Gen AI with Element

Surendra Panpaliya

PREREQUISITES

Participants should have:

Basic knowledge of **Python programming**

Familiarity with **APIs**, **JSON**, and **HTTP requests**

General understanding of **Machine Learning** and **NLP concepts**

PREREQUISITES

- Awareness of **cloud platforms** (Azure or GCP preferred)
- Prior exposure to **Jupyter Notebooks** or **VS Code**
- Knowledge of **RESTful APIs**, **Docker**, and **Git**
- Some experience with **LLMs** or **prompt engineering**

LAB SETUP REQUIREMENTS

- **Python 3.10+** installed (preferably in a virtual environment)
- **JupyterLab** or **VS Code** with Python plugin
- Access to:
- **Azure OpenAI API key** (or)
- **Google GenAI credentials** (Vertex AI Studio, PaLM/Gemini API)

📌 LEARNING OUTCOMES

Explain key concepts in

Generative AI and

Transformer-based LLMs

Build and interact with

OpenAI/Gemini models using Python

📌 LEARNING OUTCOMES

Design **RAG pipelines** integrated with

LangChain + Milvus

Apply structured prompt

engineering strategies in LLM apps

📌 LEARNING OUTCOMES

Utilize Walmart's internal

LLM Gateway and **evaluation platforms**

Create simple **Agentic applications**

with planning, execution, and tools

📌 LEARNING OUTCOMES

Troubleshoot common LLM issues

such as hallucination or bias

Build and present a functional

Excel-based report builder GenAI app

Agenda

- DAY 1: GENAI FOUNDATION & ARCHITECTURE
- DAY 2: WALMART GENAI ECOSYSTEM
- DAY 3: APPLICATION DEVELOPMENT WITH GENAI
- DAY 4: HACKATHON & DEPLOYMENT

DAY 1: GENAI FOUNDATION & ARCHITECTURE

Objective

Build conceptual clarity and

understanding of Generative AI,

Large Language Models (LLMs),

Walmart’s GenAI technology stack.

Agenda

1. What is Generative AI?

2. Neural Generative Modeling

3. Large Language Models (LLMs)

4. Transformer Architecture & Attention Mechanism

Agenda

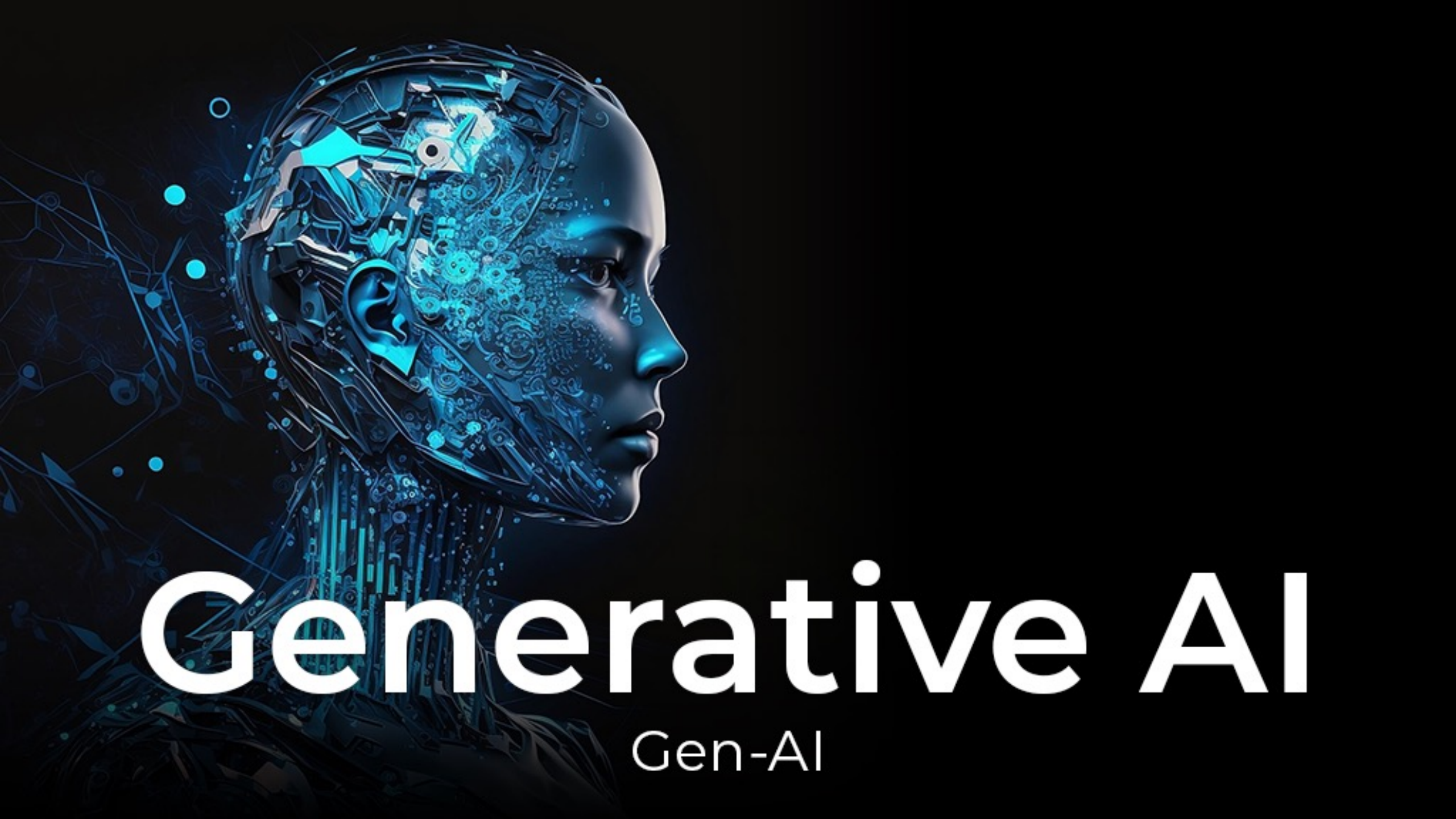
5. RLHF and DPO

6. Python Libraries Overview

7. Hands-On Session

What is Generative AI?

- Definition and significance in the AI landscape
- Evolution from traditional AI → Deep Learning → LLMs
- Demystifying the Blackbox:
- How LLMs generate human-like output


Generative AI
Gen-AI

What is Generative AI?

CLASS OF ARTIFICIAL INTELLIGENCE MODELS

CAN **GENERATE NEW CONTENT**

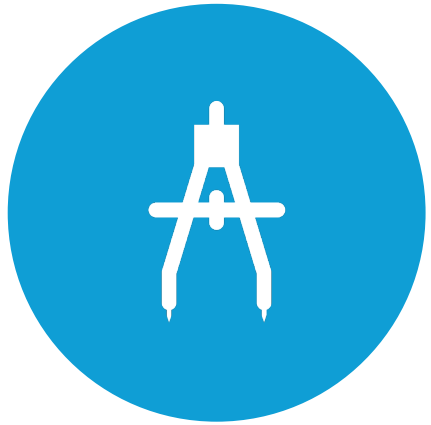
SUCH AS TEXT, IMAGES, AUDIO, VIDEO, OR CODE

BY LEARNING PATTERNS FROM EXISTING DATA.

What is Generative AI?

MODELS DON’T JUST ANALYZE DATA

THEY **CREATE** SOMETHING NEW



MEANINGFUL FROM IT.

What is Generative AI?

At the core of GenAI are **foundation models**

GPT (for text),

DALL·E or **Stable Diffusion** (for images),

MusicLM (for audio),

Codex (for code generation), and more.

What is Generative AI?

MODELS USE DEEP LEARNING,

ESPECIALLY **TRANSFORMERS**,

TO UNDERSTAND AND GENERATE CONTENT

THAT MIMICS HUMAN INTELLIGENCE.

What is Generative AI?

Like teaching a very smart machine

to **read millions of examples** and

then create **something similar but original**.

What is Generative AI?

https://www.youtube.com/watch?v=rwF-X5STYks

Why It Matters?

HELPS **AUTOMATE TASKS**

THAT REQUIRE CREATIVITY OR

COMMUNICATION.

Why It Matters?

SAVES TIME,

REDUCES MANUAL WORK

GIVES PERSONALIZED EXPERIENCES.

Why It Matters?

