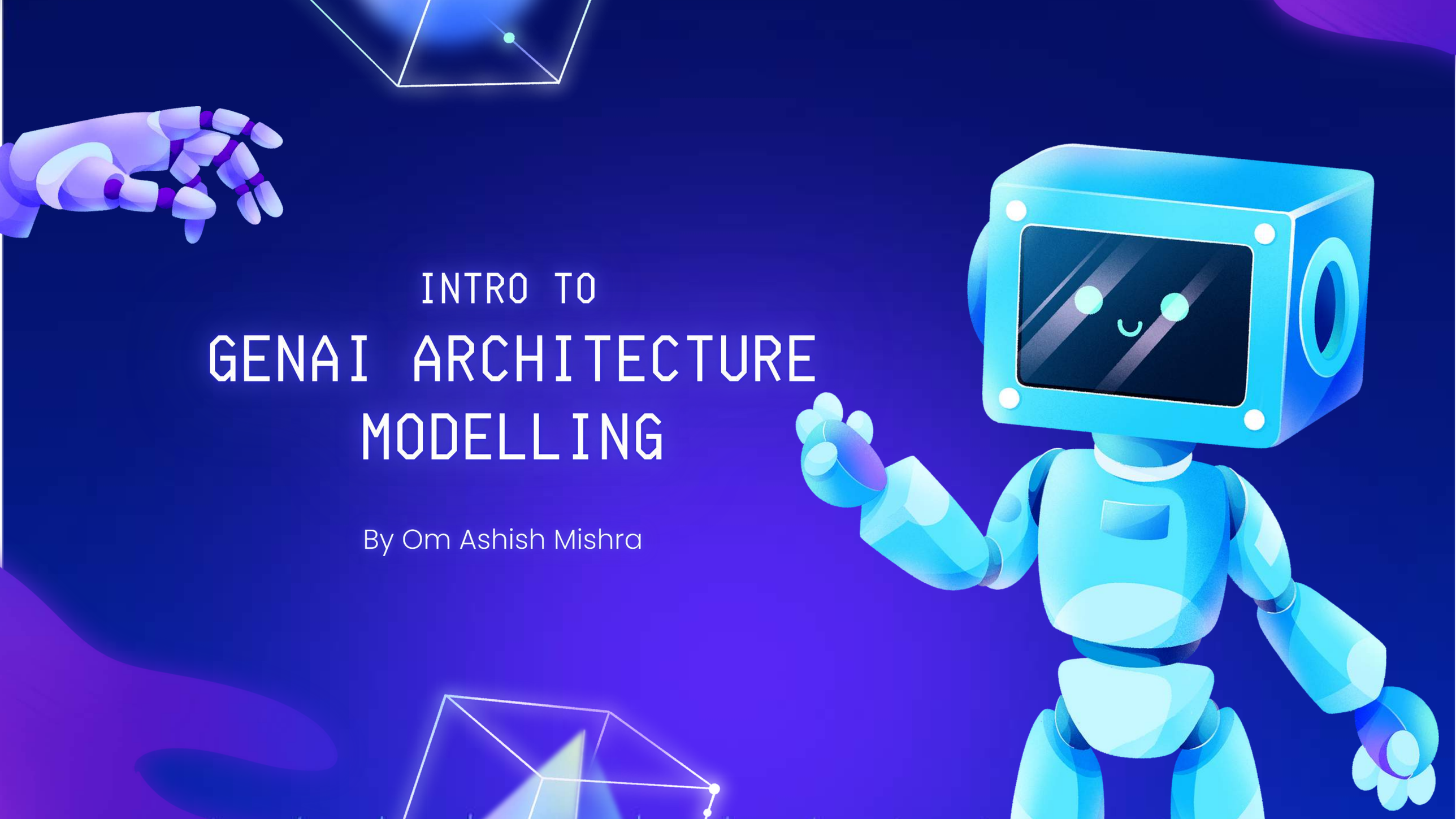
