical bonds without altering the elements themselves, and is often represented by a chemical equation.

# Machine Translation

Translate following English sentences in Hindi sentences.

Particle accelerator is a machine that uses electromagnetic fields to propel charged particles to very high speeds and energies to contain them in well defined beams.

कण त्वरक एक मशीन है जो विद्युत चुम्बकीय क्षेत्रों का उपयोग करके आवेशित कणों को बहुत अधिक गति और ऊर्जा प्रदान करती है और उन्हें अच्छी तरह से परिभाषित किरणों में सीमित रखने के लिए।



# Prompt Engineering

- Prompt
  - Inference
  - Completion
- 
- 
- 

# Prompt

- Instructions to LLM
- Context for Question
- Question

Answer following question using given context.

Context: Everyday examples of particle accelerators are cathode ray tubes found in television sets and X-ray generators. These low energy accelerators use a single pair of electrodes with a DC voltage of a few thousand volts between them.

Question: Give examples of low energy accelerators.

\*Meta-Llama-3.1-8B-Instruct

# Inference

- Large Language Model
- Model prediction

Translate following English sentences in Hindi sentences.

Particle accelerator is a machine that uses electromagnetic fields to propel charged particles to very high speeds and energies to contain them in well defined beams.

कण त्वरक एक मशीन है जो विद्युत चम्बकीय क्षेत्रों का उपयोग करके आवेशित कणों को बहुत अधिक गति और ऊर्जा प्रदान करती है और उन्हें अच्छी तरह से परिभाषित किरणों में सीमित रखने के लिए।



# Prompt Completion

- Inference without context

कण त्वरक एक मशीन है जो विद्युत चम्बकीय क्षेत्रों का उपयोग करके आवेशित कणों को बहुत अधिक गति और ऊर्जा प्रदान करती है और उन्हें अच्छी तरह से परिभाषित किरणों में सीमित रखने के लिए।

# In Context Learning

- Context window
- Task example/s
- Zero shot inference
- One shot inference
- Few shot inference



# Context Window

- Window size
  - ChatGPT 4 Turbo - 128K
  - Llama 3.1 405B - 128K

Answer following question using given context.

Context: The Large Hadron Collider (LHC) particle collider is the world's largest and highest-energy particle accelerator. It was built by the European Organization for Nuclear Research (CERN).

Question: What is the Large Hadron Collider?

# Task Examples

- One or more examples
- In context learning
- Align LLM with task

Identify sentiment of following sentence as either positive or negative.

Sentence: I like book very much.

Sentiment: Positive

Sentence: I do not like book.

Sentiment: Negative

Sentence: Harry Potter and the Goblet of Fire book is good for readers.

# Zero Shot Inference

- No examples

Identify sentiment of following sentence as either positive or negative.

Sentence: Harry Potter and the Goblet of Fire book is good for readers.

# One Shot Inference

- One example

Identify sentiment of following sentence as either positive or negative.

Sentence: I like book very much.

Sentiment: Positive

Sentence: Harry Potter and the Goblet of Fire book is good for readers.

# Few Shot Inference

- Two or more examples

Identify sentiment of following sentence as either positive or negative.

Sentence: I like book very much.

Sentiment: Positive

Sentence: I do not like book.

Sentiment: Negative

...

Sentence: Harry Potter and the Goblet of Fire book is good for readers.



# Training Large Language Models

- Model Pre-training
- Instruction Fine Tuning
- Reinforcement Learning from Human Feedback (RLHF)

# Model Pre-training

- Self supervised learning
  - Next word or token prediction
  - Self annotations – (X – y)
    - I love ice cream
      - X = I love ice
      - y = cream
  - Learn language syntax
  - Master language grammar
- Large corpus of text data
  - Books, articles, and websites
  - Web scraping
  - Vocabulary size
    - Number of tokens
    - 15.6 Trillion tokens
  - Clean dataset
    - 1% – 3% original tokens

# Instruction Fine Tuning

- Supervised learning
- Specific down stream task
- Instructions and Responses
  - Question and Answer
  - Text and Summary
  - English and Hindi sentences
  - Text and Sentiment
- Generate accurate responses
- Generate specific responses
- Specific domain knowledge
  - Source code
  - Medical documents
  - Legal documents
  - Financial documents

# Reinforcement Learning from Human Feedback

- Align model with human values
- Reinforcement learning
- 3H – Helpful, Honest, Harmless
- Helpful answer
- Honest answer
- Harmless answer
- Responsible AI
- Agent
  - Instruct LLM model
- Environment
  - LLM context
- Objective
  - Generate aligned text
- Reward model
  - Supervised learning

# Retrieval Augmented Generation

- Retrieval
  - Retrieve relevant information from source documents
- Augmented
  - Augment input query with retrieved relevant information
- Generation
  - Generate response using augmented input and LLM



# Retrieval Augmented Generation

- Improve accuracy
  - Closed book test v/s Open book text with Index page
- Increase transparency
  - Source references – Retrieved documents
- Reduce hallucination
  - Augmentation – Retrieval + Generation
- Up to date information
  - Ingest documents – No expensive model training



