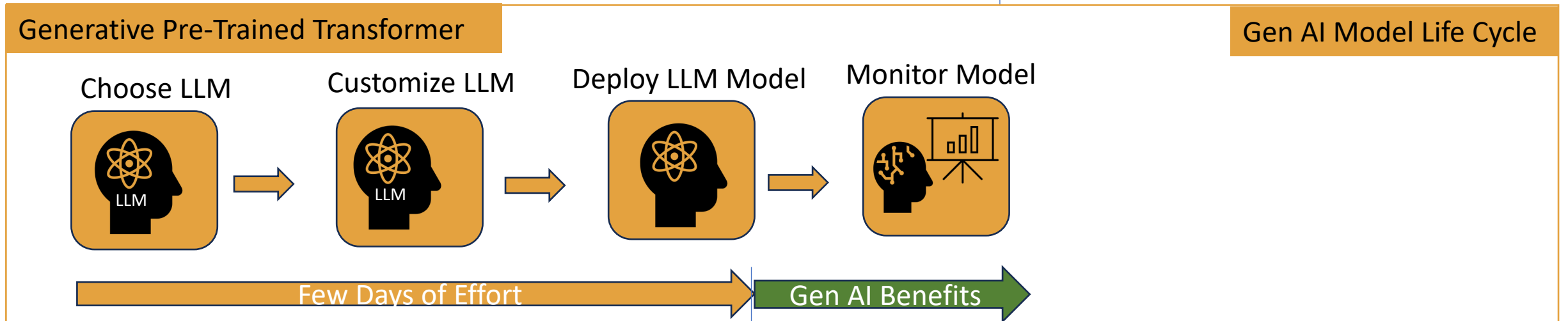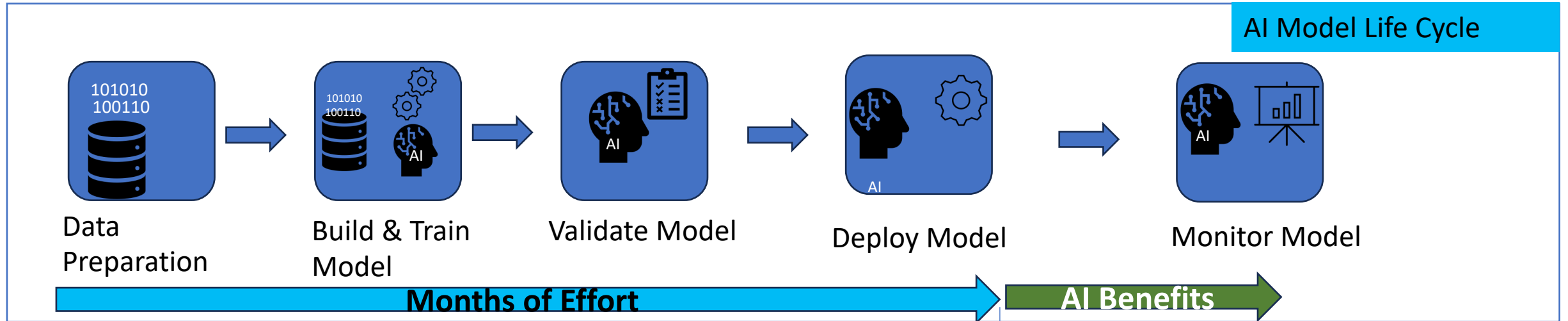# Generative AI foundation to Real World use cases

Prasad Venkatachar

Sr Director – Product & Solutions @ Pliops
IEEE Sr member & BCS Fellow

AI VS Gen AI Life Cycle

AI Model Life Cycle

Data Preparation → Build & Train Model → Validate Model → Deploy Model → Monitor Model

Months of Effort → AI Benefits

Generative Pre-Trained Transformer

Gen AI Model Life Cycle

Choose LLM → Customize LLM → Deploy LLM Model → Monitor Model

Few Days of Effort → Gen AI Benefits

Key Pillars to Generative AI

Enterprise Use Cases & Business Outcomes

Gen AI Stack

Hardware + Software

Security & Regulatory Compliance

Finops for Gen AI
Cost of Development & Serving

Cost of Scaling

New Trends & Upto-Date
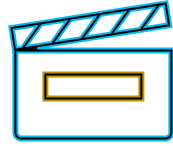
Gen AI Opportunity Everywhere