USED IN CHATBOTS,

SEARCH ENGINES,

PRODUCT RECOMMENDATIONS,

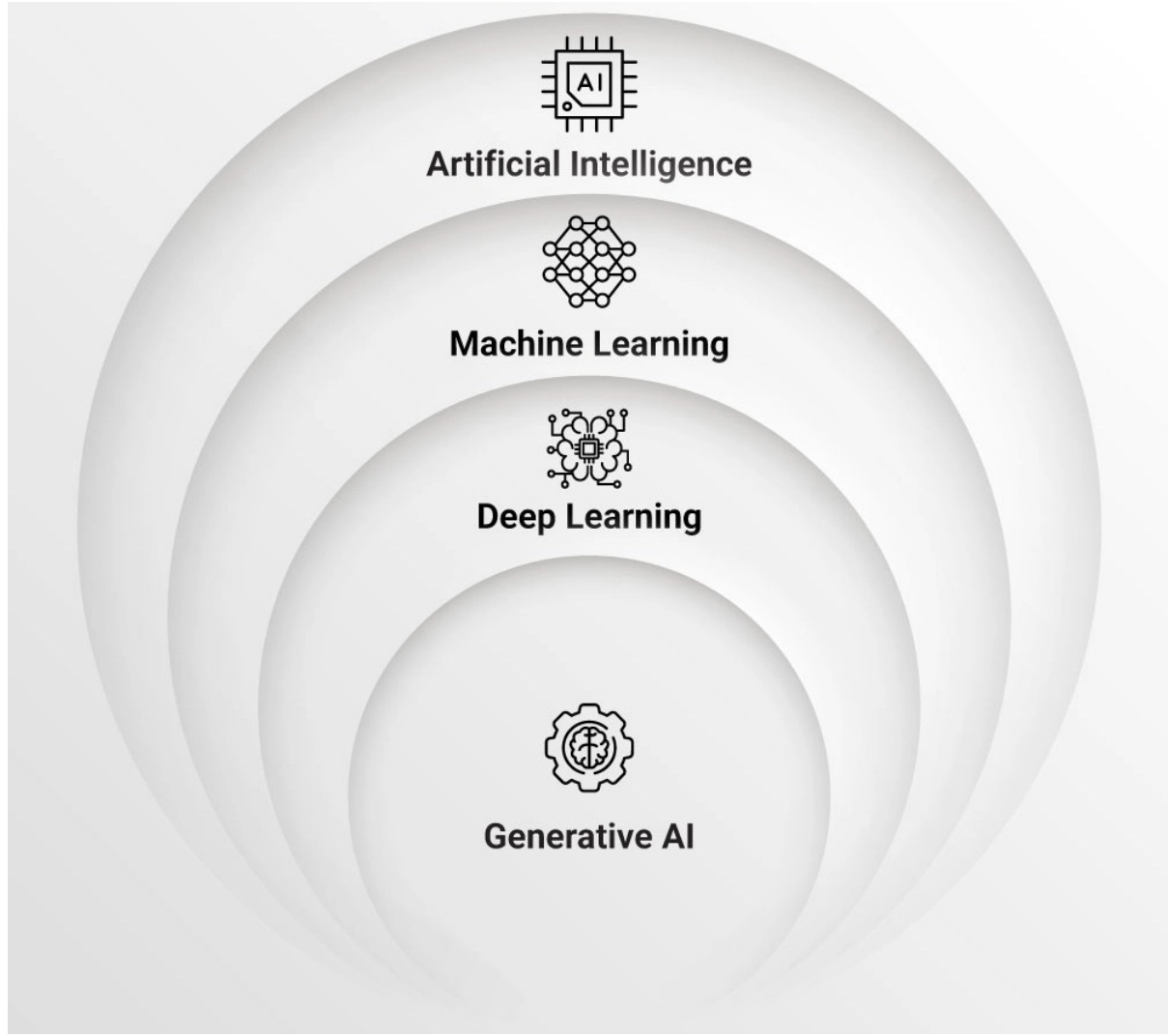
MARKETING COPY

Generative AI vs Traditional AI

Feature	Traditional AI	Generative AI
Goal	Make predictions or classifications	Generate new content: text, images, code, etc.
Input → Output	Input → Fixed output (e.g., yes/no, number)	Input → Creative output (e.g., paragraph, summary)
Training	Rule-based or statistical	Trained on vast internet-scale datasets
Example Tool	ML model to predict brake failure	LLM (e.g., GPT-4) to generate a troubleshooting guide

Evolution of AI → Deep Learning → LLMs

Stage	What It Means	Simple Analogy
Traditional AI	Rule-based systems	Like writing “if this, then that” instructions manually
Deep Learning	Learns patterns from lots of data	Like training a child to recognize cats vs dogs by showing many images
Large Language Models (LLMs)	Learns to understand and generate natural language	Like teaching an assistant to read millions of books and answer questions like a human


AI
Artificial Intelligence
Machine Learning
Deep Learning
Generative AI

Artificial Intelligence

Broad concept of machines

doing tasks that typically require

human intelligence like

understanding language,

recognizing images, decision making

Walmart Example

A basic AI system might decide:

“If order is delayed by 3 days,

send a compensation coupon.”

Logic = manually programmed rules

Machine Learning

ML IS A **SUBSET OF AI** WHERE

MACHINES **LEARN FROM DATA**

INSTEAD OF BEING EXPLICITLY PROGRAMMED.

How it works?

FEED IN DATA

TRAIN A MODEL

MAKE PREDICTIONS
ON NEW DATA

Types of ML

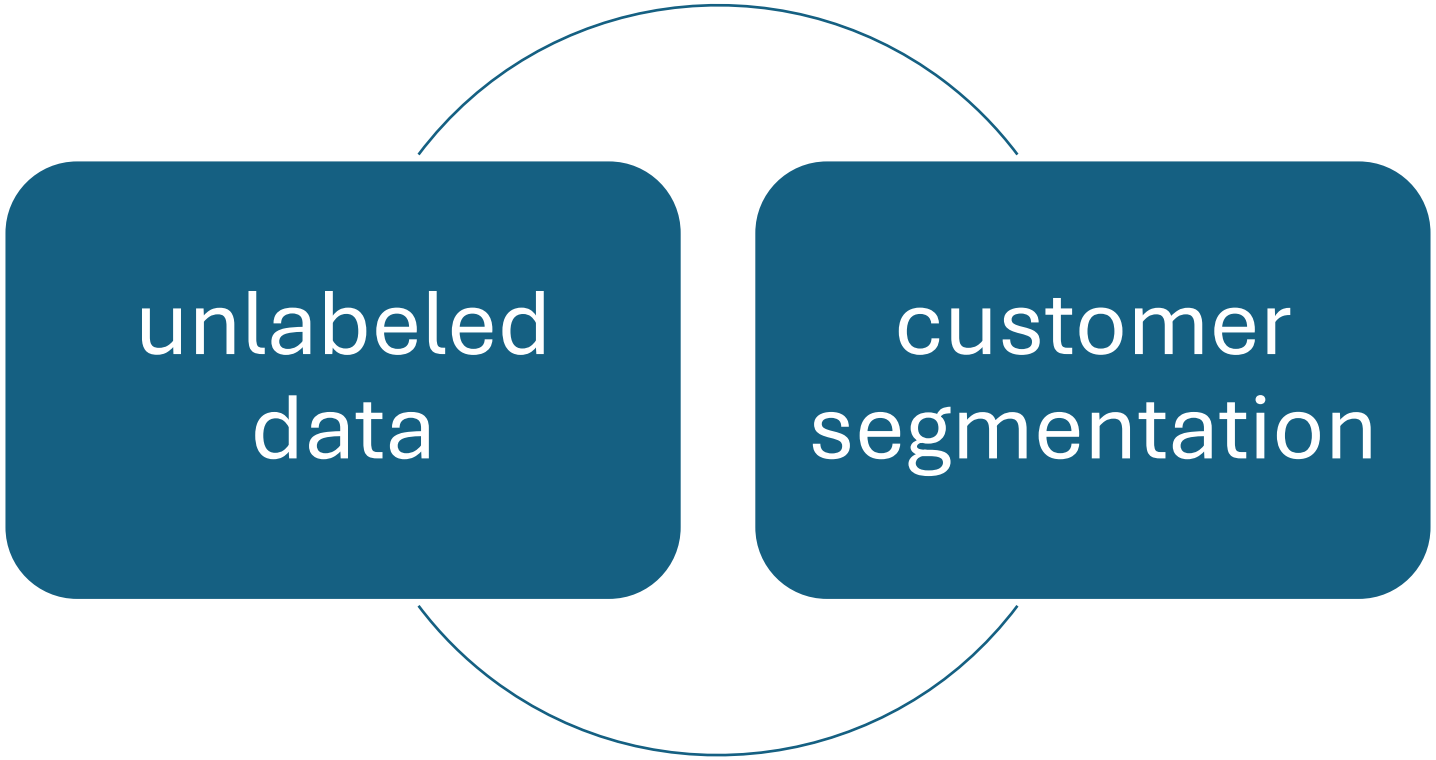
Supervised

LABELED DATA

PREDICTING PRICE

BASED ON FEATURES

Unsupervised



Reinforcement

LEARNING VIA REWARDS

DYNAMIC PRICING BOTS

Walmart Example

Train a model to predict:

Will this customer return a product?

based on past behavior, product type, and location.

Learns patterns from historical data

Deep Learning (DL) – Neural Networks with Many Layers

Subset of ML that

uses **artificial neural networks**

(like the human brain)

with **multiple layers**

to solve complex problems.

Strengths

Handles huge datasets

Great for unstructured data (text, images, speech)

Walmart Example

Use DL to analyze:

Product reviews and extract sentiment:

“This shoe feels uncomfortable” → 🔴 Negative

“This dress is amazing!” → 🟢 Positive

Generative AI

A branch of DL that focuses on

generating new content

like text, images, music, or code

by learning from existing data.

Examples

ChatGPT,

DALL·E,

Midjourney,

GitHub Copilot

Walmart Example

Automatically write product descriptions for 10,000+ items:

This microwave oven features

30L capacity, 5 cooking modes,

and child lock safety.

Large Language Models

MASSIVE GENERATIVE AI MODELS

TRAINED ON **BILLIONS OF TEXT EXAMPLES**

TO UNDERSTAND AND GENERATE

HUMAN-LIKE LANGUAGE.

Examples

GPT (OpenAI)

Claude (Anthropic)

Gemini (Google)

LLaMA (Meta)

Walmart Example

LLM chatbot answers:

When will my order arrive?”

--

Your order will reach you in Pune by Tuesday.

