THASMAI
INFOTECH

Advanced Gen AI & Agentic AI

From Basics to Practical Implementation

Training Duration : 10 Days | 80 Hours

Sudarshana Karkala

Director of Engineering

Thasmai Infotech Private Limited

✆ +91 9845561518 | ✉ ThasmaiInfotech @ gmail.com | ThasmaiInfotech.com

EV Engineer
CPR Software Systems

Topics - Gen AI & Agentic AI

THASMAI
INFOTECH

1 Introduction to Generative AI

2 NLP vs LLMs — Foundations

3 Transformer Architecture

4 LLM Training — How they learn

5 Prompt Engineering

6 Embeddings & Vector Databases

7 Retrieval Augmented Generation (RAG)

8 Agents & Agentic AI

9 LangGraph — Orchestrated AI Workflows

10 Model Context Protocol (MCP)

11 Multimodal GenAI

12 GenAI for Engineering

13 Model Deployment, Inference & Optimization

14 Fine-Tuning & Custom Models

15 Safety, Governance & Ethics

16 Evaluation & Testing

17 Hands-On (Python + APIs)

18 Capstone Projects

EV.Engineer
CAR Software Systems

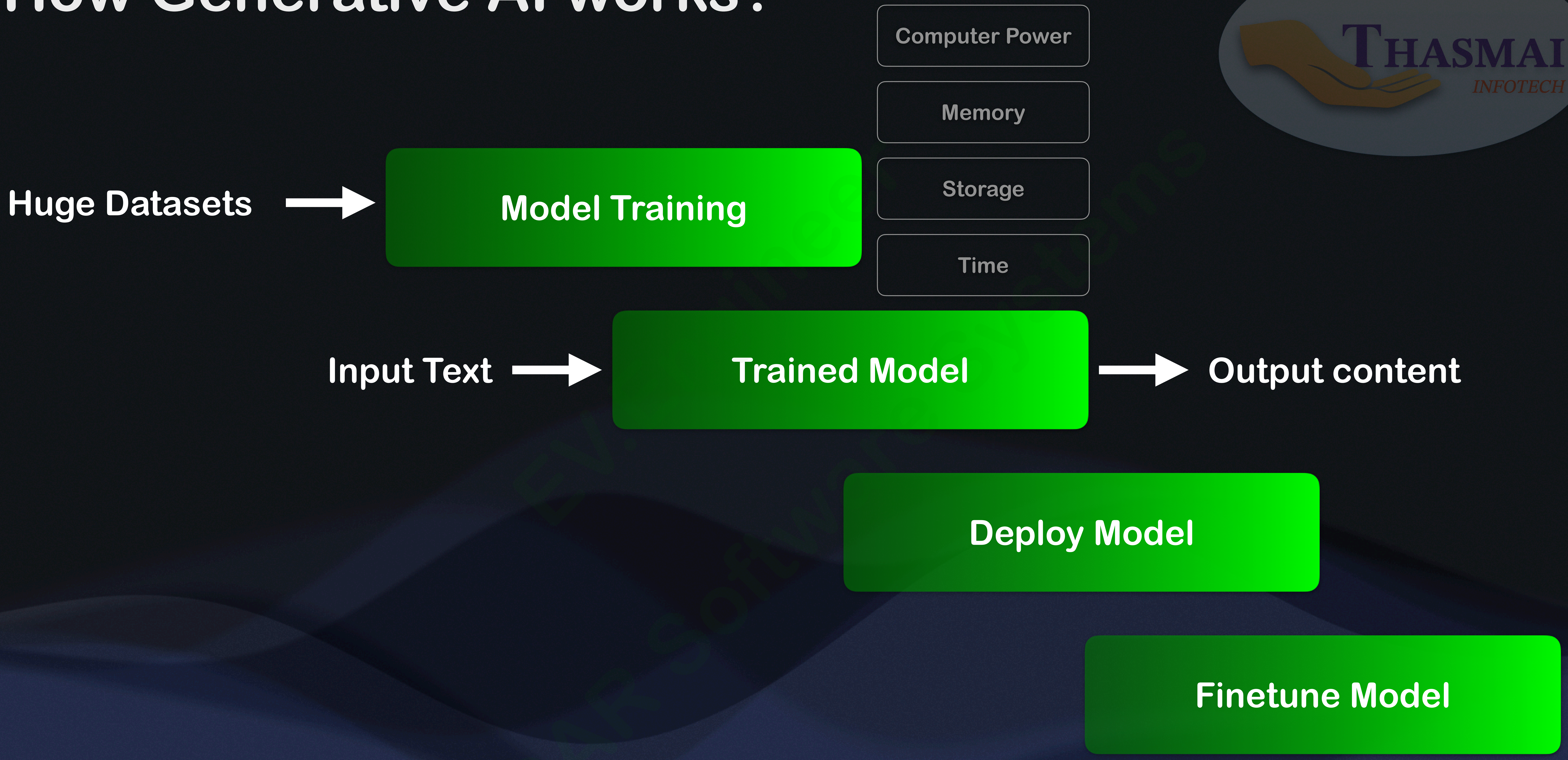
Agentic AI - Design Patterns

THASMAI INFOTECH

1 Prompt Chaining
2 Routing
3 Parallelization
4 Reflection
5 Tool Use
6 Planning
7 Multi-Agent Collaboration
8 Memory Management
9 Learning and Adaptation
10 Model Context Protocol (MCP)
11 Goal Setting and Monitoring
12 Exception Handling and Recovery
13 Human-in-the-Loop
14 Knowledge Retrieval (RAG)
15 Inter-Agent Communication (A2A)
16 Resource-Aware Optimization
17 Reasoning Techniques
18 Guardrails Safety Patterns
19 Evaluation and Monitoring
20 Prioritization
21 Exploration and Discovery

EV Engineer
CAR Software Systems

How Generative AI works?