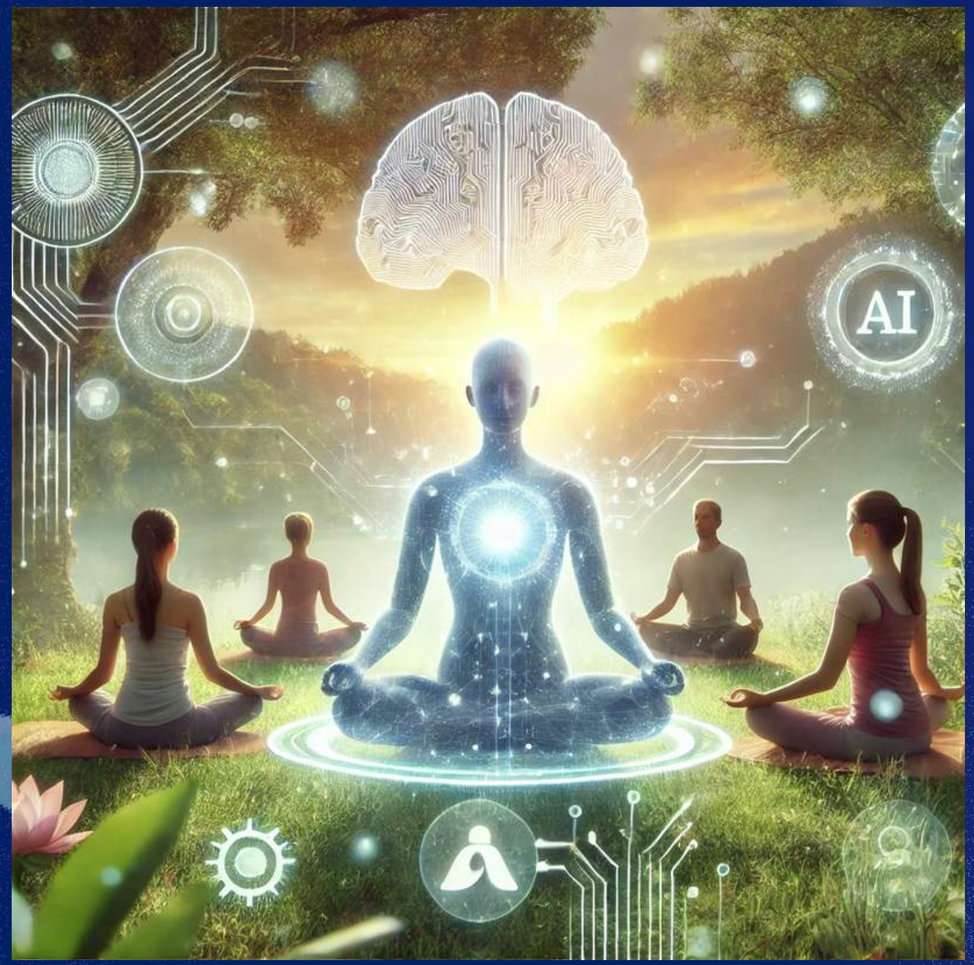