# Sahaayak – Retrieval Augmented Generation

Document  
Sources



Load  
Sources

XXXXXX XXXXXX  
XXXXXX XXXXXX  
  
XXXXXX XXXXXX  
XXXXXX XXXXXX  
  
XXXXXX XXXXXX  
XXXXXX XXXXXX  
  
XXXXXX XXXXXX  
XXXXXX XXXXXX

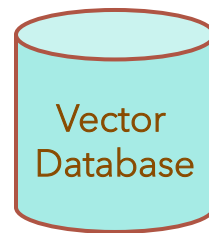
Transform  
Sources

XXXXXX XXXXXX  
XXXXXX XXXXXX  
  
XXXXXX XXXXXX  
XXXXXX XXXXXX  
  
XXXXXX XXXXXX  
XXXXXX XXXXXX  
  
XXXXXX XXXXXX  
XXXXXX XXXXXX

Compute  
Embeddings

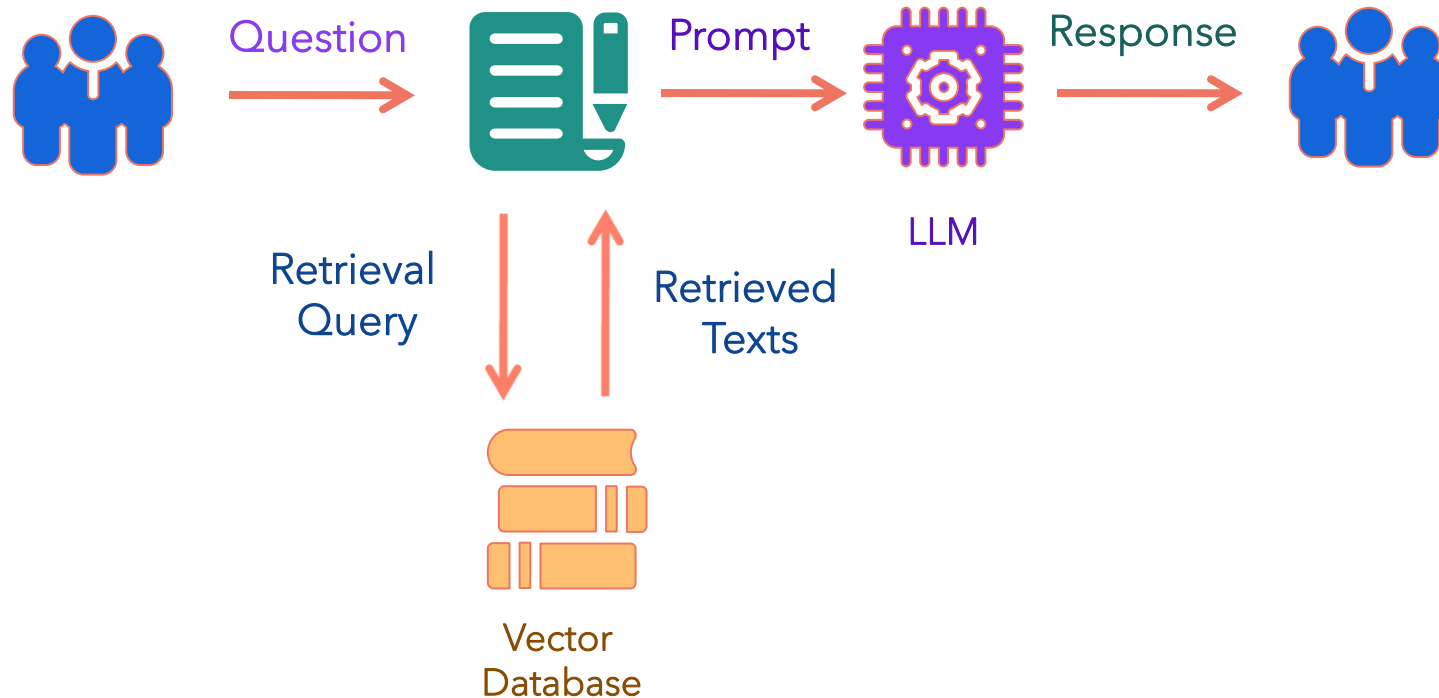
{0.25, 0.12,  
0.56, 0.78}  
  
{0.46, 0.23,  
0.79, 0.68}  
  
{0.78, 0.43,  
0.34, 0.84}  
  
{0.68, 0.24,  
0.98, 0.01}

Store





# Sahaayak – Retrieval Augmented Generation



\* Sahaayak – Gaurika Gupta, Mohini Laghate & Pritam Shete

# Sahaayak – BARC Documents

[INST]

<<SYS>> You are a helpful, respectful, and honest assistant. ... If you don't know the answer to a question, please don't share false information. <</SYS>>

CONTEXT:

Magnesium -18's Unique Decay Process: ...

provides not only new academic insights but also practical solutions ...

Collaboration Identifies Rare Nuclear Decay in Long - Lived ...

Question: Describe decay process of Magnesium-18.

[/INST]

Sure! Based on the provided context, Magnesium-18 undergoes a unique multi-proton decay process, where it sequentially emits two proton pairs. ...

This process differs from traditional radioactive decay and highlights a complex phenomenon in the quantum world. ...