Tracking ID: #923XX

Walmart Example

Learns context,

tone, facts

Generates clear,

natural replies

AI vs ML vs DL vs GA vs LLMs

Feature	AI	ML	DL	GA	LLMs
Definition	Machines mimicking human intelligence	Learning from data	Neural networks with layers	Generate content from data	Giant models trained on text
Input	Rules / logic	Data + Labels	Lots of data (text, images)	Text, images, music	Massive text datasets
Output	Actions / decisions	Predictions / labels	Features / patterns	New content (text, image)	Fluent, context-aware language

AI vs ML vs DL vs GA vs LLMs

Feature	AI	ML	DL	GA	LLMs
Walmart Use Case	Rule-based discount logic	Predict product return	Understand sentiment in reviews	Auto-write product copy	Answer customer queries, search
Human-Like Creativity	❌	❌	⚠️ Limited	✅ Yes	✅✅ Real-time fluent responses
Example Model / Tool	Rule engine, chatbot	scikit-learn, XGBoost	CNN, RNN, BERT	DALL·E, Copilot, ChatGPT	GPT-4, Claude, Gemini, LLaMA

Imagine building a Walmart Assistant

Layer	Role
AI	“Do X if Y happens” — a fixed, rule-based helper
ML	Learns from past orders to predict behavior
DL	Sees patterns in reviews/images to detect issues
GA	Writes personalized product ads in seconds
LLM	Talks to customers like a trained human agent

Short Video

- 🎬 AI vs ML vs DL vs Generative AI – Quick Visual Explanation
- (Useful for learners and corporate teams)

How LLMs Work?
(Demystifying the Black Box)

Imagine a **super-intelligent Walmart associate** that has:

Read **every Walmart product manual,**

customer query, receipt, and

return policy

How LLMs Work ?
(Demystifying the Black Box)

UNDERSTOOD HOW PEOPLE ASK QUESTIONS

WHERE’S MY ORDER?

WHAT’S THE RETURN POLICY?

How LLMs Work?

Learned to give helpful answers,

even summarizing reviews or

translating product info

into Spanish or Hindi

How LLMs Work?

That’s what an LLM does.

It doesn’t just **search** information,

it **understands context** and

generates human-like responses.

Walmart’s GenAI Stack in Action

Scenario

A customer asks:

Is this T-shirt available in medium, and

how fast can it be delivered?

What Happens Behind the Scenes

Step	Tech Involved	Explanation
1. Query Understanding	LLM (e.g., OpenAI GPT or Walmart-trained model)	The model understands the question: size + availability + delivery
2. Information Fetching	Walmart's Product API + Inventory DB	It checks stock in nearby warehouses or stores

What Happens Behind the Scenes

Step	Tech Involved	Explanation
3. Response Generation	LLM formats a human-like answer	“Yes, this T-shirt is available in Medium and can be delivered to Pune by Tuesday if you order now.”
4. Personalization	GenAI + Customer Context	If the user is a loyal member, it might offer: “Free delivery for Walmart+ members”

Benefits to Walmart

Reduces manual queries to call centers

Enhances customer satisfaction

with **24/7 smart assistance**

Enables **multilingual support** and

auto-generated product summaries

Neural Generative Modeling

TOKENIZATION: BREAKING DOWN INPUT TEXT

EMBEDDING: REPRESENTING LANGUAGE IN VECTOR SPACE

GENERATION: SAMPLING AND DECODING TECHNIQUES

What is Neural Generative Modeling?

HOW **LARGE LANGUAGE MODELS (LLMS)**

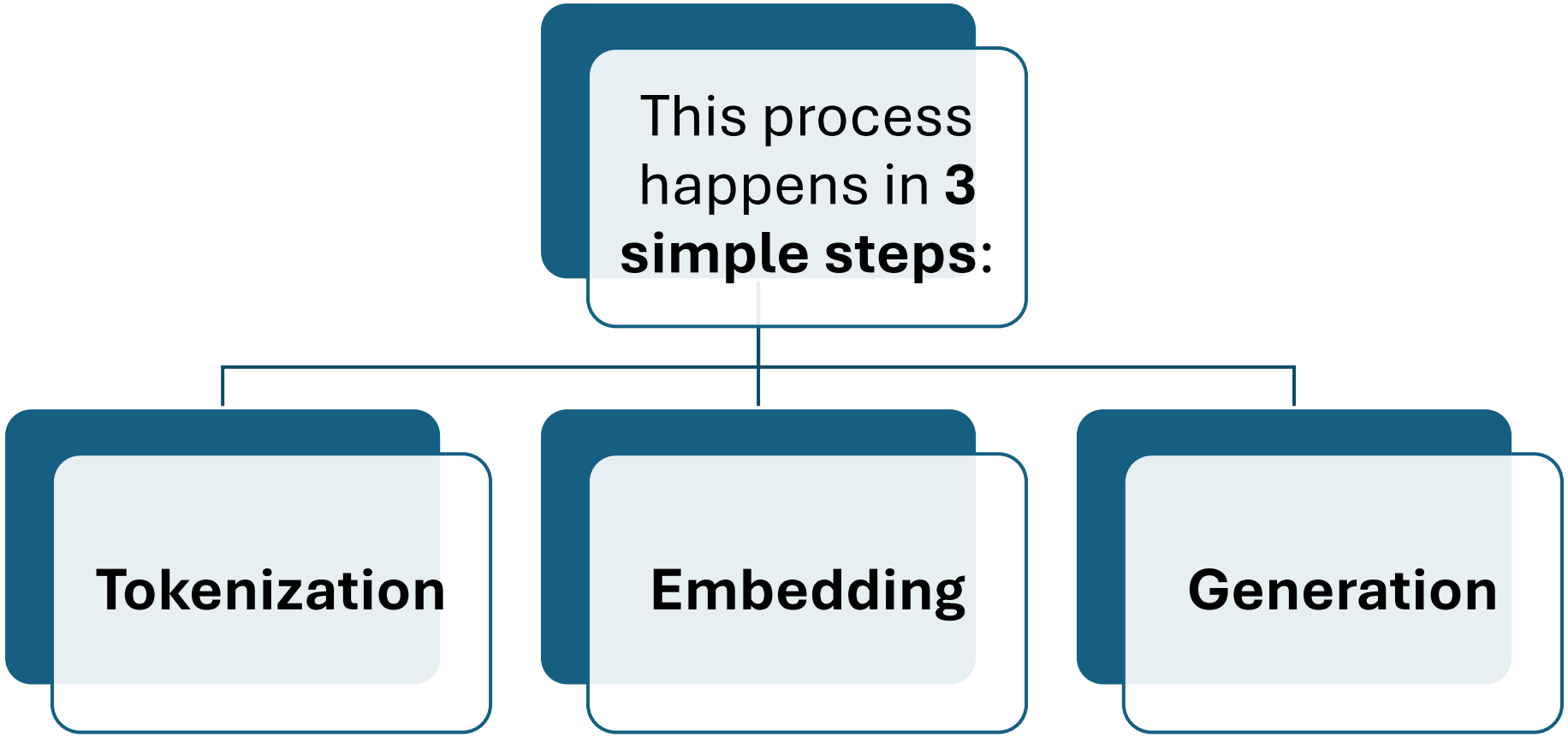
LIKE GPT WORK **BEHIND THE SCENES**

TO TURN CUSTOMER INPUTS

LIKE “WHERE’S MY ORDER?

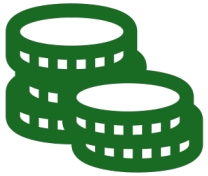
INTO SMART, HUMAN-LIKE RESPONSES.

What is Neural Generative Modeling?



Step 1: Tokenization

Break the input sentence



into **smaller units (tokens)**

like words or parts of words

that a computer can understand.

Walmart Example:

A customer types:

Is this laptop available in Mumbai?

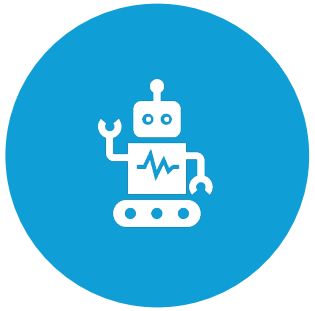
The model **tokenizes** this into:

[“Is”, “this”, “laptop”, “available”, “in”, “Mumbai”, “?”]

Step 2: Embedding

CONVERT EACH TOKEN INTO A

VECTOR (A LIST OF NUMBERS)



SO THE MODEL CAN “UNDERSTAND”

ITS **MEANING AND CONTEXT** MATHEMATICALLY.

Step 2: Embedding

These vectors capture **relationships** between words.

For example,

“Mumbai” will be **closer in meaning**

to “city” or “location”

than “price” in this vector world.

Walmart Example

The model now knows that

“Mumbai” is a **city**, and “available” means **check stock**, and “laptop” is a **product**.

So it prepares to **look up inventory and context.**

Step 3: Generation

The model generates a smart response

using **decoding techniques** like:

Greedy Search

choosing most probable word at each step

Step 3: Generation

Beam Search

considering multiple likely responses

Top-k or Top-p sampling

for diversity

Walmart Example

Now, based on internal inventory and location info,

it generates:

“Yes, this laptop is in stock at our Mumbai store and

can be delivered by Monday.”

Real Impact for Walmart

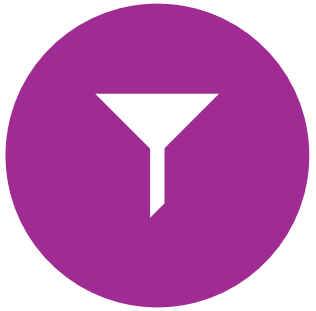
Component	Example in Walmart GenAI Tech Stack
Tokenization	Parsing customer queries: “refund”, “late order”, “COD?”
Embedding	Understanding what “return window” or “express shipping” means
Generation	Creating replies like “Your refund will be processed in 3 days.”