Industry Verticals

Content & Media

Horizontal Functions

AI High Performers Response

Banking & FSI

HR & Finance

Retail

	Reduce Costs in Core Business	Create New Business and/or sources of Revenue	Increase Revenue from Core Business	Increase value of Offerings by Integrating AI-Based Features or Insights
Respondents at AI High Perfomers	19%	23%	27%	30%

Customer Support

Gen AI Stack

SECURITY

FULL-STACK PLATFORMS

Google Vertex AI

Azure AI Studio

AWS Bed Rock

Content Generation

Summarization

Classification

Translation

Gen AI Application Use-Cases

Monitoring & Observability: Vertex ML-ops, DataRobot, Arize, PromptLayer

Vector Databases: AlloyDB Omni, Oracle23ai, SQL Server, MongoDB, PostgreSQL, Pinecone,CosmosDB

Orchestration Frameworks: Langchain, LamaIndex,HayStack

Model Providers: Google Model Garden, Hugging Face, Anthropic, Mistral AI, together AI, Fireworks.AI

Development to Deployment : Cloud Infrastructure/Hybrid/On-Premise

GOVERNANCE

Gen AI : Adoption Maturity

Generic Problems

Prompt Engineering

Prompt Templates & LLM APIs

Text

100110
100110

Pre-training

Image

Foundation Model

LLM

Specific Problems

RAG (Retrieval Augmented Generation)

Enterprise Dataset in Vector Databases + LLMs

Video

Specific Problems

Fine Tuning Model

Fine Tune with Selective Enterprise Dataset

Prompt Engineering – Use cases

LLM

- Sentiment Analysis
- Document Summarization
- Call center Chatbot
- Code Generation

UI|SDK|API

Prompt Request

Define Task

Multi Modal LLM

- Generate Images - Google Imagen, Diffusion Models
- Generate text captions from Images
- Generate text from speech
- Generate speech from Text
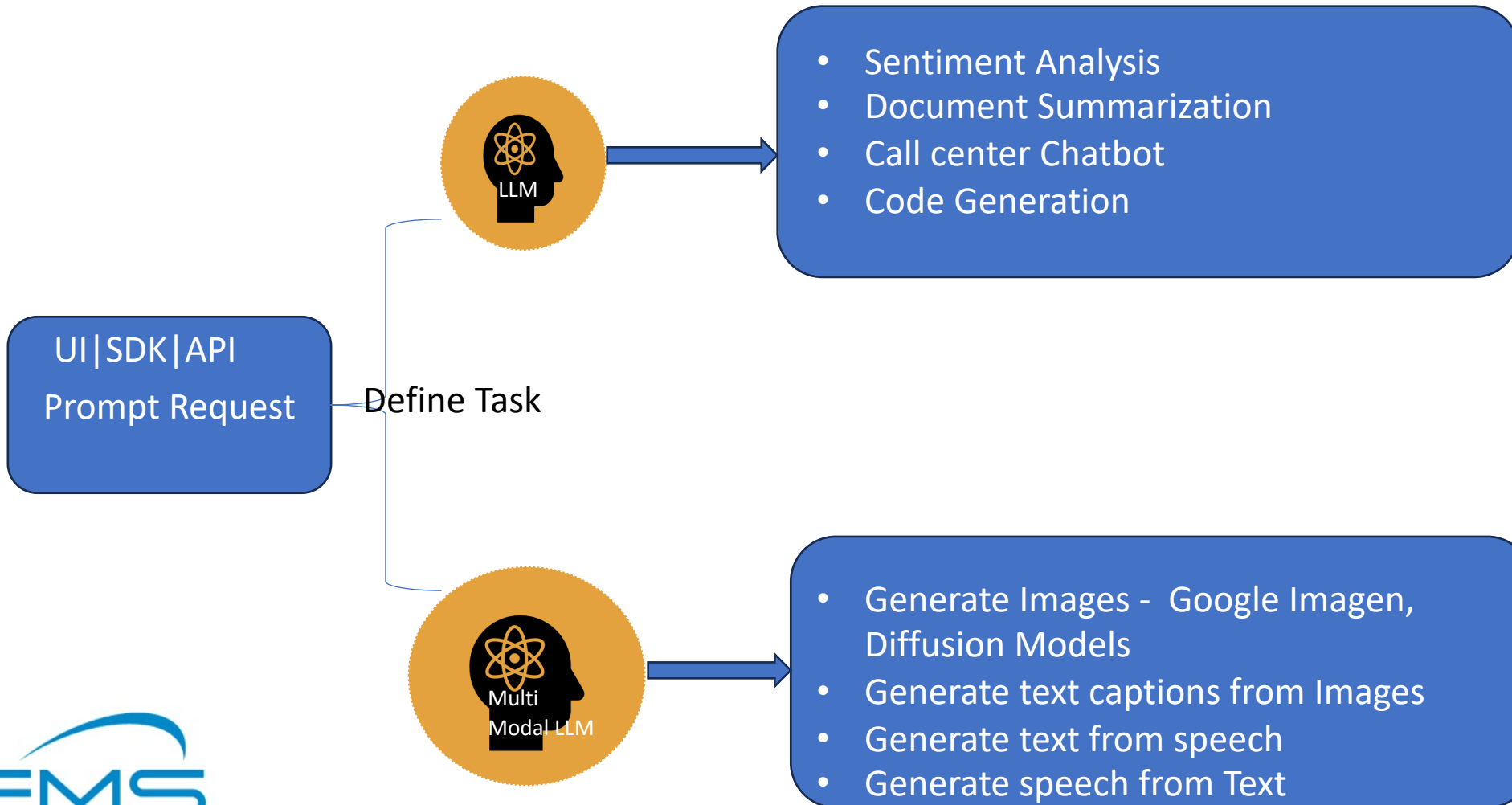
Electric Bike Chatbot Example

```
ElectricBikes = [
{
"brand": "VELECTREC",
"type": "Mountain",
"features": {
"motor" :"1000w",
"Max Range(Miles)":"40-50",
"Max Speed(MPH)":"25",
"Number of speeds: "7",
"brakes": "Front & Rear Disc
Brakes",
"Charging Time(HRS)": "5-6",
"frame Type": "Aluminum",
"suspension": "Dual",
"tires": "27.5-inch all-terrain",
},
"price": 699,
"Sold by": "velectric inc"
},
```