The study, published in the journal Nuclear Science and Techniques, provides ...

To summarize, Magnesium-18 decays through a chain of events, starting with the emission of two proton pairs, followed by the decay of Neon -16, ...



# Sahaayak – BARC Documents

## Q1. What is PWR?

Based on the context text provided, PWR stands for Pressurized Water Reactor.

## Q2. What is PHWR?

Based on the context text provided, PHWR stands for Pressurized Heavy Water Reactor.

## Q3. What is AHWR?

Based on the context text provided, AHWR stands for Advanced Heavy Water Reactor.

## Q4. What is MSBR?

Based on the context provided, MSBR stands for Molten Salt Breeder Reactor.



# Pragya – Machine Learning Cluster

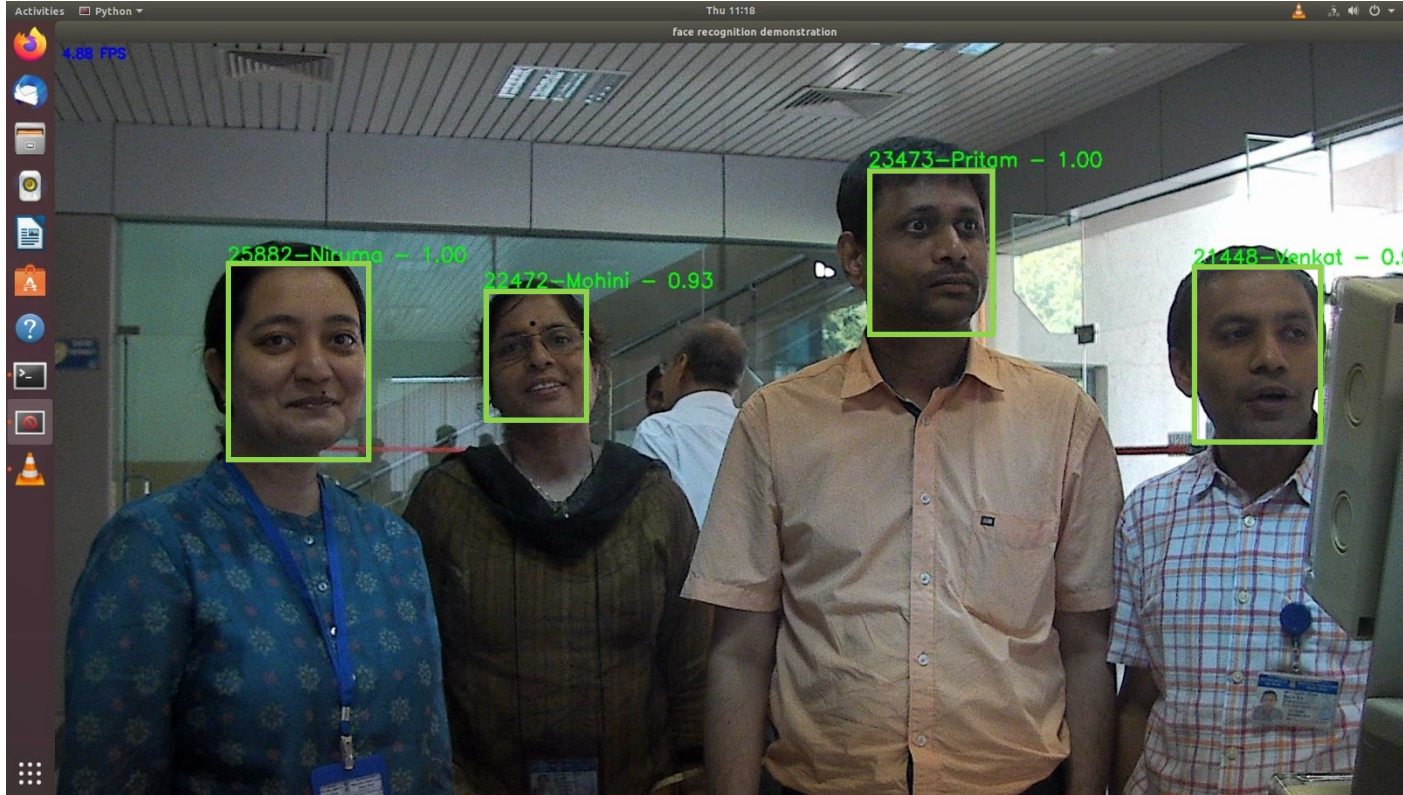
- 12 workstations
- Each workstation
  - Dual AMD EPYC Processors – 48 cores
  - 4 NVIDIA A100 GPUs
  - 6400 GB NVMe SSD storage
  - 1TB Memory
- NVIDIA A100 GPU
  - 6912 CUDA cores
  - 80 GB memory

# Falcon 180B On Pragya

- Falcon 180B Chat GPTQ
  - 180 Billion parameters
  - Instruction fine-tuned model
- Model compression
  - 180B parameters - 16 bit float - 360 GB memory
  - 180B parameters - 4 bit precision - ~94 GB memory
- Model parallelization
  - Divide model among multiple GPUs
  - ~24 GB - 4 A100 GPUs - ~94 GB / 4 GPUs

\* [TheBloke/Falcon-180B-Chat-GPTQ](#)