Real Impact for Walmart

THESE MODELS ARE ALSO USED IN:

AUTO-GENERATED **PRODUCT DESCRIPTIONS**

CUSTOMER SERVICE CHATBOTS



SMART SEARCH (“CHEAP SHOES FOR KIDS” → FILTERS BY PRICE + AGE)

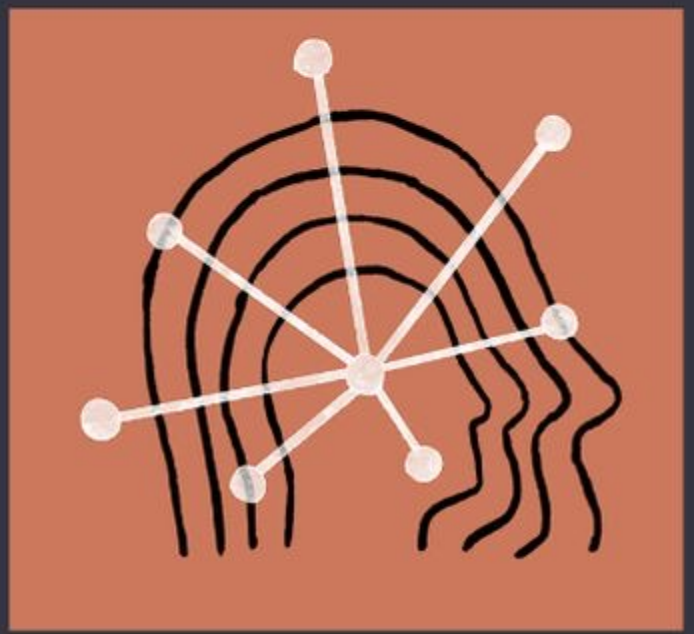
Summary

Step	What It Does	Walmart Use
Tokenization	Break text into pieces	“Where’s my order?” → [Where, is, my, order]
Embedding	Convert to math meaning	Understand “order” = “delivery” context
Generation	Produce response text	“Your package will arrive by 6 PM today.”

3. Large Language Models (LLMs)

Overview of popular models:

- GPT (OpenAI)
- Claude (Anthropic)
- Gemini (Google)
- LLaMA (Meta)



Gemini vs ChatGPT vs Claude

LLMs: GPT, Gemini, Claude

Objective:

Help decision-makers understand how to

leverage LLM like GPT (OpenAI),

Gemini (Google), and Claude (Anthropic)

to drive innovation, efficiency, and competitive edge in automotive.

What Are LLMs?

AI SYSTEMS TRAINED

ON MASSIVE
AMOUNTS OF DATA

BOOKS, MANUALS,
WEB CONTENT, CODE

What Are LLMs? (In Simple Words)

UNDERSTAND NATURAL LANGUAGE

ANSWER QUESTIONS

GENERATE TEXT, CODE, DOCUMENTATION

PERFORM REASONING AND SUMMARIZATION

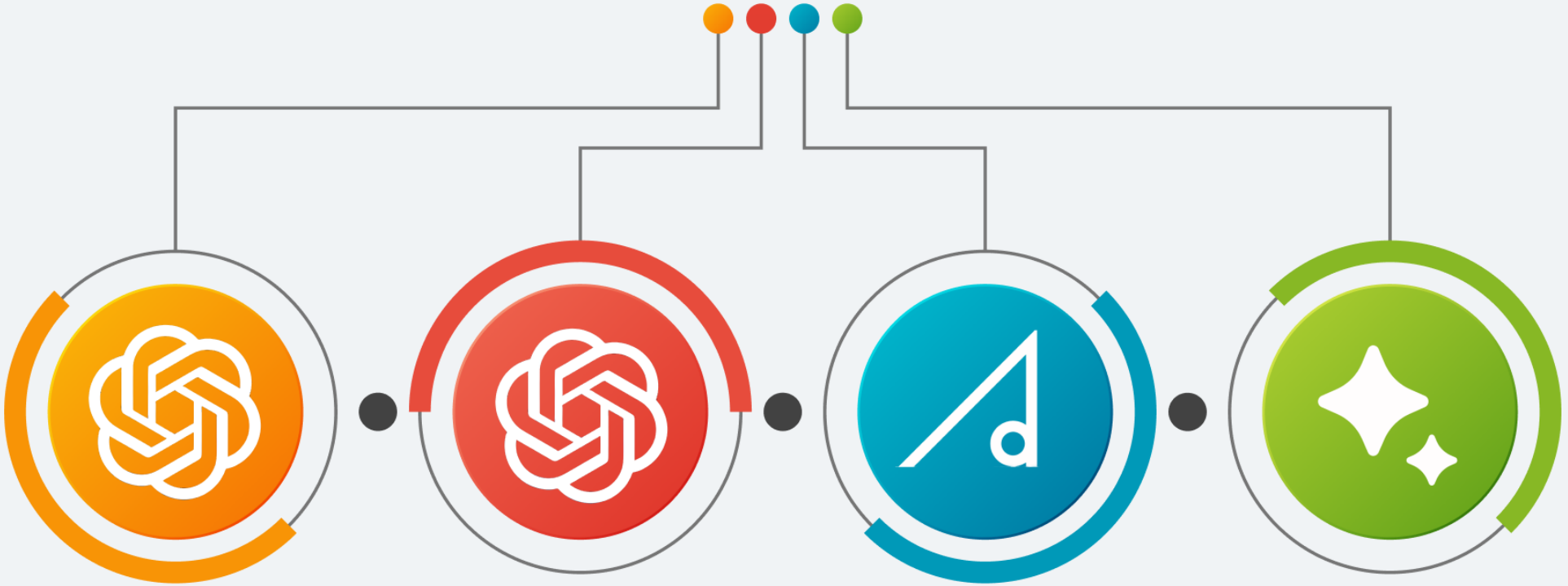
ACT AS A “CO-PILOT” FOR WORK

Top 3 LLMs at a Glance

Model	Developed By	Key Strengths	Unique Features
GPT-4	OpenAI (Microsoft-backed)	Industry-leading accuracy & creativity	Plugins, Code Interpreter, Copilot integration
Gemini	Google (DeepMind)	Multimodal (text + image + video)	Deep GCP integration, powerful search context
Claude	Anthropic	Safer, ethical, large context window (100k+)	Friendly tone, long document understanding

orangemantra

TOP GENERATIVE AI TOOLS



GPT-4

ChatGPT

AlphaCode

Bard

ChatGPT vs. Gemini vs. Claude vs Llama

Which AI is better in 2025?

https://www.youtube.com/watch?v=5KiDabAa9JY

1. Core Strengths & Use Cases

ChatGPT-4 (GPT-4o / 4.5)

Best overall performer in coding,

creative writing, and

multimodal interactions.

Excelled in calculus

problem-solving (94.7% success).

https://openai.com/index/introducing-gpt-4-5/

1. Core Strengths & Use Cases

Google Gemini (2.5 Pro/Flash)

Strong in multimodal context

(text+image+audio/video),

reasoning-heavy tasks, and

cost-effective free usage.

https://gemini.google.com/app

1. Core Strengths & Use Cases

Anthropic Claude (4 Sonnet / Opus)

Tops safety, ethical alignment,

long-context reasoning

best accuracy in reading comprehension

without hallucinations.

https://claude.ai/onboarding?returnTo=%2F%3F

1. Core Strengths & Use Cases

Meta Llama 4 (Scout & Maverick)

Competitive open-source option,

excels in reasoning and

massive-context tasks,

rivaling GPT-4o in benchmarks.

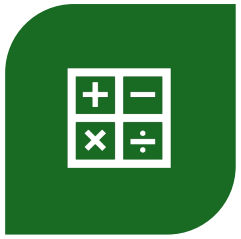
Meta Llama 4 (Scout & Maverick)

Meta_UI Interface

https://www.meta.ai/?utm_source=ai_meta_site&utm_medium=web&utm_content=AI_nav&utm_campaign=06112025_moment

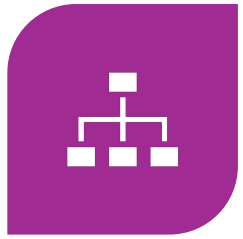
2. Performance Highlights

BENCHMARKS & ACCURACY



CHATGPT-4 LEADS IN CODING & CALCULUS

CLAUDE PRO FOLLOWS CLOSELY.



ON OLYMPICARENA, GPT-4O AND

CLAUDE-3.5-SONNET TOP GEMINI-1.5-PRO.

3. Strengths & Weaknesses Table

Feature	ChatGPT-4/4.5	Gemini 2.5	Claude 4	Llama 4
Coding	🔝 Excellent	✅ Very strong	⚙️ Moderate to strong	Good (via Code Llama)
Creative Writing	🔝 Superior prose	✅ Good (free-tier leader)	✅ Solid clarity & consistency	Moderate
Reasoning & Math	🔝 Top performance	✅ Strong	✅ Excellent accuracy	✅ Competitive

3. Strengths & Weaknesses Table

Feature	ChatGPT-4/4.5	Gemini 2.5	Claude 4	Llama 4
Multimodality	✅ Text + Images	🔝 Natively multimodal	✅ Image + document support	✅ Multimodal support
Safety & Ethics	✅ Good	⚠️ Moderate	🔝 Rigorous guardrails	Open-source caution

3. Strengths & Weaknesses Table

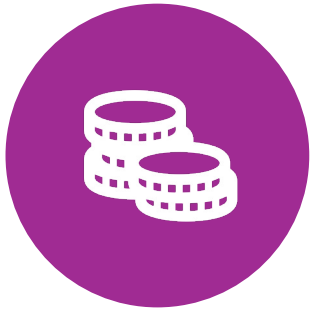
Feature	ChatGPT-4/4.5	Gemini 2.5	Claude 4	Llama 4
Cost & Access	Paid tiers	Free-tier strong	Free/open beta	Open-source
Long Context Handling	✅ Large	✅ Million-token window	✅ Opus with 200k+	✅ 1M token window

4. Emerging Trends & Developments

GEMINI 2.5 PRO

INTRODUCES “DEEP THINK”

REASONING MODE AND



MILLION-TOKEN CONTEXTS.