E-Bike Conversation Assistant:
List Bikes: Provide list of latest models bikes matching customer query:Price,Range, Charging Time etc
addToOrder: If the Bike on the list do addToOrder, then summarizeOrder, then confirmOrder.
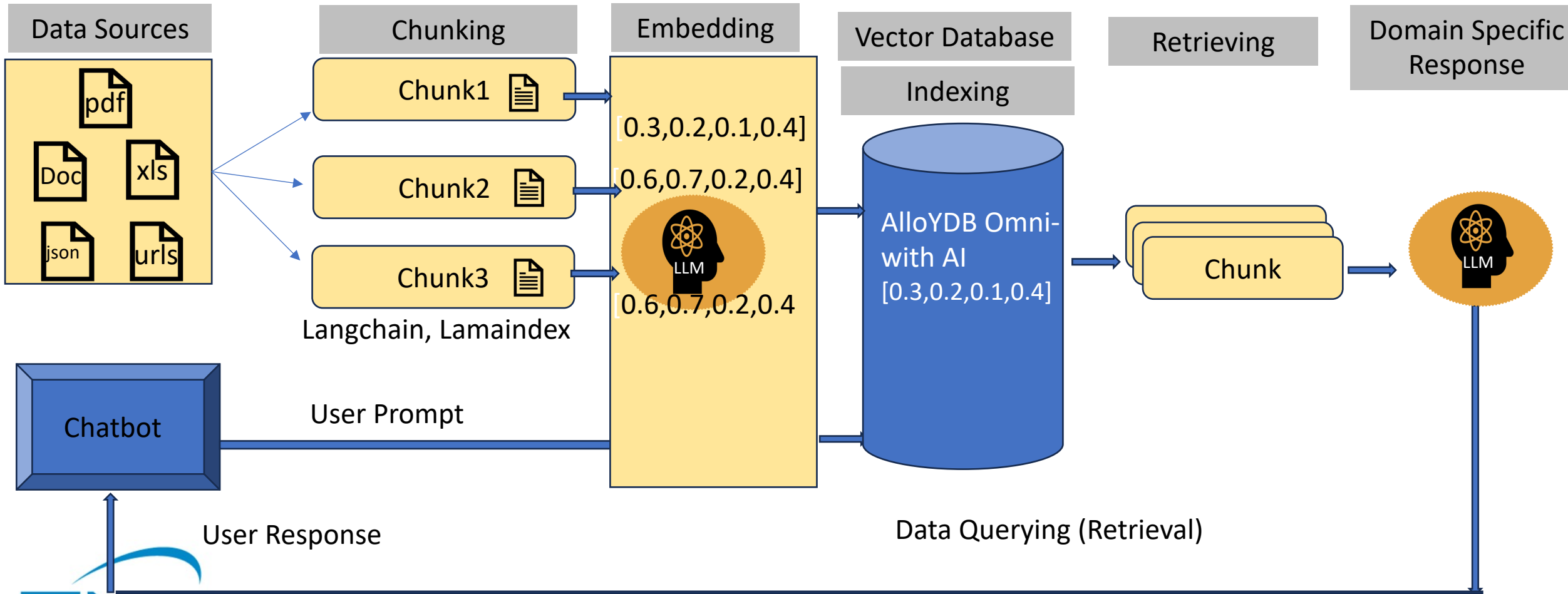**changeItem:** Replace bike with another type: Range, Price
summarizeOrder: If the customer has added to the order, list each menu item and modifier added to the order. If there has been nothing ordered, redirect.
confirmOrder: Ask the customer to confirm the order details are correct.
finishOrder: Tell the user the order has been sent to the Delivery