Computer Power

Memory

Storage

Time

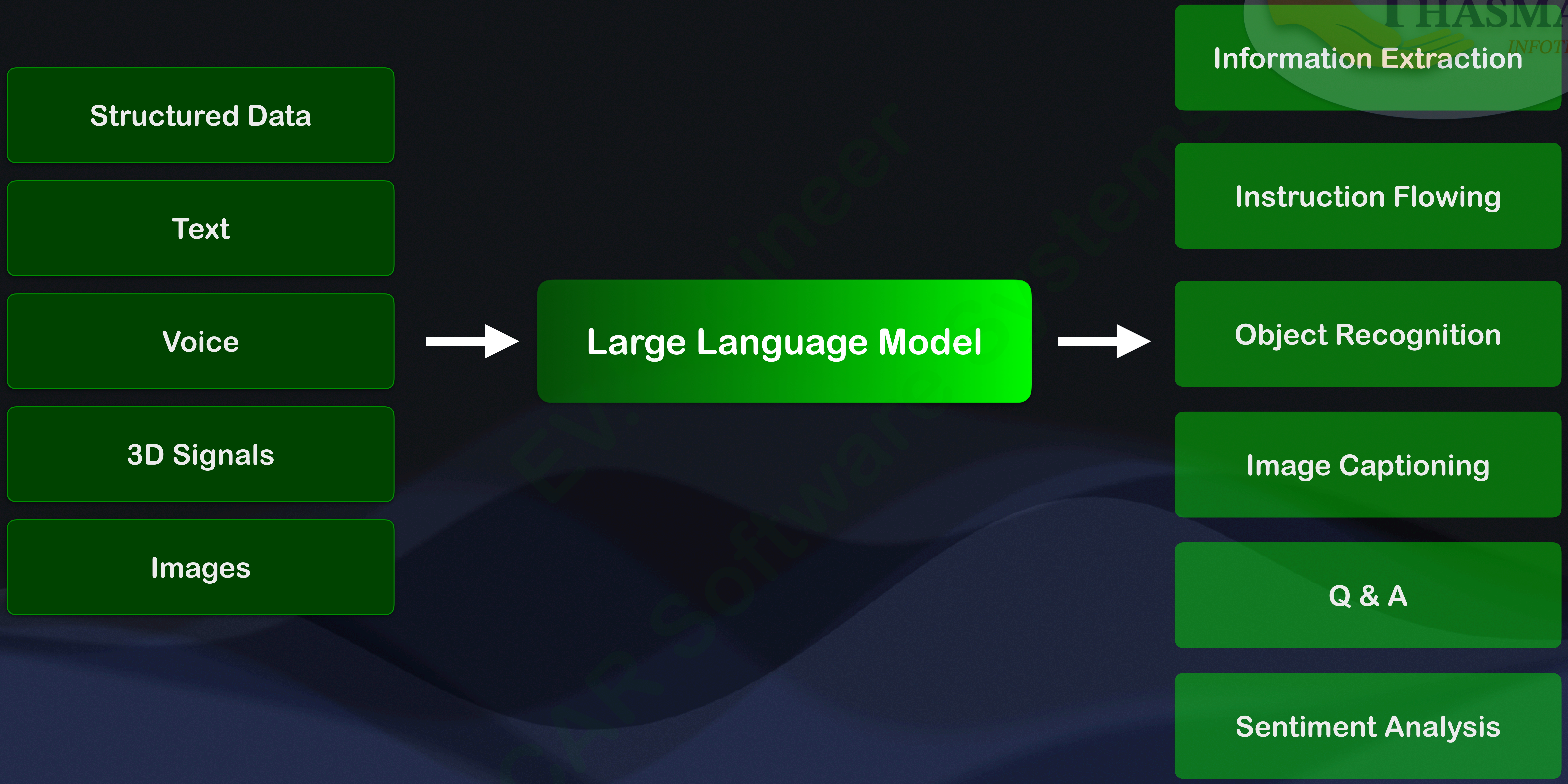
Huge Datasets → Model Training

Input Text → Trained Model → Output content

Deploy Model

Finetune Model

Large Language Models (LLMs)