# Pehchaan – AI based Face Recognition System



# Talaash – Multimodal Query and Retrieval

- Query & retrieval system
  - Natural language
  - Input image
- Search & track users
  - User attire
  - Facial features
- Vision Language Models
  - Text & image embeddings
  - FashionVLM
- Vector database
  - Efficient storage
  - Retrieval of images





# Fashion Captioning Dataset (FACAD)

- Largest dataset of fashion items
  - 130K fashion item captions
  - 993K images
  - 990 attributes
  - 78 different categories
- Rich captions
- Expressive vocabulary
- Different age groups
- Different seasons



colourful bloom and minimalist style define an everyday backpack with a convenient exterior pouch interior slip pocket and a signature logo patch at the front



flower rendered in monochromatic metallic jacquard illustrate this bateau neck dress finished with a pleat flared skirt



a notched top line accentuates the modern drama of this lofty ankle strap sandal



sparkling diamond accent a delicate pendant cast in 18 karat white gold suspended from a lovely chain necklace

# FashionVLM

- BLIP-2
  - Bootstrapping Language Image Pre-training
- Image encoder
  - Vision Transformer (ViT)
- Text encoder
  - Large Language Model (LLM)
- Connect Image and Text encoders
  - Querying Transformer (Q-Former)



# FashionVLM – Evaluation on FACAD

| Models                 | Evaluation Metrics (%) |                |               |               |
|------------------------|------------------------|----------------|---------------|---------------|
|                        | BLEU-4                 | CIDEr          | ROUGE-L       | METEOR        |
| Tang (2023)            | 10.0                   | 81.8           | <u>23.0</u>   | <u>11.9</u>   |
| Moratelli (2023)       | <u>10.6</u>            | <u>84.5</u>    | 22.4          | 11.6          |
| OPT-6.7<br>Stage One   | 12.331                 | 101.205        | 26.746        | 14.279        |
| OPT-6.7<br>Stage Two   | 13.409                 | 111.396        | 27.756        | 14.946        |
| OPT-6.7<br>Stage Three | <u>14.881</u>          | <u>123.515</u> | <u>28.667</u> | <u>15.419</u> |

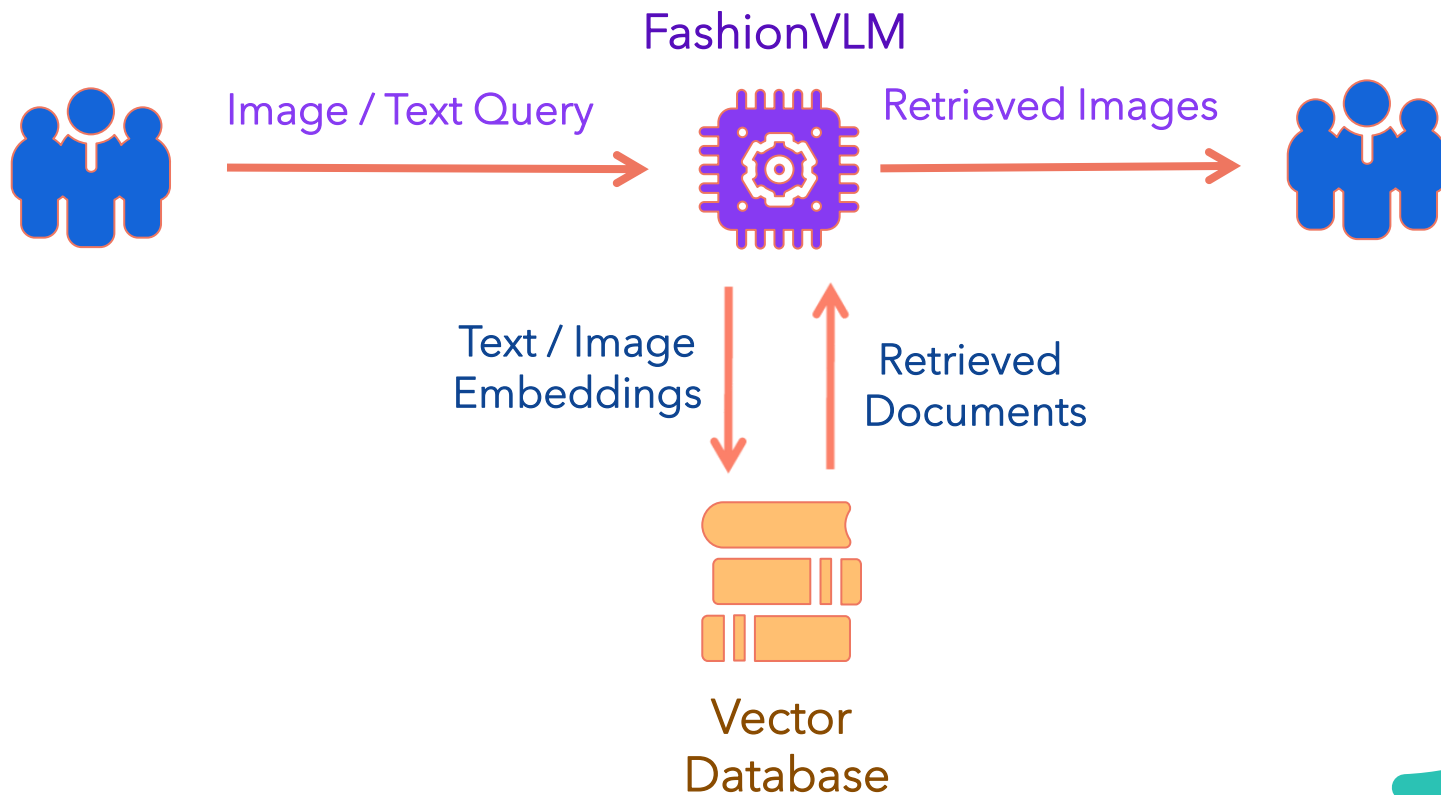
\* FashionVLM (2024) – Gaurika Gupta & Pritam Shete