INTRO TO
GENAI ARCHITECTURE MODELLING
By Om Ashish Mishra

PLAN OF ACTION

1. Understanding GenAI
2. Basic Architecture Flow
3. Types of Design Patterns
4. Takeways



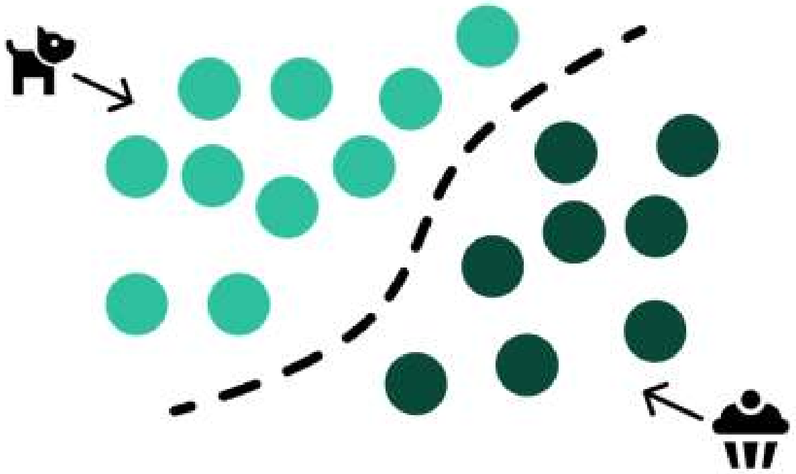
UNDERSTANDING GENAI

DISCRIMINATIVE MODEL

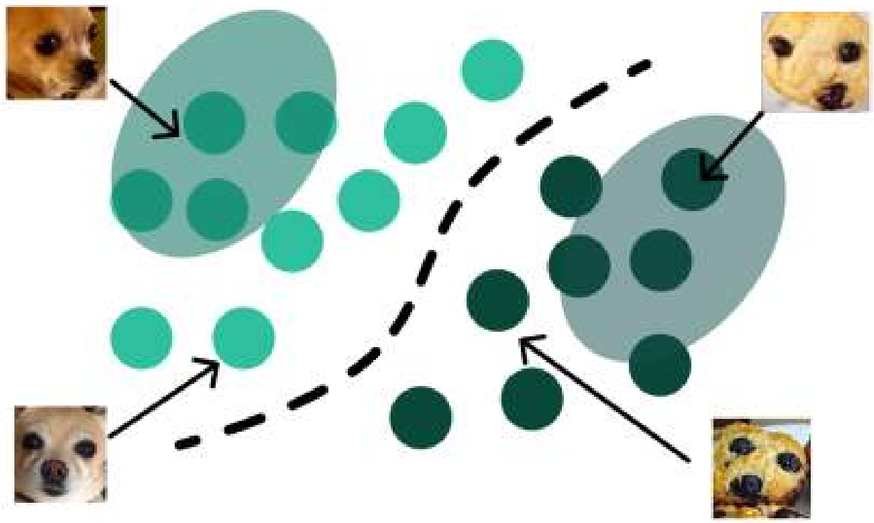
VS

GENERATIVE MODEL

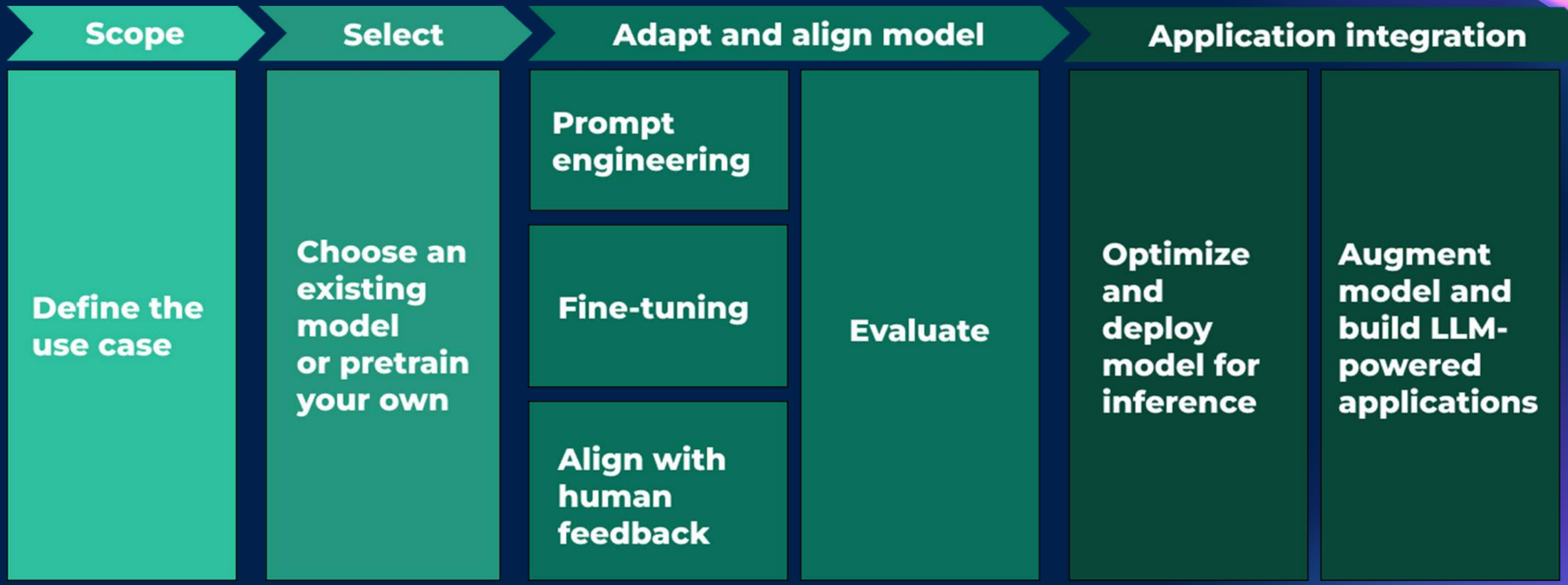
Learns decision boundaries between classes in the data by learning differences i.e what it is vs what it isn't



Holistic model learns the overall structure and distribution of the data, i.e. only what it is