4. Emerging Trends & Developments

CLAUDE 4 SONNET/OPUS

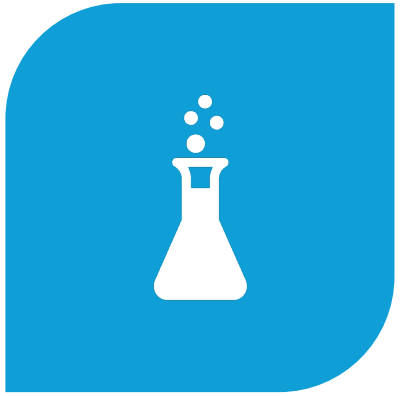
ADDS CODE EXECUTION AND

BROWSING WITH HIGH SAFETY-FOCUS.

4. Emerging Trends & Developments

LLAMA 4 SCOUT/MAVERICK

ACHIEVE GPT-4 PARITY ON REASONING BENCHMARKS



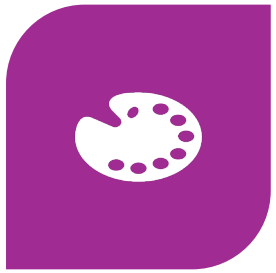
USING MIXTURE-OF-EXPERTS ARCHITECTURE.

2025 Pick by Use Case

OVERALL BEST: CHATGPT-4,

WITH UNMATCHED

VERSATILITY AND



CREATIVE CAPABILITY.

2025 Pick by Use Case

BEST FREE AND MULTIMODAL TOOL

GEMINI 2.5 FLASH

STRONG MULTIMODAL AND

GENERAL PERFORMANCE.

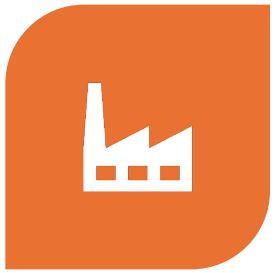
2025 Pick by Use Case

SAFEST AND MOST ACCURATE

FOR LONG-FORM OR LEGAL WORK:

CLAUDE 4.

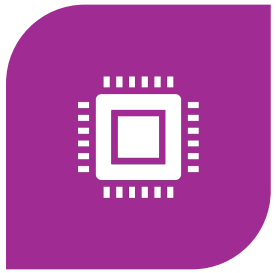
2025 Pick by Use Case



BEST OPEN-SOURCE POWERHOUSE

LLAMA 4 (MAVERICK),

ESPECIALLY FOR



ON-PREM/SELF-HOSTED USE.

Final Take

Each AI is best-in-class for specific goals:

🔷 **ChatGPT-4** – elite all-rounder.

🟦 **Gemini** – multimodal & free utility.

🟩 **Claude** – safe, ethical, long-context.

🟨 **Llama** – scalable open-source.

🛠️ How Walmart Might Use These LLMs

Model	Use Case at Walmart
GPT-4.5	Generate SEO-friendly product pages and email copy
Claude Opus	Explain complex policies or troubleshoot customer issues
Gemini Flash	Enhance store experience via voice/image queries in apps
LLaMA 4	Build proprietary, open-source solutions for internal bots

Model	Developer	Key Features	Strengths	Weaknesses
ChatGPT	OpenAI	Conversational, versatile	Strong conversational abilities, creative writing, code generation	Potential misuse, limited real-time information access
Google Gemini	Google	Multimodal, real-time information access	Technical and scientific knowledge, internet search, integrates with Google services	Can be verbose, prone to "hallucinations"
Meta AI	Meta	Contextual understanding, social interactions	Nuance in communication, social trend analysis, integrates with Meta platforms	Limited availability, potential privacy concerns
Claude	Anthropic	Reasoning, complex task handling	Analytical abilities, ethical focus, handles long-form content well	Can be slow, struggles with creative tasks

Walmart’s GenAI Use Cases

GenAI Use Case	Example
🛍️ **Customer Support Automation**	“Where is my order?”, “How do I return this product?”
🛒 **Smart Product Descriptions & Reviews**	Generate product titles, bullets, summaries
🔍 **Search & Recommendation Enhancement**	“Show me kids shoes under ₹1000 in Pune”
 **Supply Chain, Inventory, Pricing Insights**	Forecast stockouts, restocking tips, price trends

Walmart’s GenAI Use Cases

GenAI Use Case	Example
💬 **Multilingual Chat & Accessibility Support**	Translate FAQs, handle Hindi, Spanish, etc.
🎯 **Personalized Marketing & Campaigns**	Tailored email, SMS copy based on user preferences
🏧 **In-store AI Assistant (voice/image/text)**	“Scan product, ask price/stock via app or kiosk”

LLM Comparison for Walmart's Needs

Feature / Model	GPT (OpenAI)	Claude (Anthropic)	Gemini (Google)	LLaMA (Meta)
Model Strength	Creative writing, reasoning, coding	Safe, ethical reasoning, summarization	Multimodal (text+image+voice) + fast recall	Lightweight, open-source, fine-tuned easily
Best Use in Walmart	Chatbots, marketing, product generation	Support FAQs, policy reasoning	Mobile app assistant, search UX	Internal AI agents, warehouse automation
Accuracy (Reasoning)	✅✅✅✅	✅✅✅✅✅	✅✅✅	✅✅ (depends on fine-tuning)

LLM Comparison for Walmart's Needs

Feature / Model	GPT (OpenAI)	Claude (Anthropic)	Gemini (Google)	LLaMA (Meta)
Response Tone	Friendly + versatile	Professional + safety-first	Neutral, Google-style	Customizable
Multilingual Support	50+ languages	Limited	100+ languages + voice synthesis	Depends on training data
API Access	✅ (OpenAI platform, Azure OpenAI)	✅ (Claude API, Amazon Bedrock)	✅ (Gemini API, Google Cloud Vertex AI)	✅ (open-source via Hugging Face)

LLM Comparison for Walmart's Needs

Feature / Model	GPT (OpenAI)	Claude (Anthropic)	Gemini (Google)	LLaMA (Meta)
Multimodal (Image, Text, Voice)	GPT-4V supports images	❌ (text only)	✅ (images, speech, documents)	❌ (text only by default)
Customization	✅ (via fine-tuning, instructions)	Limited customization	Limited, focused on safety	✅✅ (open source, retrainable)
On-Premise Possibility	❌ (cloud only)	❌	❌	✅✅ (host internally)

LLM Comparison for Walmart's Needs

Feature / Model	GPT (OpenAI)	Claude (Anthropic)	Gemini (Google)	LLaMA (Meta)
Cost (for scale)	₹₹₹₹ (high, pay-per-token)	₹₹₹ (moderate via Bedrock/Claude.ai)	₹₹₹ (competitive via Google Cloud)	₹ (free, but infra cost applies)
Security & Privacy	Enterprise-grade, SOC2, GDPR	Focused on safety & data privacy	Google's enterprise-grade security	Customizable per deployment

Walmart-Specific Recommendations

Use Case	Best Model(s)	Why
Customer Chatbot / Support Agent	Claude / GPT	Claude for safe & accurate answers; GPT for flexibility
Product Content Generation	GPT	Excellent at creative text, variations, SEO content
Voice + Visual Product Search	Gemini	Multimodal capability ideal for mobile retail app UX
Warehouse AI Agent (internal use)	LLaMA	Open-source, can be tuned on logistics data

Walmart-Specific Recommendations

Use Case	Best Model(s)	Why
Policy Assistant for Managers	Claude / LLaMA	Claude for natural policy language; LLaMA for cost-effective automation
Personalized Marketing Messages	GPT	Personalization + creativity + tone tuning
Multilingual Regional FAQs	Gemini / GPT	Gemini: wider language + audio support

Access & Integration

Platform	Access URL	Integration Tips
GPT	https://platform.openai.com	Use Azure OpenAI for enterprise SLAs
Claude	https://claude.ai	Available via Anthropic API or Bedrock
Gemini	https://gemini.google.com	Use Vertex AI or Gemini Pro APIs
LLaMA	https://llama.meta.com	Download via Hugging Face for customization

Final Recommendations for Walmart

Customer-Facing:

Use GPT or Claude for **chat and content.**

Mobile/Store Assistants:

Use Gemini for **multimodal app interactions.**

Final Recommendations for Walmart

Internal/Backoffice AI:

Deploy LLaMA to save cost and **retain full control**.

Final Recommendations for Walmart

For **critical reasoning + safety**,

Claude is best.

For **scale + creativity**,

GPT remains top performer.