# Talaash – Image Registration

- Input images
- Load images
- Transform images
- Compute image embeddings
- Store embeddings in vector database



# Talaash – Multimodal Search and Retrieval



# Chain-of-Thought Prompting

- Break down complex task into series of steps
- Focus on one step of task at a time
- Include responses to intermediate steps leading to final answer
- Improve accuracy of LLM's responses
- Improve relevance of LLM's responses
- Reduce hallucination



# Chain-of-Thought Prompting

Let us solve this step-by-step.

What information do we need to find?

What operations might help us to solve this?

Explain your reasoning and prove the answer.

If there are 10 birds on a branch and 3 birds fly away. How many birds are remaining?

Let's break down the problem step-by-step.

What information do we need to find? We need to find the number of birds remaining on the branch after 3 birds fly away.

What operations might help us to solve this? To solve this problem, we can use a simple subtraction operation. We know the initial number of birds (10) and the number of birds that fly away (3). By subtracting the number of birds that fly away from the initial number, we can find the number of birds remaining.

Why subtraction? Subtraction is the correct operation because we are removing a certain number of birds (3) from the initial total (10). By subtracting 3 from 10, we are essentially "taking away" 3 birds from the branch, leaving us with the remaining number of birds.

Let's calculate!  $10$  (initial number of birds)  $- 3$  (number of birds that fly away)  $= 7$

Therefore, there are 7 birds remaining on the branch.

# Code Generation

Write Python function to find a factorial of an input number.

```
def factorial(n):  
    if n < 0:  
        raise ValueError("Factorial is not  
defined for negative numbers")  
    elif n == 0 or n == 1:  
        return 1  
    else:  
        return n * factorial(n-1)
```





# Code Generation

Correct following Python code.

```
def factorial(n):  
    if n == 0 or n == 1:  
        return 1  
    else:  
        result = 1  
        for i in range(2, n + 1):  
            result *= i  
        return result
```

```
def factorial(n):  
    if n < 0:  
        raise ValueError("Factorial is not  
defined for negative numbers")  
    elif n == 0 or n == 1:  
        return 1  
    else:  
        return n * factorial(n-1)
```

# Creative Writing

Write a story on atom for first standard student.

## The Adventures of Atom

Once upon a time, in a tiny world, there was a tiny creature named Atom. Atom was so small that you needed a special tool to see him.

...

One day, Atom decided to go on an adventure. He packed a small bag and said goodbye to his family.

...

## The End

This story is designed to be simple and easy to understand for a first standard student.

\*Meta-Llama-3.1-8B-Instruct

# Education

In six sentences, explain proton, neutron and electron to first standard student.

Here's an explanation of protons, neutrons, and electrons that a first standard student can understand:

## The Tiny Building Blocks of Atoms

You know how we're made up of tiny things called atoms? Well, atoms are made up of even smaller things called protons, neutrons, and electrons. Protons are like the strong and brave soldiers that live in the center of the atom. Neutrons are like the quiet and calm friends who live with the protons in the center. Electrons are like the fast and energetic runners who zoom around the outside of the atom. All three of them work together to make the atom strong and stable!

# Deploying Large Language Models

| GPU             | CUDA Cores | Memory | FP32 FLOPS   | Power | Cost     |
|-----------------|------------|--------|--------------|-------|----------|
| NVIDIA H100     | 14,592     | 80 GB  | 67 TFLOPS    | 700W  | \$30,000 |
| NVIDIA A100     | 6,912      | 80 GB  | 19.5 TFLOPS  | 500W  | \$15,000 |
| NVIDIA L40      | 18,176     | 48 GB  | 90.5 TFLOPS  | 300W  | \$10,000 |
| NVIDIA RTX 4090 | 16,384     | 24 GB  | 82.58 TFLOPS | 450W  | \$3,000  |
| NVIDIA RTX 4080 | 10,240     | 16 GB  | 48.74 TFLOPS | 320W  | \$1,000  |

# Conclusions

- Large Language Models
- Prompt Engineering
- In Context Learning
- Training Large Language Models
- Large Language Model Applications
- Deploying Large Language Models