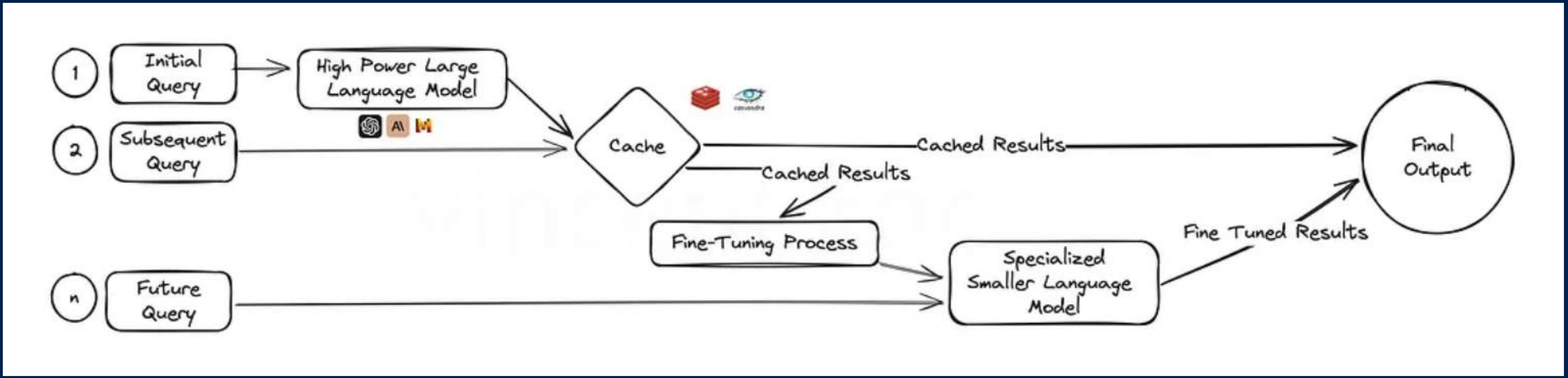
GENERATIVE AI SOLUTION LIFECYCLE



DIFFERENT GEN AI DESIGN PATTERNS

1. Layered Caching Strategy Leading To Fine-Tuning
2. Multiplexing AI Agents For A Panel Of Experts
3. Fine-Tuning LLM’s For Multiple Tasks
4. Blending Rules Based & Generative
5. Utilizing Knowledge Graphs with LLM’s
6. Swarm Of AI Agents
7. Modular Monolith LLM Approach With Composability
8. Approach To Memory Cognition For LLM’s
9. Red & Blue Team Dual-Model Evaluation

1. LAYERED CACHING STRATEGY LEADING TO FINE-TUNING



Applications

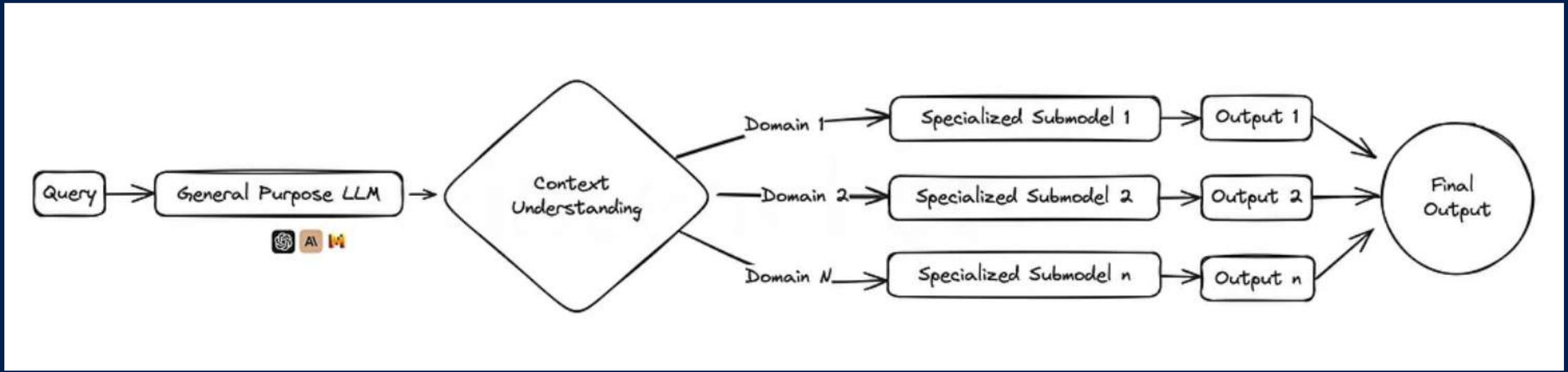
Where Precision is paramount

- Customer Service
- Personalized Content Generation

Tools

- GPTCache
- Redis
- Memcache
- Apache Cassandra

2. MULTIPLEXING AI AGENTS FOR A PANEL OF EXPERTS



Applications

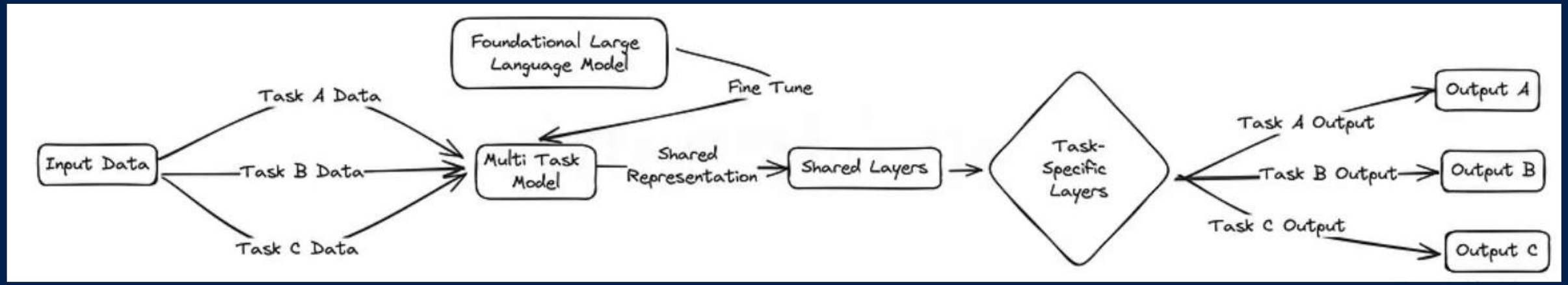
Complex problem Solving scenarios where different aspects of a problem require different expertise