Structured Data

Text

Voice

3D Signals

Images

Large Language Model

Information Extraction

Instruction Flowing

Object Recognition

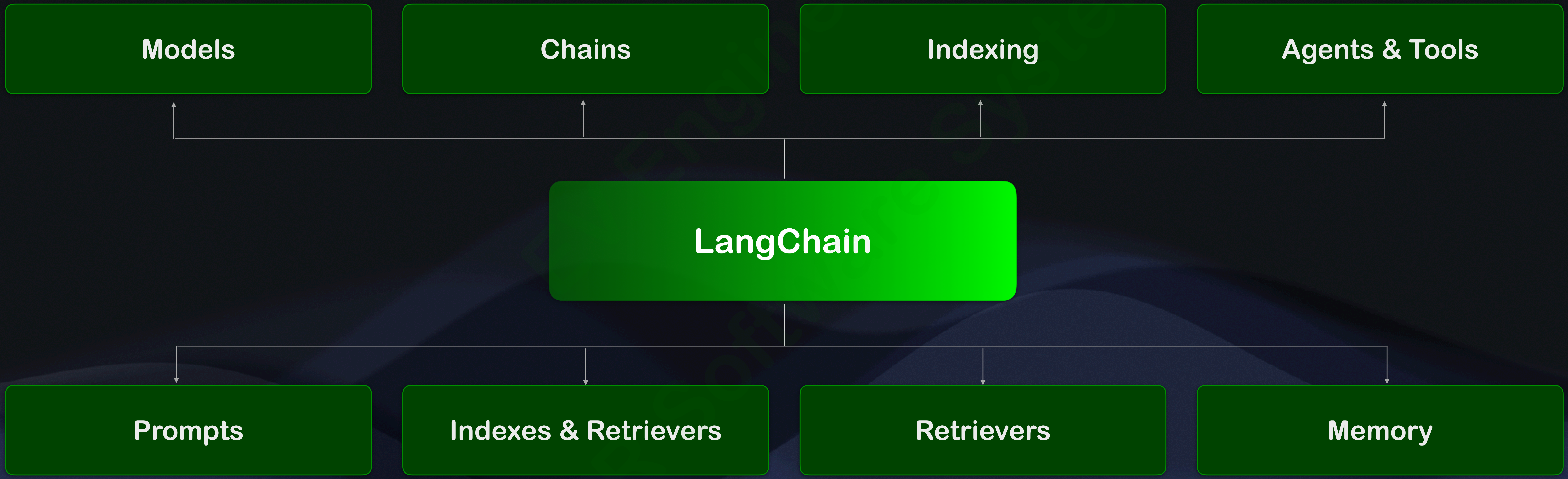
Image Captioning

Q & A

Sentiment Analysis

Components of LangChain

An Open Source framework for building LLM powered apps



LangChain