Thank you

Pritam Prakash Shete

Telephone - 022-25595212

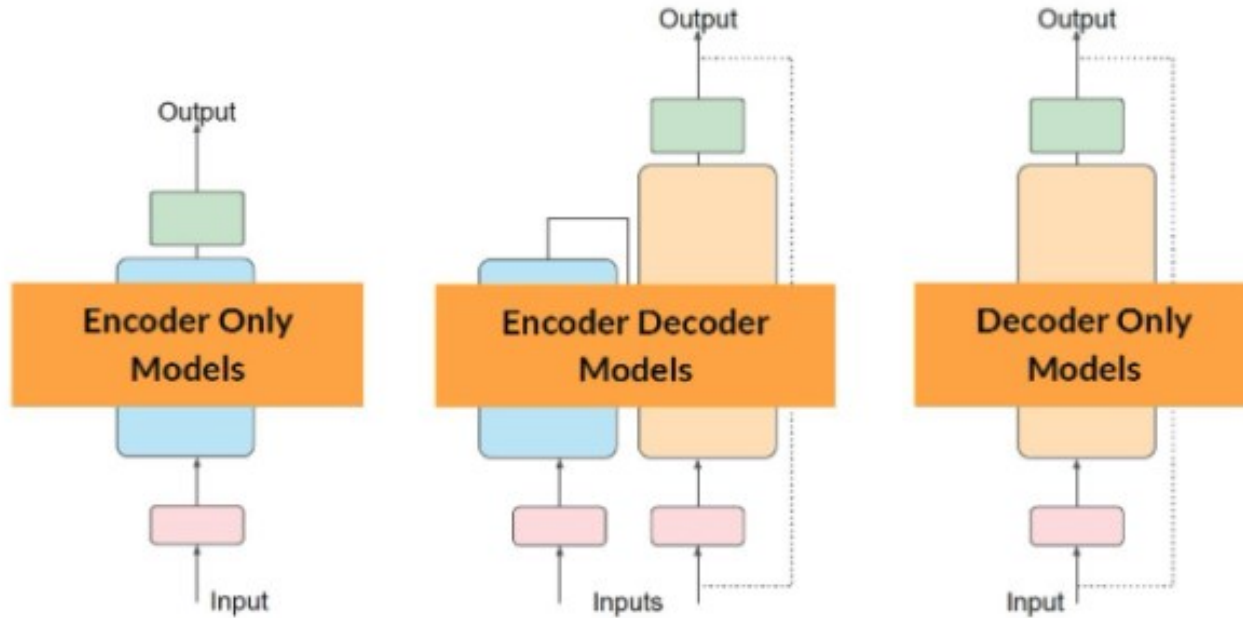
Email - [ppshete@barc.gov.in](mailto:ppshete@barc.gov.in)

# Word v/s Token

- Token – Word or sub-word
- LLM inference cost
  - Number of tokens
  - Input tokens
    - Less cost
  - Output tokens
    - More cost
- Words with single token
  - Atomic or Primitive tokens
  - yes, no, run, red, blue, love
- Words with two tokens
  - Compound words
  - unhappy, dislike, replay, sunset
- 300 Words – 400 Tokens



# Large Language Model Architectures



\* Source: Vijay Chaudhary



# LLM – Model Pre-training

- Encoder only LLM
  - Auto-encoding models
  - Masked Language Modeling
  - Reconstruct text
  - De-noising objective
  - Bidirectional context
- Applications
  - Sentiment analysis
  - Word classification
- Examples
  - BERT model
  - ROBERTA model

# LLM – Model Pre-training

- Decoder only LLM
  - Autoregressive models
  - No encoder model
  - Causal Language Modeling
  - Predict next token
  - Unidirectional context
  - Statistical representation of language
- Applications
  - Text generation
  - Zero-shot inference
- Examples
  - OpenAI GPT
  - Meta Llama

# LLM – Model Pre-training

- Encoder-Decoder LLM
  - Encoder and Decoder
  - Sequence-to-sequence models
  - Span corruption - T5 model
    - Mask random input tokens
    - Reconstruct masked input tokens
  - Bidirectional context
- Applications
  - Machine translation
  - Text summarization
  - Question & Answering
- Examples
  - Text-to-Text Transfer Transformer - T5 model

# Mixture of Experts – (MoE)

- Multiple specialized models
  - Work together
- Gating model
  - Select best expert
- Examples
  - Doctor for medical issues
  - Mechanic for car problems
  - Chef for cooking
- Train Large Language Model
  - Computational resources
- Mixture of Experts
  - Break down large model
  - Smaller specialized models



# LLM – Computational Challenges

- LLM inference

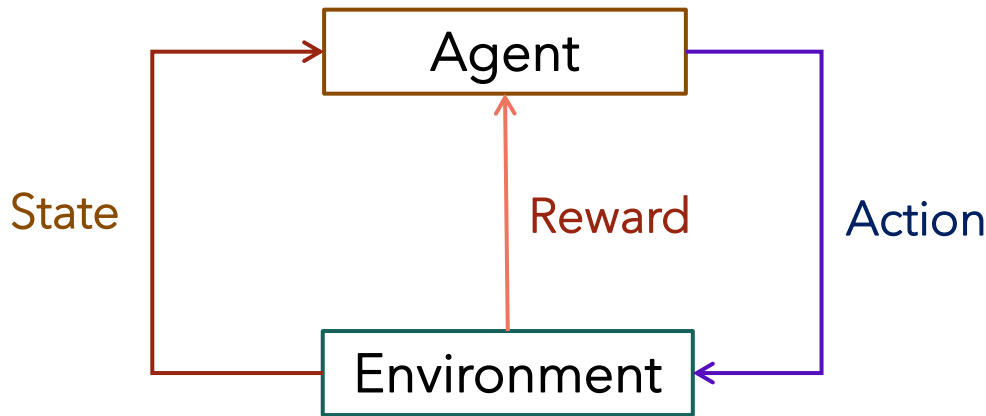
- 1 parameter – float 32 – 4 bytes
- 1B parameters –  $10^9$  parameters
- 1B parameters –  $4 \times 10^9$  bytes
- 1B parameters – 4 GB memory