- Personalized Medical Diagnosis
- Comprehensive Research & Analysis

Tools

- PHI-2
- TinyLlama

3. FINE-TUNING LLM'S FOR MULTIPLE TASKS



Applications

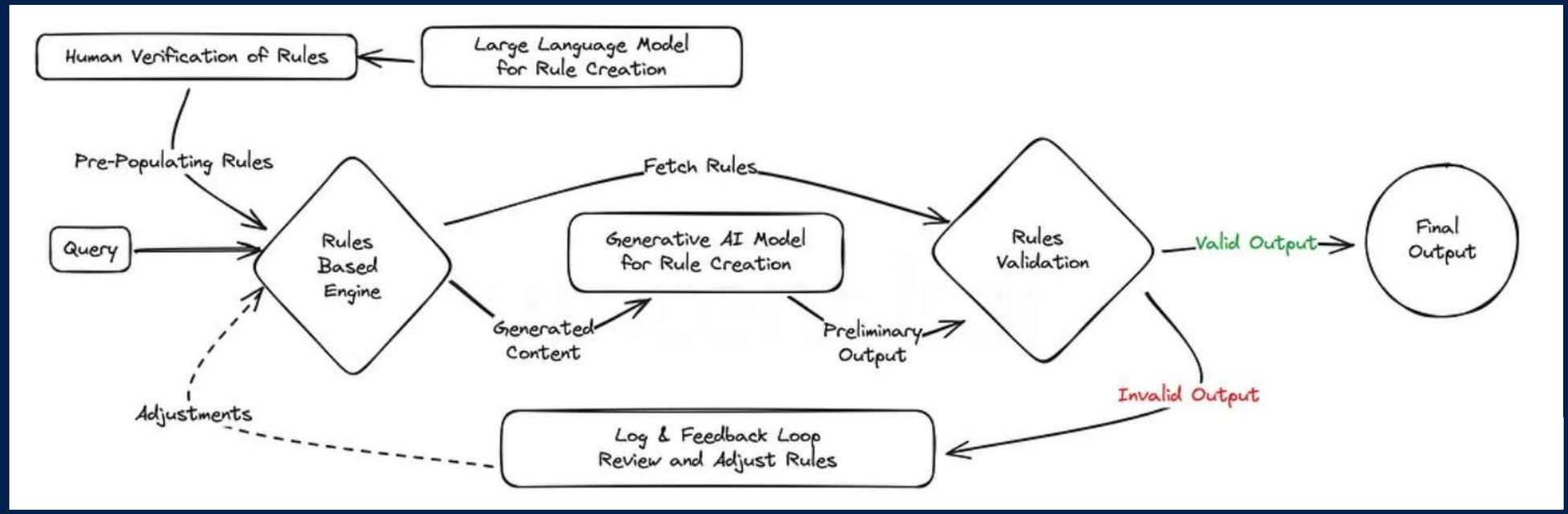
Platforma that need to handle a variety of tasks with a high degree of competence