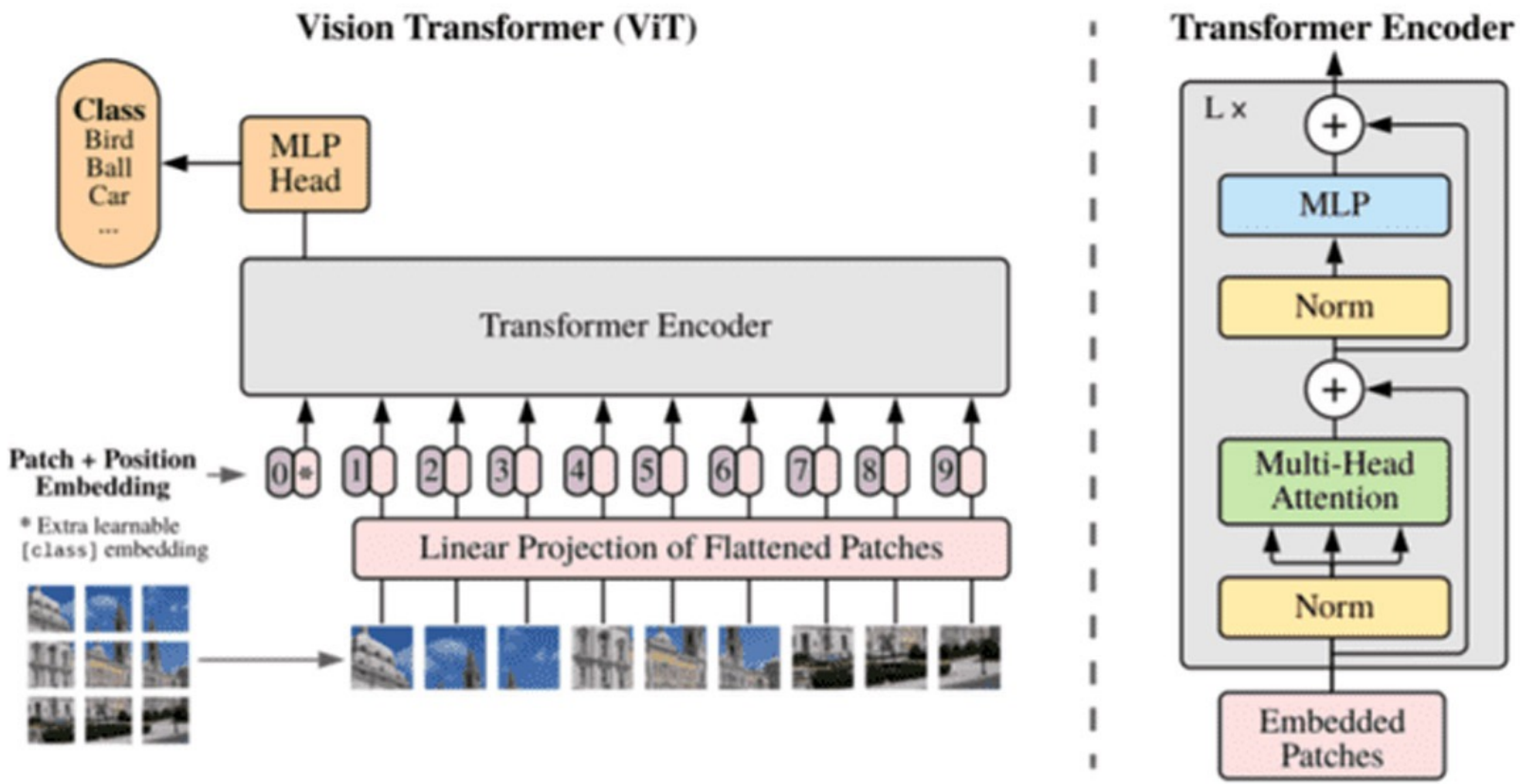
4. Transformer Architecture & Attention Mechanism

Self-Attention and its role in LLMs

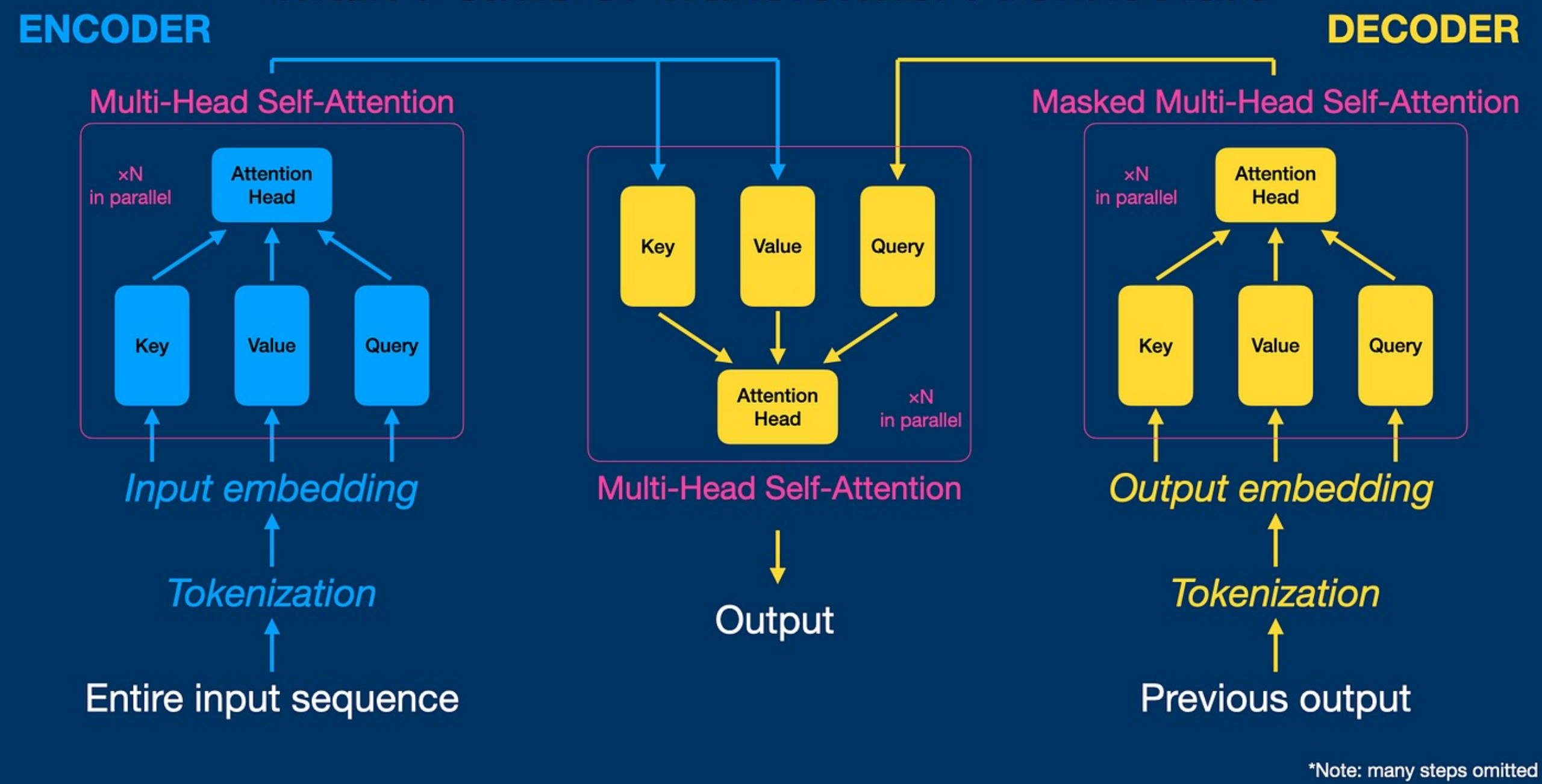
Encoder-only, decoder-only, and encoder-decoder variations

Use cases and model examples for each type

Transformers



Main Points of Transformer Architecture



*Note: many steps omitted

Self-Attention: The Core Mechanism

WHAT IT DOES?

EACH WORD EXAMINES ALL OTHER WORDS

IN A SENTENCE TO DECIDE WHAT MATTERS MOST.

WORDS BECOME **QUERIES**, **KEYS**, AND **VALUES**.

Self-Attention: The Core Mechanism

Why it's important?

Enables understanding of context

regardless of word order

like correctly relating

"it" to "laptop" or "delivery"

Walmart Example

A question “When will my return arrive?”

“return” attends to “my” and

“arrive” to grasp intent,

rather than just the words themselves.

Transformer Variants

a) Encoder-Only

Structure: Stack of self-attention layers (like BERT)

Best for:

Understanding tasks—classification,

search embedding, document summarization.

Transformer Variants

Walmart Use:

Create vector representations

of product reviews

for sentiment analysis or

classification.

Transformer Variants

b) Decoder-Only

Structure:

Masked self-attention + feed-forward layers (like GPT)

Best for:

Generating text, chatbots, email copy.

Transformer Variants

WALMART USE:

CHAT SUPPORT,

AUTO-GENERATING PRODUCT DESCRIPTIONS OR

SMS NOTIFICATIONS.

Transformer Variants

c) Encoder-Decoder

Structure:

Encoder processes input →

Decoder uses both self- and cross-attention

(original “Attention is All You Need”)

Transformer Variants

Best for:

Tasks converting one text to another

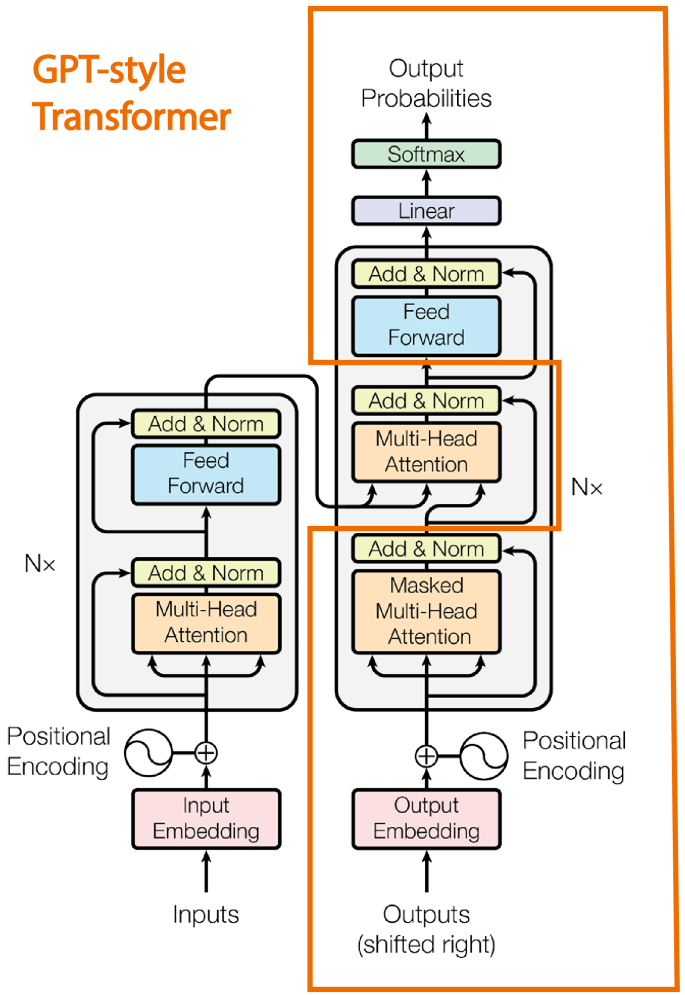
translation, summarization.