- LLM training

- 1 parameter – float 32 – 4 bytes
- Model parameters – 4 bytes
- Optimizer – Two states – 8 bytes
- Gradients – 4 bytes
- Activations – 4 bytes
- Temporary variables – 4 bytes
- 20 times number of parameters

# Reinforcement Learning

- Agent
- Environment
- Action
- State
- Reward or Penalty
- Maximize reward



# Reinforcement Learning

- Playout or Rollout
- Exploration
- Exploitation
- Reward hacking



# Retrieval Augmented Generation – Evaluation

- Ground Truth (GT) by Human
- Character based evaluation
  - Edit distance
- Word based evaluation
  - WER, BLEU
- Embedding based evaluation
  - BERT score, Mover score
- Ground Truth (GT) by LLM
- Mathematical Framework
  - RAGAS framework
- Experimental based Framework
  - Number of tasks and datasets
  - Number of aspects
  - GPT score

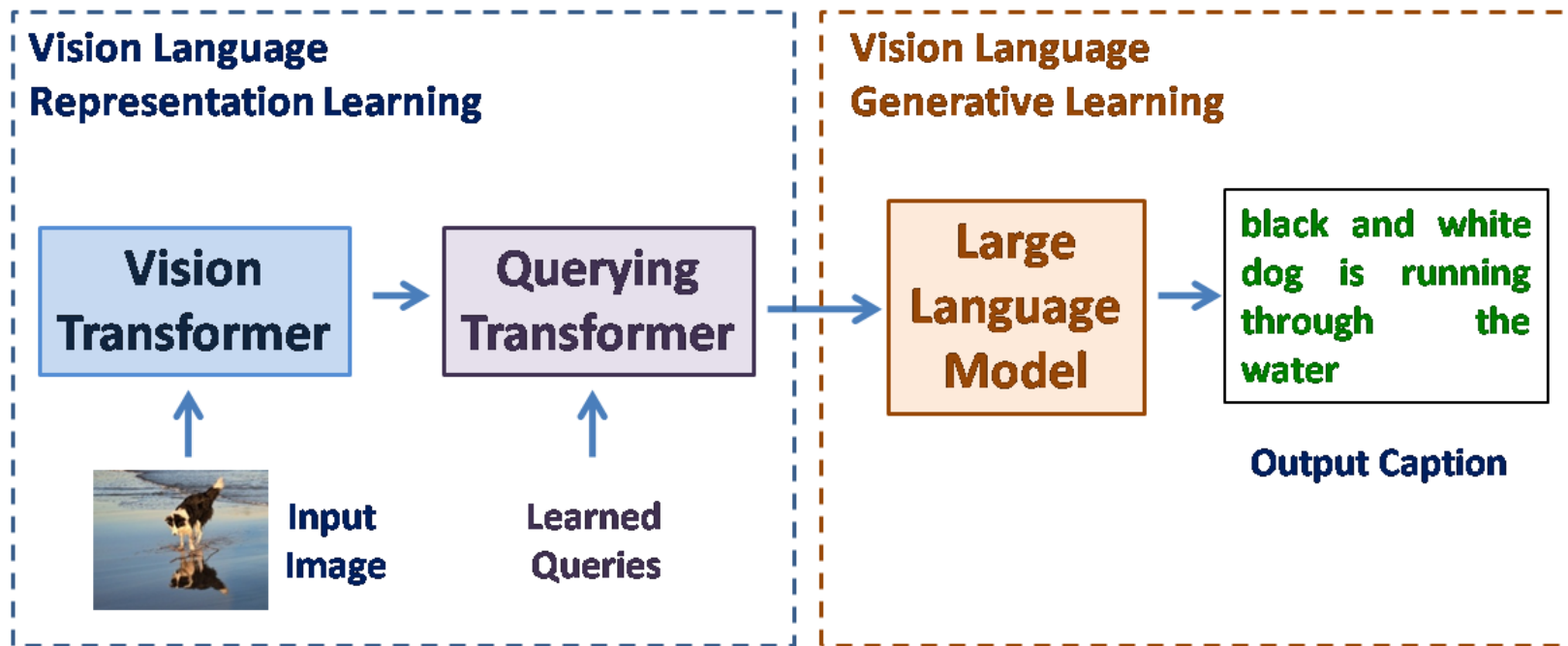




# Image Captioning Models

| Model    | Encoder      | Decoder       | Q-Former | Number of Parameters |
|----------|--------------|---------------|----------|----------------------|
| CNN-LSTM | CNN (VGG)    | LSTM          | No       | 138 Million          |
| CLIP     | ResNet / ViT | Transformer   | No       | 33 Million           |
| BLIP     | ViT          | Transformer   | No       | 583 Million          |
| BLIP-2   | ViT          | OPT / FLAN T5 | Yes      | 188 Million          |