- Virtual Assistant
- AI powered research tool

Tools

- DeepSpeed
- Hugging Face's Transformer Library

4. BLENDING RULES BASED & GENERATIVE

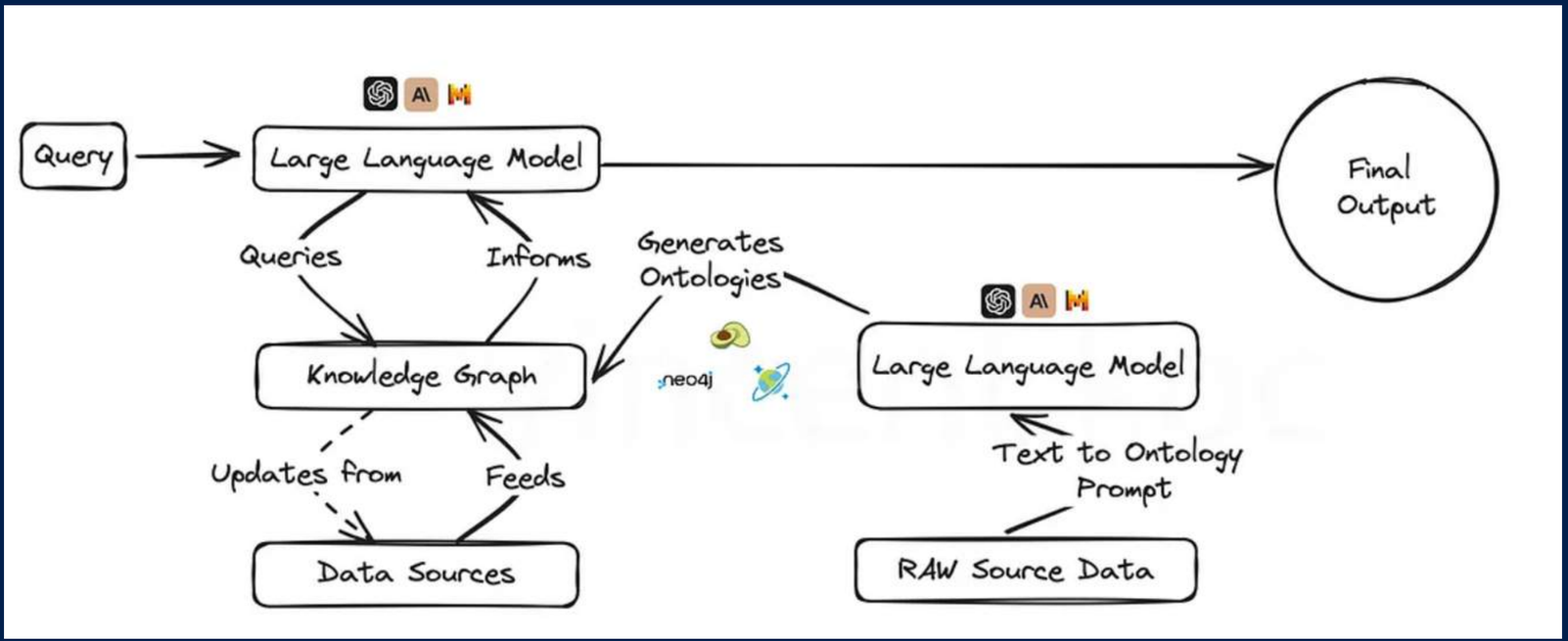


Applications

Where outputs must adhere to strigent standards or regulations, ensuring the AI remains within the bounds of desired parameters while still being able to innovate and engage

- Phone call IVR
- Traditional Chatbot rules based

5. UTILIZING KNOWLEDGE GRAPHS WITH LLM'S



Applications

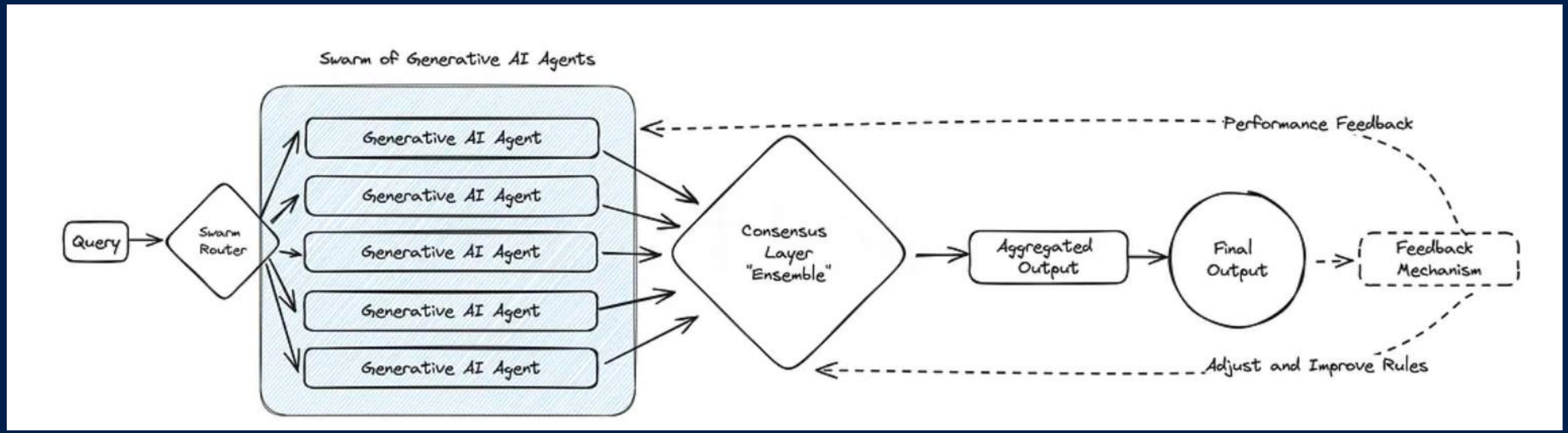
Truth and accuracy are non-negotiable

- Education Content Creation
- Medical Advice

Tools

- Neo4j
- Amazon Neptune
- Azure Cosmos
- ArangoDB
- Google Enterprise Knowledge Graph API
- PyKEEN Datasets
- Wikidata