Walmart Use:

Summarize long returns policies

into a concise FAQ answer.

GPT-style Transformer

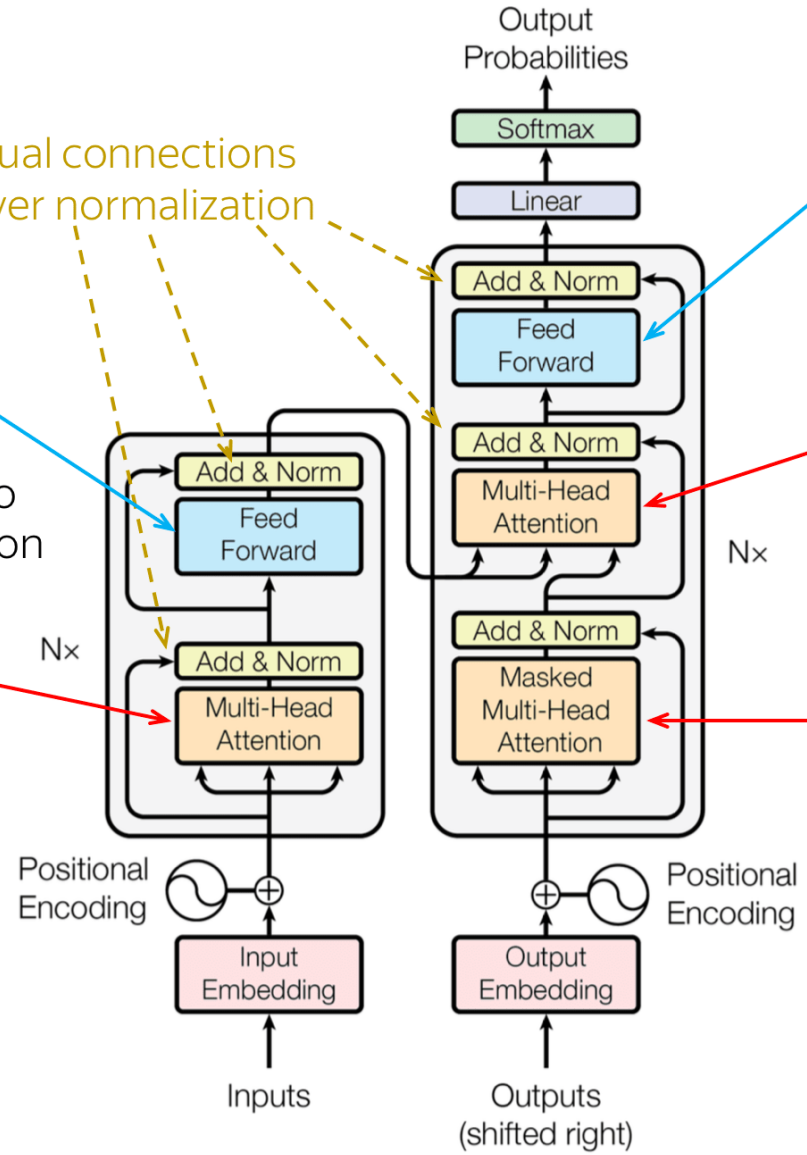


Residual connections and layer normalization

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Encoder self-attention: tokens look at each other

queries, keys, values are computed from encoder states



Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Decoder-encoder attention: target token looks at the source

queries – from decoder states; keys and values from encoder states

Decoder self-attention (masked): tokens look at the previous tokens

queries, keys, values are computed from decoder states

Attention Flavors

Self-Attention:

Words attend to words

within the same sentence

(question or policy).

Attention Flavors

Masked Self-Attention:

Decoder-only models only

look at **prior words**

keeps output coherent in auto-generation

Attention Flavors

Cross-Attention:

Decoder aligns each output token

with the encoded input

from the encoder

Walmart Example

Customer asks: “Translate this review to Hindi.”

Encoder reads English review,

Decoder cross-attends to that context and

outputs Hindi translation.

Simplified Architecture Flow

Encoder-Only Flow:

Input → Tokenize →

Embed → Self-Attention →

Feed-Forward →

Output embeddings.

Simplified Architecture Flow

Decoder-Only Flow:

Prompt → Masked Self-Attention →

Feed-Forward →

Generate next word.

Simplified Architecture Flow

Encoder-Decoder Flow:

Input → Encoder → Context →

Decoder + Cross-Attention →

Output.

Walmart-Style Use Cases

Model Type	Core Feature	Walmart Example
Encoder-only	Deep understanding, embeddings	Classifying returns as ‘damaged’, ‘late’, ‘exchange’
Decoder-only	Autoregressive text generation	Creating chat responses (“Your order ships tomorrow.”)
Encoder-decoder	Translate, summarize, convert	Summarize long policy into brief customer-friendly answer

Why It Matters for Walmart?

Scalability:

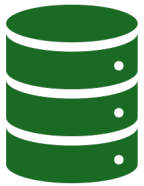
Models parallelize attention,

processing hundreds of requests

in real-time

Why It Matters for Walmart?

Accuracy:



Capture long-distance dependencies

“return” relates to both

“refund” and “policy window”.

Why It Matters for Walmart?

FLEXIBILITY:

APPROACH WORKS ACROSS TASKS

CHATBOTS, SEARCH, TRANSLATION,

SUMMARIZATION, AND MORE.

5. RLHF and DPO

RLHF (Reinforcement Learning with Human Feedback)

Feedback loop for model alignment

DPO (Direct Preference Optimization)

Enhancing response quality without RL

RLHF – Reinforcement Learning with Human Feedback

A PROCESS WHERE AN LLM LEARNS

TO IMPROVE ITS OUTPUTS BASED

ON **HUMAN PREFERENCES**.

RLHF

Three stages:

Collect examples: Humans choose better responses over worse ones.

Train a reward model to predict preferences.

Fine-tune the language model using reinforcement learning (e.g., PPO), guided by the reward model .

Walmart Example

A bot drafts two versions of a refund response:

A: “Refund processed in 3 days.”

B: “Your ₹500 refund will be credited within 3 business days to your original payment method.”

Walmart Example

Humans prefer B.

The model learns this and

gradually favors clearer,

more customer-friendly phrasing.