Models

Chains

Indexing

Agents & Tools

Prompts

Indexes & Retrievers

Retrievers

Memory

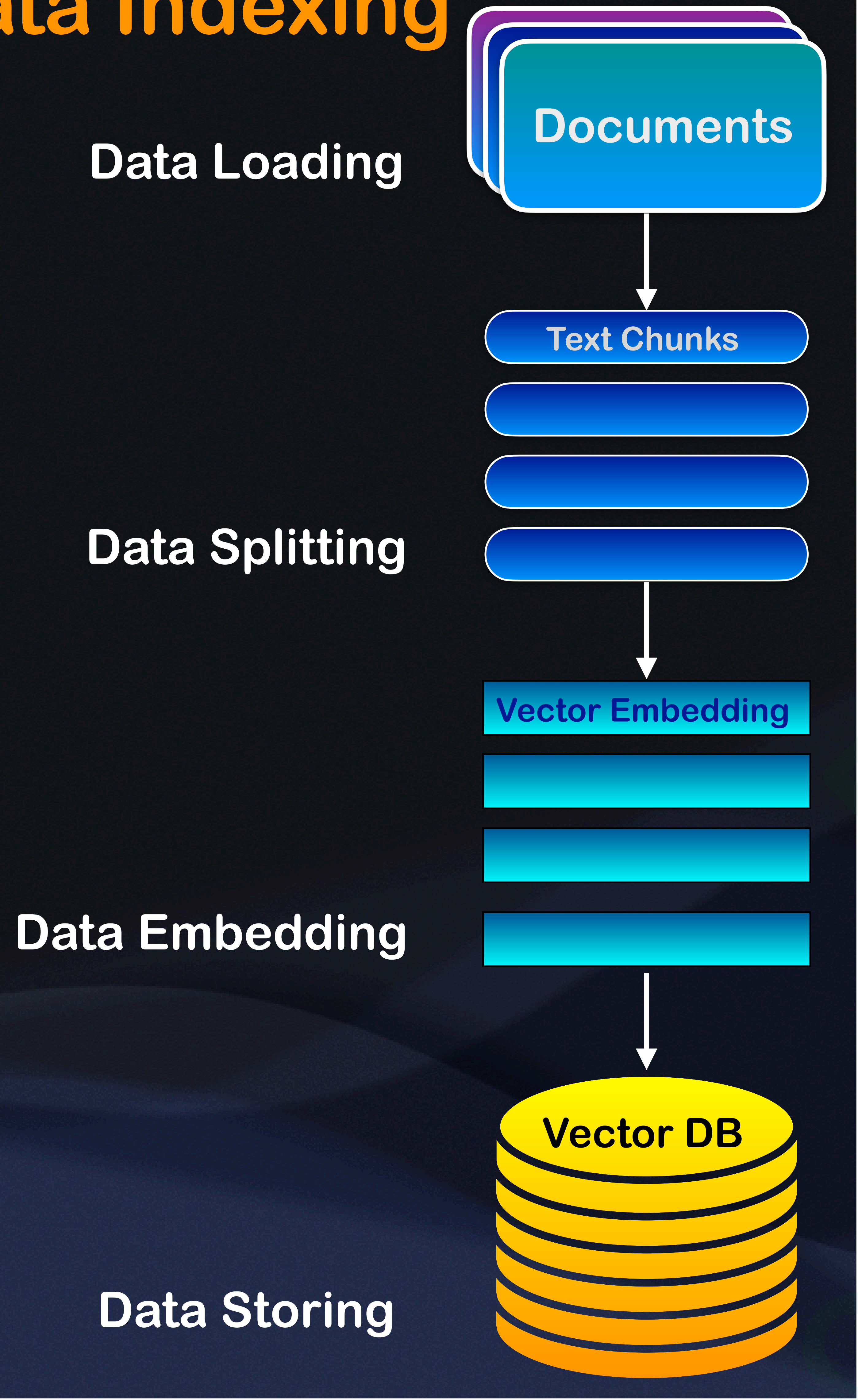
Components of LangGraph

START
State
NODE 1
State
Conditional Edge
State
NODE 4
State
Conditional Edge
State
State
State
State
NODE 2
State
NODE 3
State
END
State
State
NODE 5