6. SWARM OF AI AGENTS



Applications

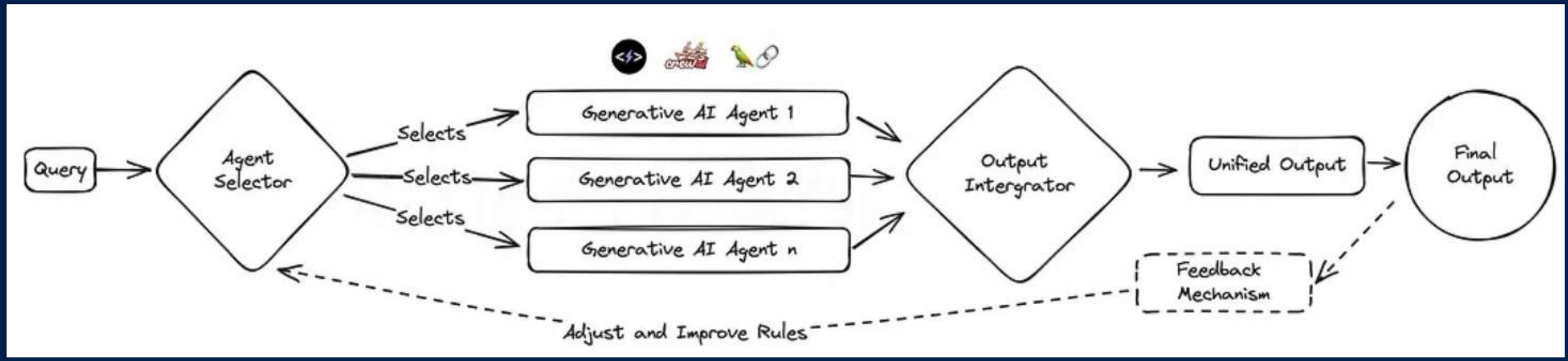
Advantageous in scenarios that require a breath of creative solutions or when navigating complex datasets

- Reviewing a research paper from a multiple “experts” point of view
- Assessing customer interactions for many usecases at once from fraud to offers

Tools

- Apache Kafka

7. MODULAR MONOLITH LLM APPROACH WITH COMPOSABILITY



Applications

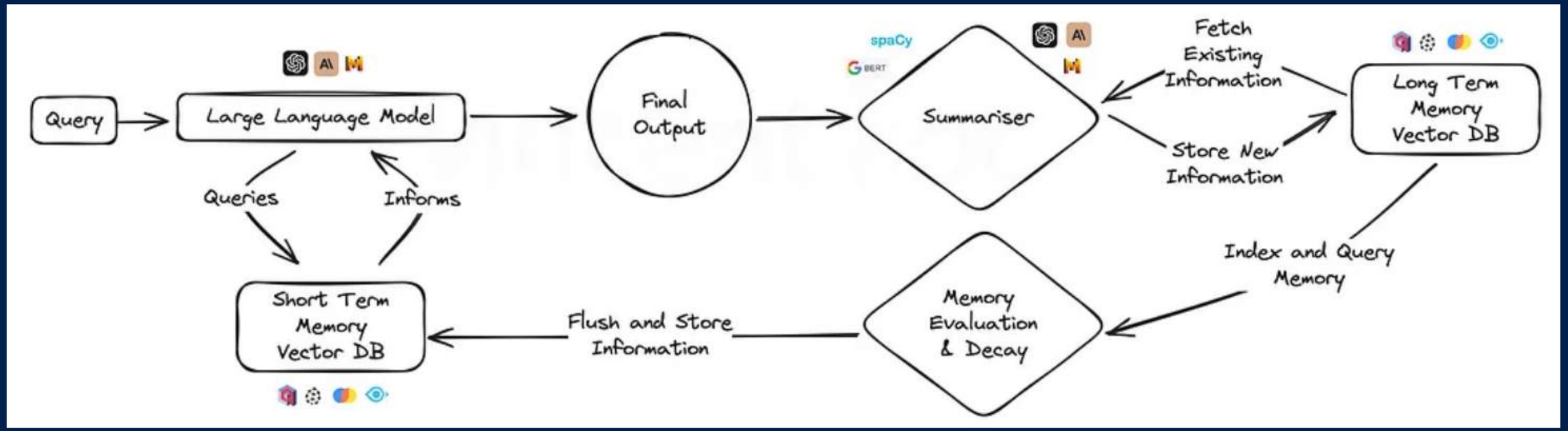
Tailor-made solutions for varying customer interactions or product needs