Source: Amazon Electric Bike URL

AlloyDB AI – Vector Database

Data Sources

pdf
Doc xls
json urls

Chunking

Chunk1
Chunk2
Chunk3

Langchain, Lamaindex

Embedding

[0.3,0.2,0.1,0.4]

[0.6,0.7,0.2,0.4]

LLM

[0.6,0.7,0.2,0.4

Vector Database

Indexing

AlloYDB Omni-with AI
[0.3,0.2,0.1,0.4]

Retrieving

Chunk

Domain Specific Response

LLM

Chatbot

User Prompt

User Response

Data Querying (Retrieval)

RAG Architecture

LLMs: Buy, Build or Best of Breed

$ $ $ $ $

Multiple GPU Cost X Months of Tuning Duration

Pre-Train
(Build from Scratch)

$ $ $

LLMs * GPU Cost * Tuning Duration

Fine Tune Model

$ $

RAG (Retrieval Augmented Generation)

$

Prompt + Completion Tokens

Prompt Engineering

Inference Cost

RAG +
LLM Inference Cost

Fine Tuning Cost
+ Inference cost

Pre-Training Cost
+
Inference Cost

Hosting Cost : On-Premise or Cloud
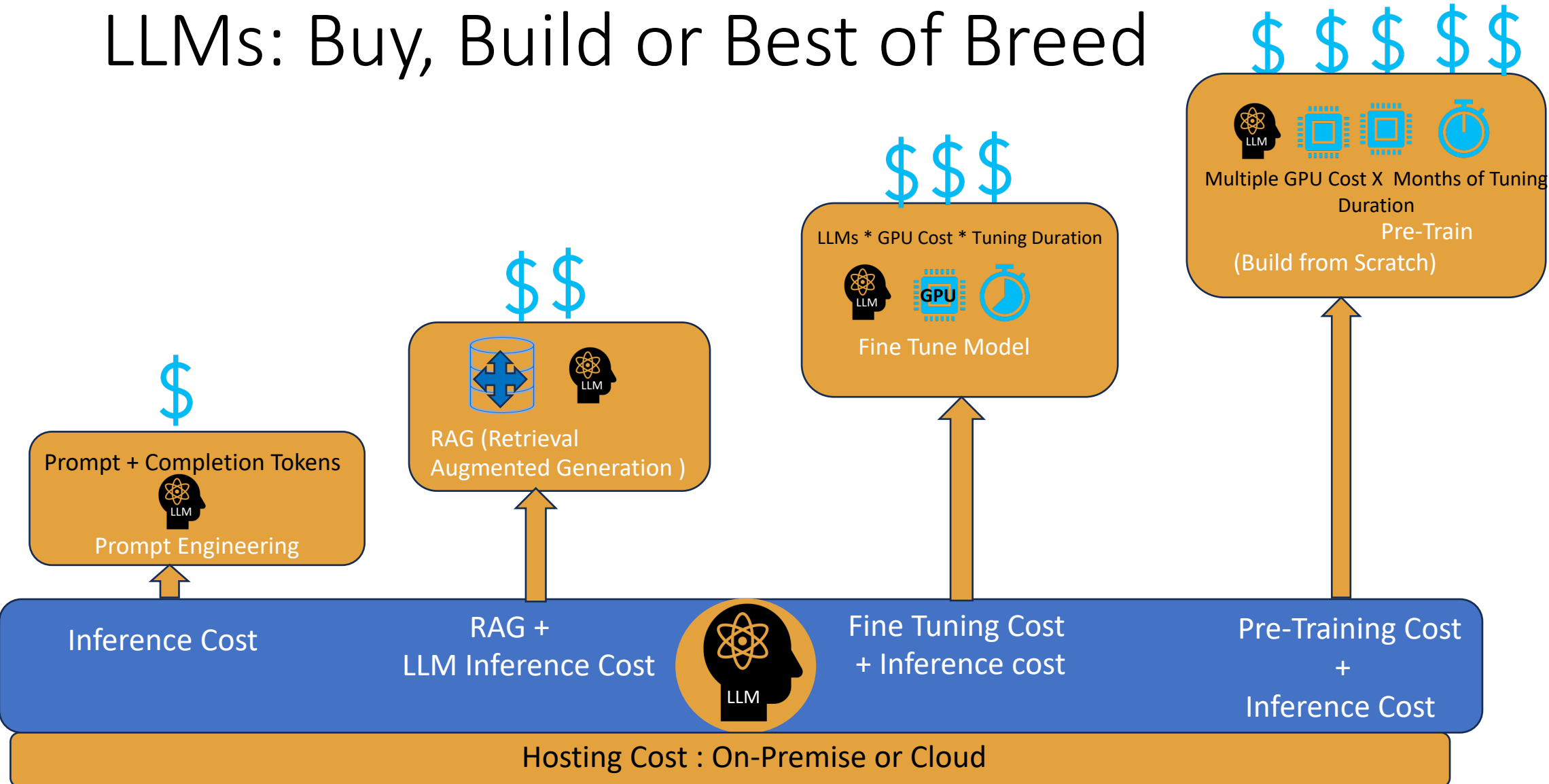
One Platform: OLTP/Analytics/Gen AI