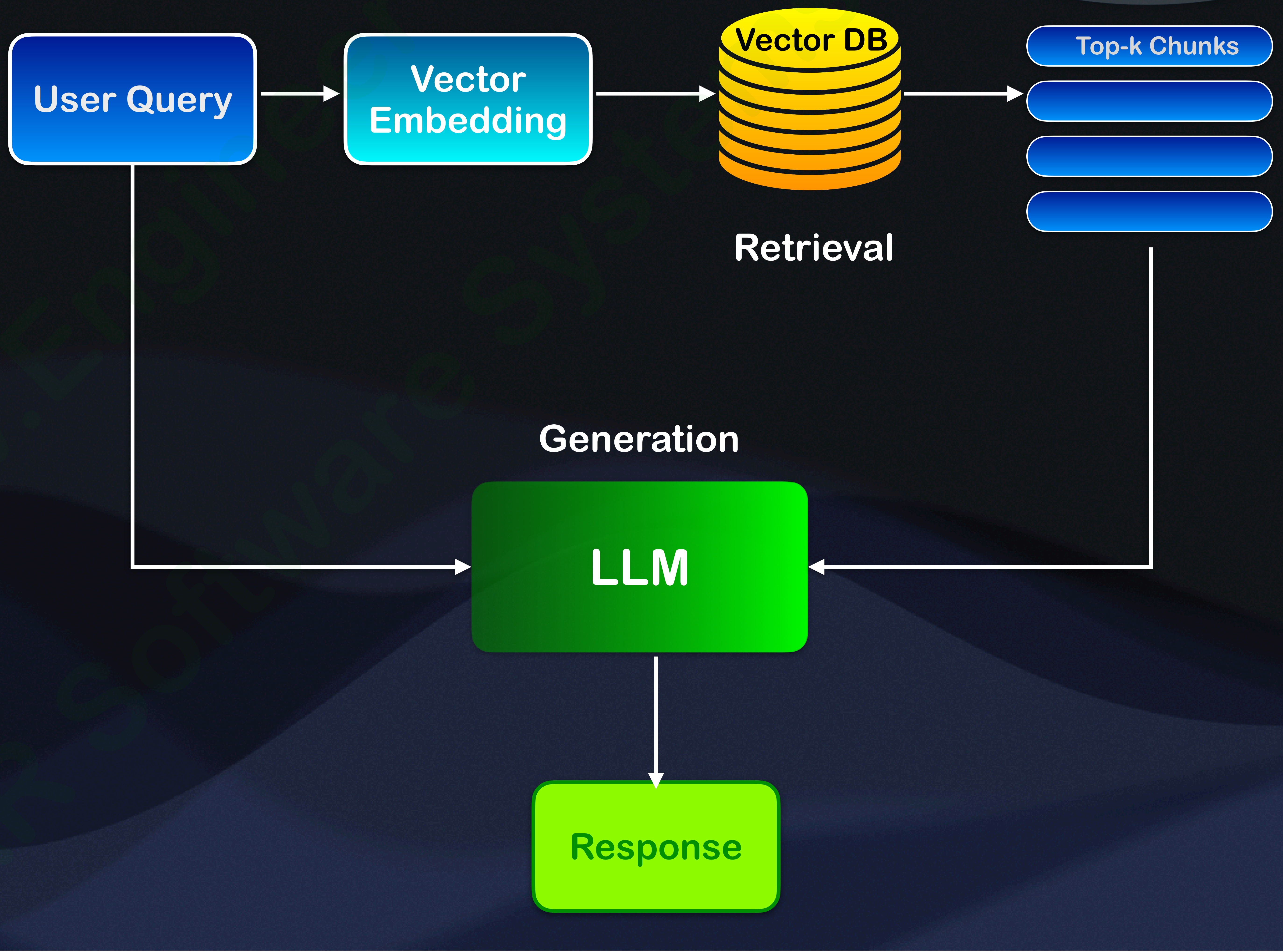
THASMAI INFOTECH

RAG Pipeline | Retrieval-Augmented Generation

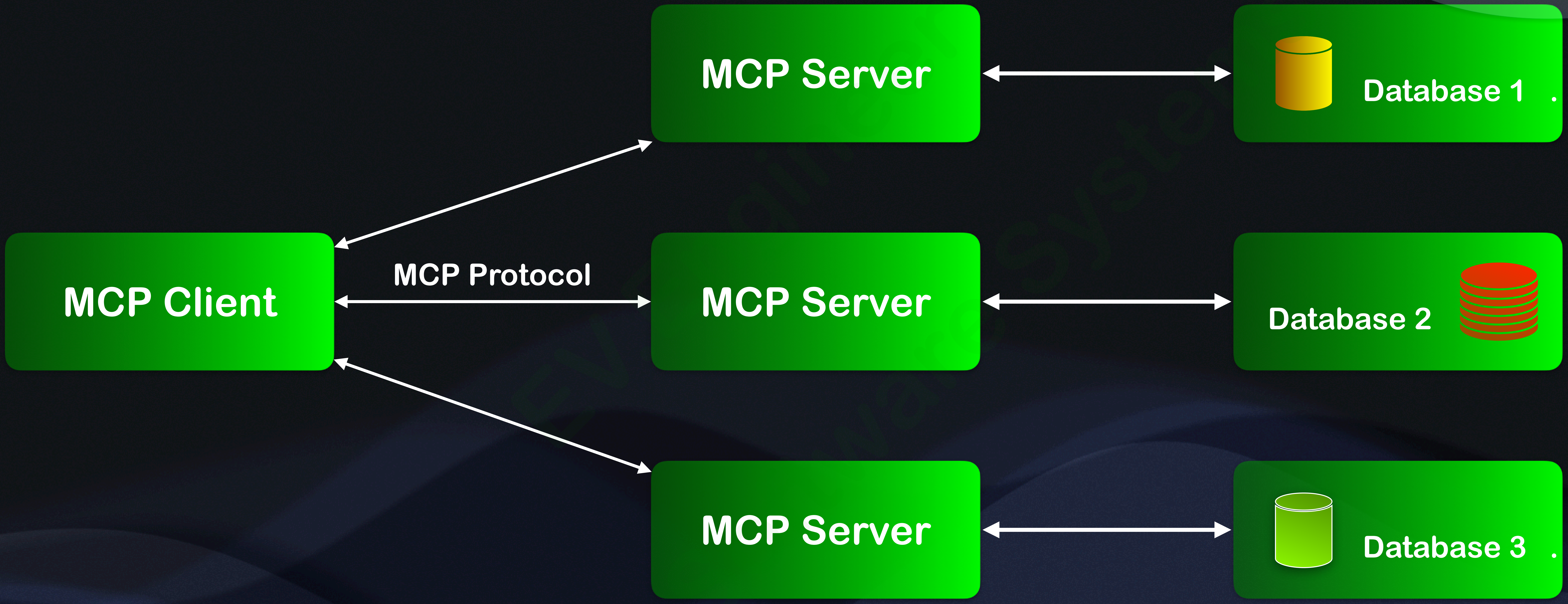
Data Indexing



Data Retrieval & Generation



MCP AI architecture



1. Introduction to Generative AI

1.1 What is Generative AI ?

1.2 Differences: AI vs ML vs DL vs GenAI

1.3 History & Evolution of LLMs

1.4 Why Generative AI matters for engineering

1.5 Key Terminologies (Token, Embedding, Context Window, Latent Space)

EV.Engineer
CAR Software Systems

2. NLP vs LLMs — Foundations

2.1 What is NLP ?

2.2 What is LLM ?

2.3 Differences between NLP & LLM

2.4 Traditional NLP pipelines vs Transformers

2.5 Limitations of classical NLP

EV Engineer
CAR Software Systems

3. Transformer Architecture

3.1 What is the Transformer model ?

3.2 Attention & Self-Attention

3.3 Encoder / Decoder / Encoder-Decoder

3.4 Positional Encoding

3.5 Why Transformers changed everything ?

EV Engineer
CPR Software Systems

4. LLM Training — How they learn?

THASMAI INFOTECH

4.1 Pre-training