# BLIP2 Architecture



# FashionVLM – Model Centric AI Development

- Stage One
  - BLIP-2 pre-trained models
  - Freeze ViT & Freeze LLM
  - 224x224 image size & fp16 precision
- Stage Two
  - MS COCO dataset fine-tuned models
  - Fine-tune ViT & Freeze LLM
  - 364x364 image size & fp32 precision
- Stage Three
  - Stage One models
  - Fine-tune ViT & Freeze LLM
  - 364x364 image size and fp32 precision

# FashionVLM – Evaluation on FACAD

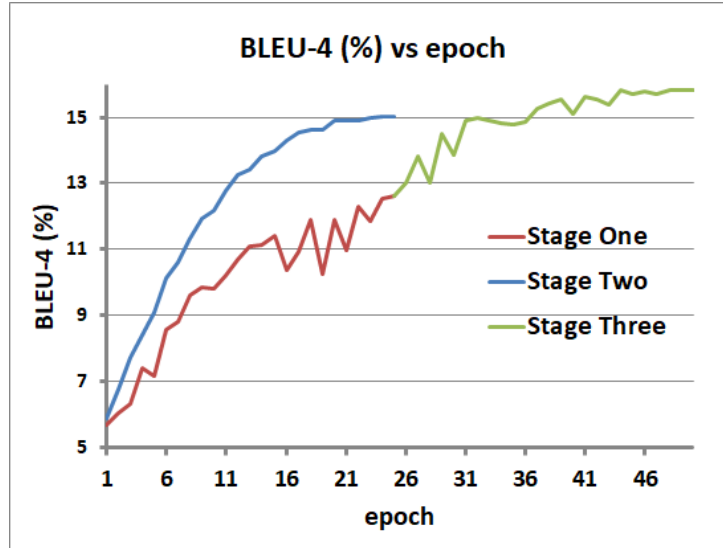


Fig. 1 - BLEU-4 validation score for OPT-2.7 based FashionVLMs

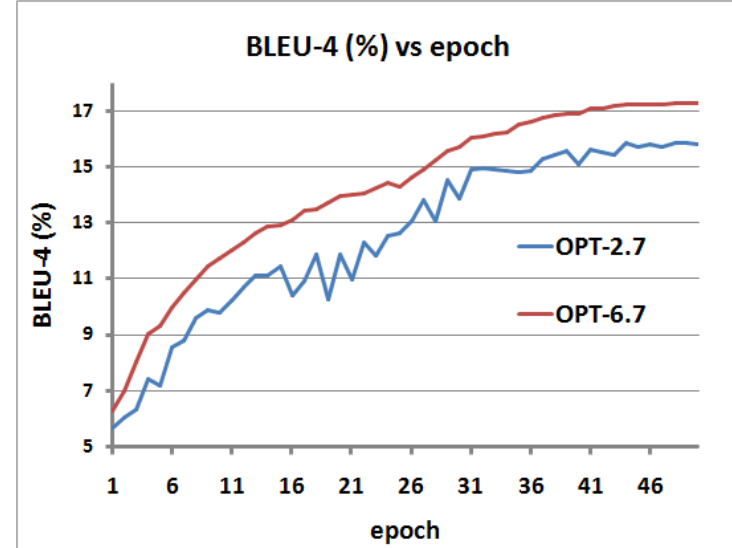


Fig. 2 - BLEU-4 validation score for OPT-2.7 and OPT-6.7 based composite FashionVLMs



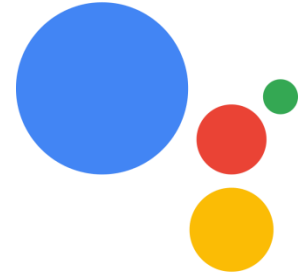
# Code Generation

- Mistral AI
  - Mistral 7B and Mixtral 8X7B
- Meta
  - CodeLlama, CodeLlama-Python, CodeLlama-Instruct
- Phind
  - Phind-CodeLlama-34B, Phind-CodeLlama-34B-Python
- BigCode Project
  - StarCoder, StarCoder2



# Virtual Assistants

- Simulate human like conversations
- Answer questions
- Provide information
- Complete simple tasks



# Medical Diagnosis

- Symptom analysis
  - Disease identification
  - Treatment recommendations
  - Medical literature analysis
  - Medical question answering
  - Medical imaging analysis
- Google Med-PaLM 2
    - Medical question answering
  - Google Med-Gemini
    - Medical question answering
  - BiomedGPT
    - Vision language model
    - Visual question answering

# Knowledge Distillation

- Knowledge transfer
- Large pre-trained model
- Set of models
- Teacher model
- Ensemble of models
- Single smaller model
- Student model
- Model compression
- Small student model
- Learn to emulate
- Large teacher model
- Leverage teacher knowledge
- Emulate thought process
- Obtain similar / higher accuracy