E-Commerce Applications

Real Time Biz Analytics

Gen AI Chatbots

Transaction Database

Columnar Engine

Vector Database

AlloyDB Omni

AlloyDB

AccelDB
PLiOPS
EXTREME DATA PROCESSOR

Lenovo

XDP Data Accelerator + SSDs
SSD SSD

Lenovo Edge Server

XDP Data Accelerator + SSDs
SSD SSD SSD SSD

Lenovo Datacenter Server

Google Cloud
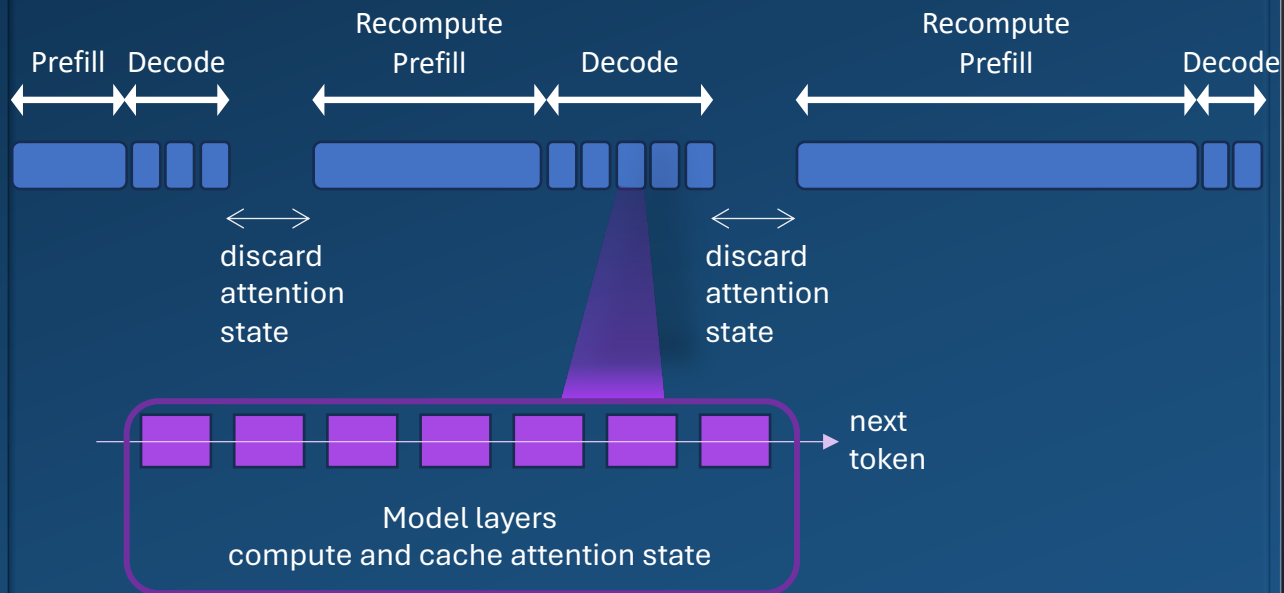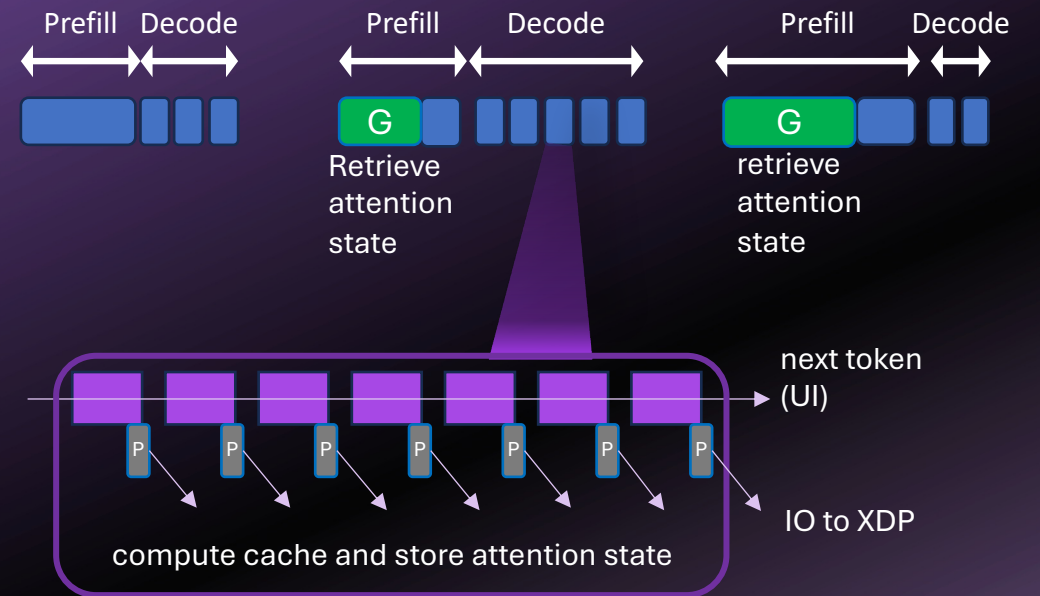
FMS

Long-Context Attention

conversation or multi-shot task agent prefills attention score based on expanding history

Today: *Compute-Based Prefill*

Prefill Decode

Recompute Prefill   Decode

Recompute Prefill   Decode

discard attention state

discard attention state

next token

Model layers
compute and cache attention state

Pliops LightningAI-Based Prefill

AccelKV

Prefill Decode   Prefill Decode   Prefill Decode

G Retrieve attention state

G retrieve attention state

next token (UI)

IO to XDP

compute cache and store attention state

P multiput to kv storage

G multiget from kv storage

Summary & Q/A

- Gen AI is rapidly adopted in Enterprises

- Identifying the use case and business outcomes is the key

- Building skillset for Gen AI adoption maturity is critical

- Enterprises has to focus on Fin-ops for Gen AI
  - Building the Gen AI LLM solution
  - Serving the Gen AI LLM Solution