4.2 Fine-tuning

4.3 Instruction tuning

4.4 RLHF

4.5 DPO / Constitutional AI

4.6 Evaluation Methods (Perplexity, BLEU, MMLU)

EV Engineer
CPR Software Systems

5. Prompt Engineering

5.1 Zero-shot

5.2 One-shot

5.3 Few-shot

5.4 Chain-of-Thought

5.5 Tree-of-Thought

5.6 ReAct (Reason + Act)

5.7 Prompt Patterns (Summarization, Extraction, Conversion)

5.8 System / User / Developer prompts

5.9 Anti-hallucination patterns

EV Engineer
CAR Software Systems

6. Embeddings & Vector Databases

6.1 What are embeddings ?

6.2 Semantic similarity

6.3 Types of Embeddings (Text, Image, Cross-Modal)

6.4 Vector DBs (FAISS, Pinecone, Chroma, Weaviate)

6.5 Chunking strategies

6.6 Indexing & retrieval evaluation

EV.Engineer
CPR Software Systems

7. Retrieval Augmented Generation (RAG)

7.1 Why RAG is needed

7.2 RAG Architecture

7.3 Retrieval → Re-ranking → Generation

7.4 RAG + Agents

7.5 RAG failure patterns & fixes

7.6 Advanced RAG

- Multi-vector RAG
- Context compression
- Graph RAG
- Fusion RAG
- Query rewriting