DPO – Direct Preference Optimization

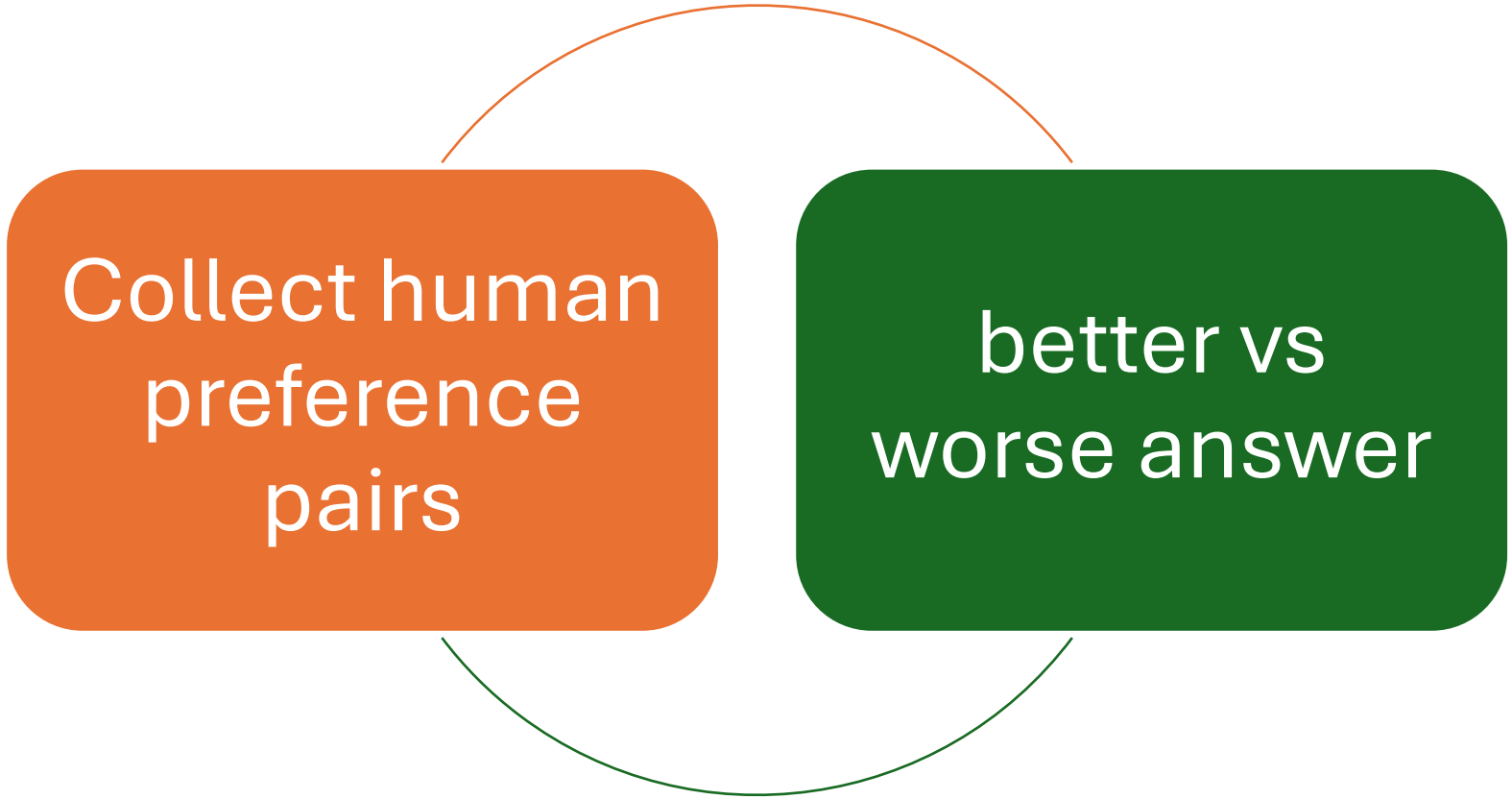
A **more efficient** way

to align models

with human preferences

no reward model, no RL

DPO – Direct Preference Optimization



DPO – Direct Preference Optimization

Use a **classification loss**

to directly bias the model

toward preferred outputs

Avoids RL instability and complexity.

Walmart Example

AFTER A/B RESPONSES:

USES METRIC:

LOG(PROB(A)/PROB(B)).

Walmart Example

Model is directly fine-tuned

to assign **higher probability**

to the better one

no extra RL stage needed.

Video Link

https://www.youtube.com/watch?v=nSrj1J6ODoM

This quick video explains

both concepts clearly and

compares their pros and cons.

RLHF vs DPO – At a Glance

Feature	RLHF	DPO
Pipeline	Complex: preference → reward → RL	Simple: direct preference → fine-tune
Computation	High (reward model + RL)	Moderate (just fine-tuning)
Stability	Can be unstable, needs tuning	More stable and easier to implement
Performance	Strong with policy flexibility	Matches or exceeds RLHF in demos

Why It Matters for Walmart?

RLHF:

Use where safety and correctness are critical

Refund policy explanations,

Sensitive customer support.

Why It Matters for Walmart?

DPO:

Ideal for scalable content tasks

like auto-generating product descriptions or

responses—with faster, leaner fine-tuning.

Summary

RLHF builds alignment via reward models +

RL—flexible but complex.

DPO directly optimizes preferences

simple, stable, efficient.

Summary

Walmart can **mix both** based on use case

RLHF for high-stakes alignment,

DPO for fast and scalable improvements.

6. Python Libraries Overview

transformers – HuggingFace for pretrained models

openai – API interface for GPT models

langchain – Orchestration for LLM workflows

vertexai – Google GenAI integration and utilities

transformers (Hugging Face)

A leading open-source library

for using and fine-tuning pretrained transformer models

text, vision, audio

in frameworks like PyTorch, TensorFlow

Walmart Use Case

Fine-tune product-specific models

to automatically generate bullet points.

Use sentiment-analysis models

to scan customer feedback for

‘return’, ‘defective’, or ‘love’.

Why it’s useful?

Thousands of ready-to-use models.

Supports training, inference,

deployment locally or in the cloud.

openai (OpenAI Python SDK)

Official library to call OpenAI models

(ChatGPT, GPT-4/4o) via API for chat,

content generation,

summarization .

Walmart Use Case

#Generate personalized marketing emails

```
from openai import OpenAI
client = OpenAI(api_key="...")
client.chat.completions.create(model="gpt-4",
messages=[{"role":"user","content":"Write email about toys sale for Pune customers"}])
```

Why it’s powerful?

Plug-and-play enterprise-grade LLMs.

Ideal for customer-facing intelligent text workflows.

langchain (LLM Orchestration Framework)

A modular toolkit to build complex applications

using LLMs—e.g., chatbots, retrieval systems,

multi-step pipelines—

with easy integration of tools like

search, databases, APIs

Walmart Use Case

Build a “Return Assistant”:

Parse user queries,

Retrieve policy from a database (RAG),

Generate an answer with GPT or Claude.

Walmart Use Case

Why it excels:

Turns simple models into powerful agents.

Supports chaining LLM calls + tool usage + memory.

vertexai (Google Cloud Vertex AI GenAI Integration)

Google’s SDK to access GenAI models

like Gemini, deploy agents, and

manage multimodal workflows and pipelines .

Walmart Use Case

Deploy an in-store assistant:

Use image input (customer scans a shelf),

Recognize product via multimodal Gemini,

Respond with availability or offer details.

Walmart Use Case

Why it’s useful?

Simplifies deploying secure,

scalable GenAI solutions

on Google Cloud.

Summary Table

Library	Core Role	Walmart Connection
transformers	Pretrained models local or cloud	Generate bullets, analyze sentiment, fine-tune SKU-specific models
openai	Cloud LLM access	Create chatbots, compose emails, answer customer queries
langchain	Orchestrate workflows	Build complex agents combining search, memory, LLMs
vertexai	Google GenAI deployment	Scale multimodal assistants, voice/image agents in-store or online

Why this Stack Matters for Walmart?

Speed:

Quickly deploy text generation or

question-answering with minimal code.

Scale:

Use vertexai to deploy globally with enterprise infrastructure.

Why this Stack Matters for Walmart?

Flexibility:

Compose powerful bots using langchain that tap multiple data sources.

Customization:

Fine-tune models locally with transformers to match Walmart’s tone and data.

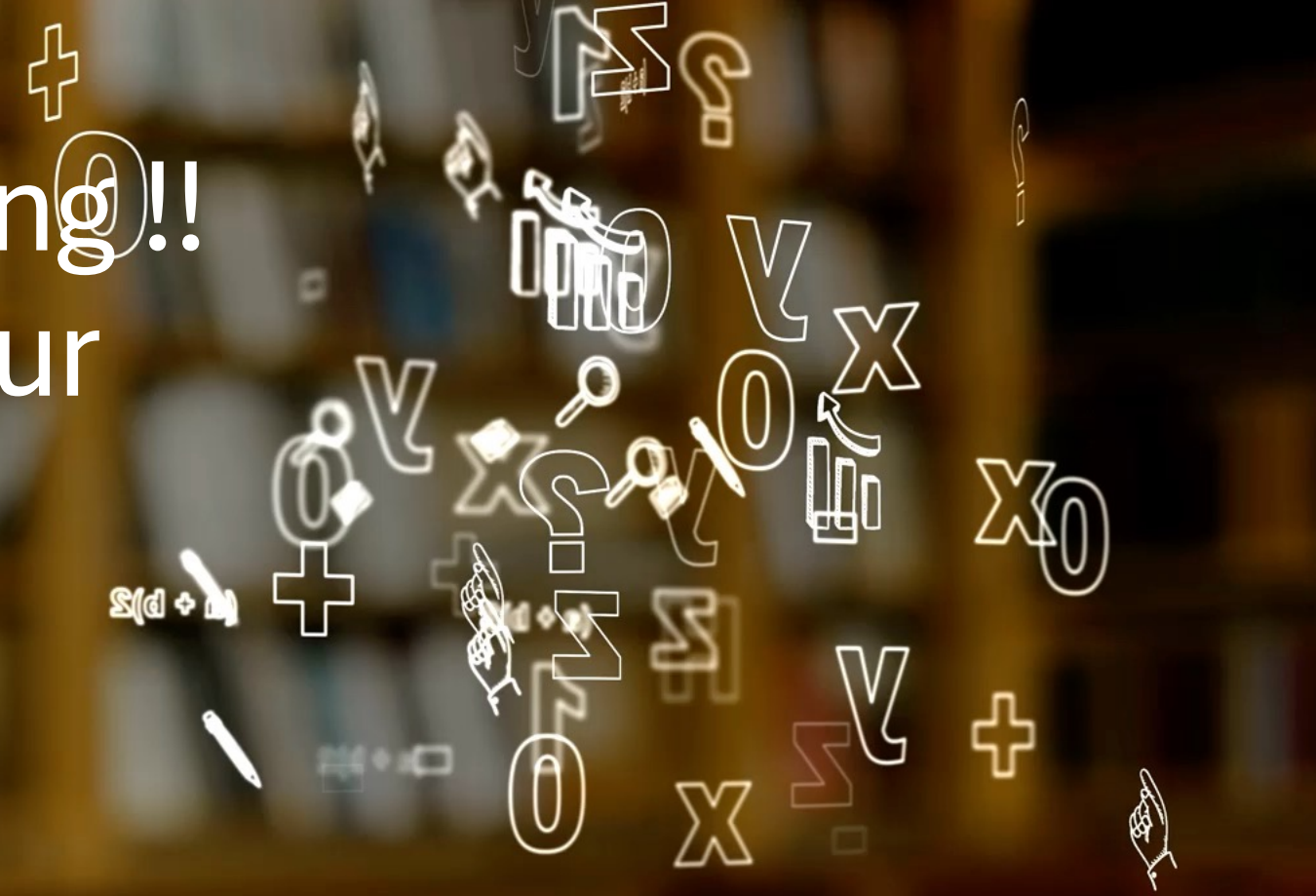
Hands-On Session

Load and interact with GPT or Gemini via Python

Tokenization → Embedding → Generation walkthrough

Explore responses with different prompt styles

Happy Learning!! Thanks for Your Patience ☺



Surendra Panpaliya

GKTCS Innovations