- Domain Specific Problem Solving

Tools

- CrewAI
- Langchain
- Microsoft Autogen
- SuperAGI

8. APPROACH TO MEMORY COGNITION FOR LLM'S



Applications

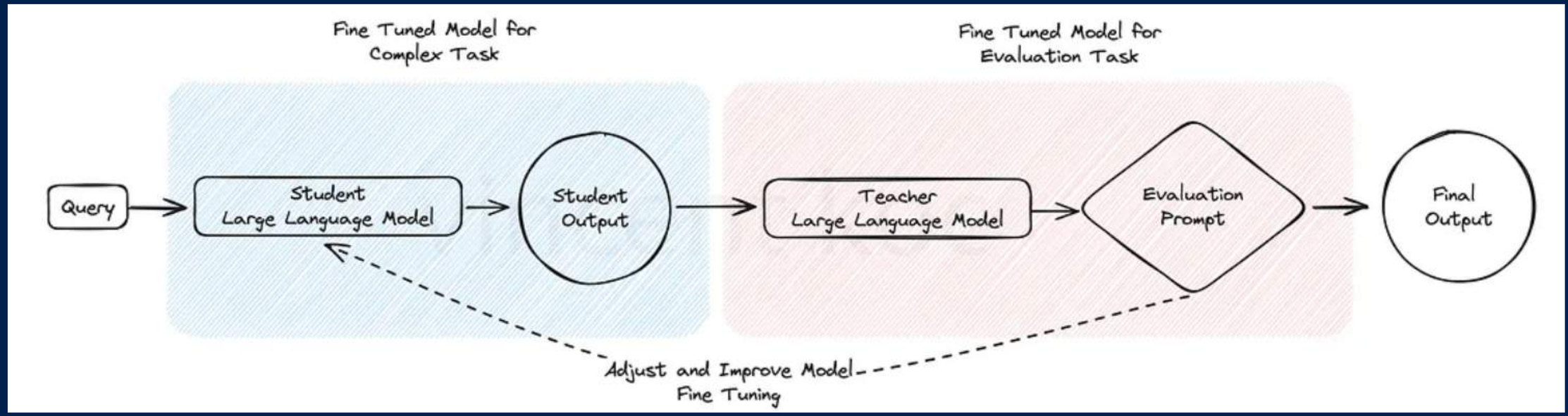
This approach introduces an element of human-like memory to AI, allowing models to recall and build upon previous interactions for more nuanced responses

- Adaptive Learning Platform
- Enhance Personalized Customer Service Chatbot

Tools

- spaCy
- BART lang model
- MemGPT

9. RED & BLUE TEAM DUAL-MODEL EVALUATION

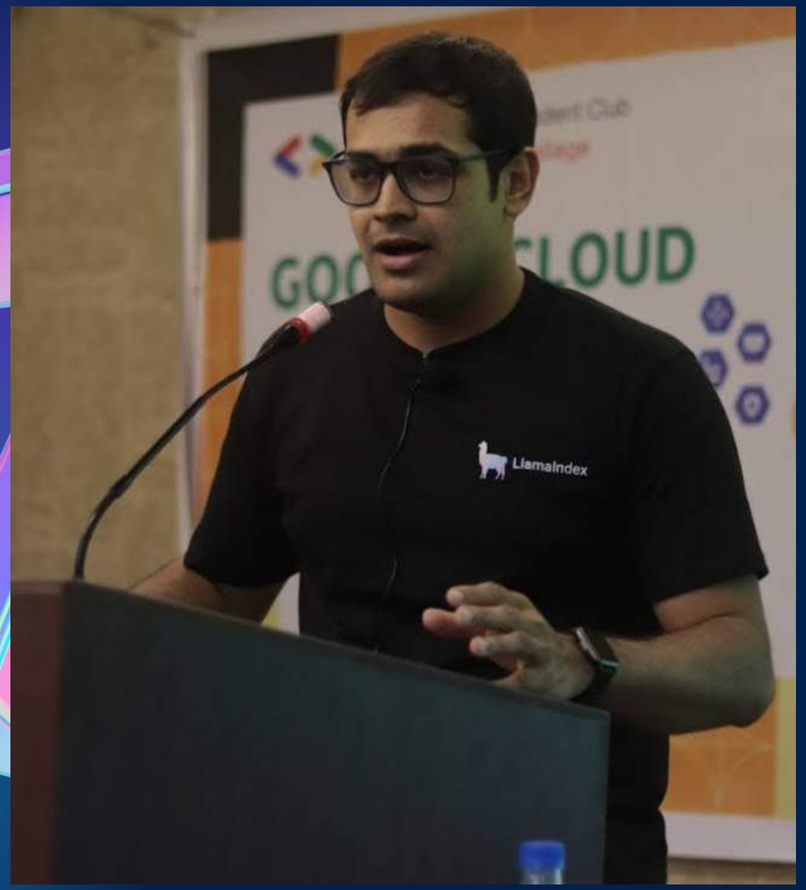


Applications

This dual-model setup is excellent for quality control, making it highly applicable in content generation platforms where credibility and accuracy are vital

- News Aggregation
- Education Material Production

Thank You



Om Ashish Mishra

https://www.omashish.com

https://twitter.com/ommishra100

https://www.linkedin.com/in/om-ashish-mishra/

https://medium.com/@OmAshishMishra