EV.Engineer
CPR Software Systems

8. Agents & Agentic AI

8.1 What is an AI Agent

8.2 Plan → Act → Observe → Update workflow

8.3 Tools, Actions, Observations

8.4 Memory types

8.5 Multi-Agent collaboration

8.6 Supervisors & Coordinators

8.7 Guardrails & deterministic agents

8.8 Human-in-the-loop agents

8.9 Popular Agent Frameworks (**LangChain** Agents, **AutoGen**, **CrewAI**)

8.4 Memory types

- Short-term
- Long-term
- Vector Memory

9. LangGraph — Orchestrated AI Workflows

9.1 What is LangGraph

9.2 Why LangGraph is needed

9.3 Nodes, Edges, State, Memory

9.4 Control Flow Patterns

9.5 Multi-agent graphs

9.6 Checkpointing & State persistence

9.7 Streaming responses

9.8 LangGraph vs LangChain Agents

9.9 RAG + LangGraph

9.10 Production workflows using LangGraph

9.4 Control Flow Patterns

- Conditional edges
- Loops
- Retries
- Branching

SPV.Engineer
SPR Software Systems

10. Model Context Protocol (MCP)

10.1 What is MCP ?

10.2 Why MCP exists ?

10.3 MCP vs OpenAI Tool Calling vs LangChain Tools

10.4 How LLMs use tools via MCP

10.5 Building MCP tools

10.6 Integrating MCP with LangGraph

10.7 Real-world use cases (file tools, database tools, API tools)

EV.Engineer
Software Systems

11. Multimodal GenAI

11.1 Text + Image models

11.2 Image generation models

11.3 Vision-Language models

11.4 Audio generation

11.5 Speech-to-text (Whisper)

11.6 Video generation (Sora, Runway)

11.7 Document understanding (PDF → JSON)

EV.Engineer
CPR Software Systems

12. GenAI for Engineering

12.1 Auto code generation

12.2 Test case generation

12.3 Log summarization

12.4 Debug assistance

12.5 Documentation creation

12.6 Firmware analysis

12.7 EV Battery analytics using LLMs

12.8 BLE packet troubleshooting using AI agents

EV Engineer
CAR Software Systems

13. Model Deployment, Inference & Optimization

13.1 Running LLMs locally (Ollama)

13.2 Quantization (GGUF 4-bit, 8-bit)

13.3 GPU/TPU/CPU inference

13.4 Containerized inference (Docker, FastAPI)

13.5 Scaling LLM apps

13.6 Caching, batching, streaming

CKV.Engineer
CKR Software Systems

14. Fine-Tuning & Custom Models

THASMAI INFOTECH

14.1 Full fine-tuning vs LoRA vs Q-LoRA

14.2 Dataset creation

14.3 Synthetic data generation

14.4 Evaluation metrics

14.5 Domain-specific LLMs (EV, automotive, IoT)

14.6 Safety considerations

EV.Engineer
CPR Software Systems

15. Safety, Governance & Ethics

15.1 Hallucination control

15.2 Copyright issues

15.3 PII & privacy protection

15.4 Bias and fairness

15.5 Red teaming & adversarial testing

15.6 Enterprise guardrails

15.7 Audit, logging, traceability

EV.Engineer
CPR Software Systems

16. Evaluation & Testing

16.1 LLM testing

16.2 RAG testing

16.3 Agent safety testing

16.4 Behavioral testing

16.5 Observability (LangSmith, WandB)

16.6 Monitoring & debugging tools

EV.Engineer
CPR Software Systems

17. Hands-On (Python + APIs)

17.1 Calling OpenAI / Gemini APIs

17.2 Building a chatbot

17.3 Token counting

17.4 JSON mode / structured outputs

17.5 Building RAG with Chroma

17.6 Building LangGraph workflows

17.7 Building MCP tools

17.8 Deploying with FastAPI + Docker

GV.Engineer
CPR Software Systems

18. Capstone Projects

18.1 Custom RAG for EV battery knowledge

18.2 LangGraph-based BLE unlock workflow

18.3 Smart lock diagnostic agent

18.4 Multi-agent code review system

18.5 EV Battery Fire Prevention AI Assistant

EV.Engineer
CPR Software Systems

Thank you

Contact Us for customised training materials & source code

EV Engineer
CAR Software Systems

Sudarshana Karkala

Director of Engineering

Thasmai Infotech Private Limited

✆ +91 9845561518 | ✉ ThasmaiInfotech @ gmail.com | ThasmaiInfotech.com