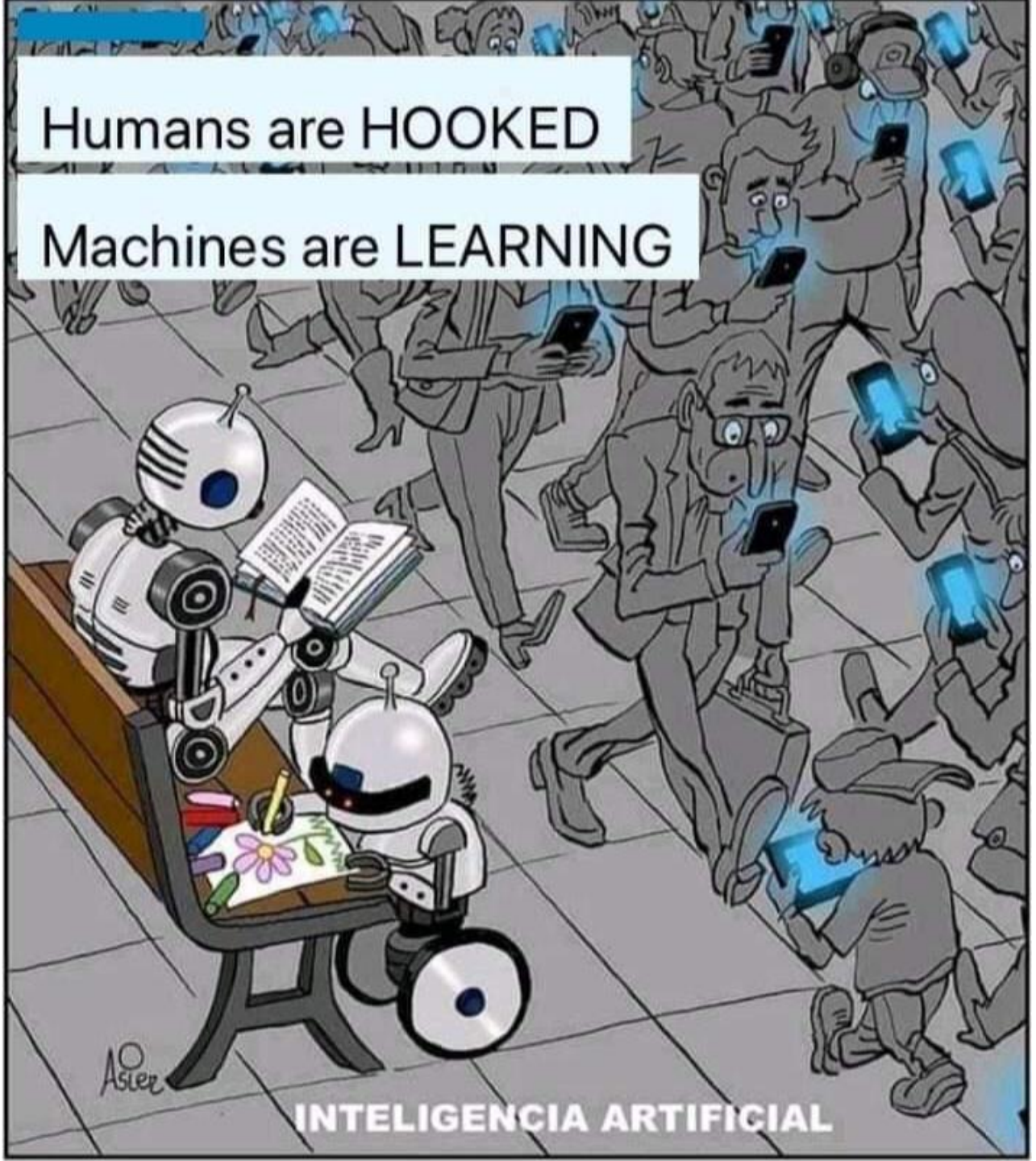
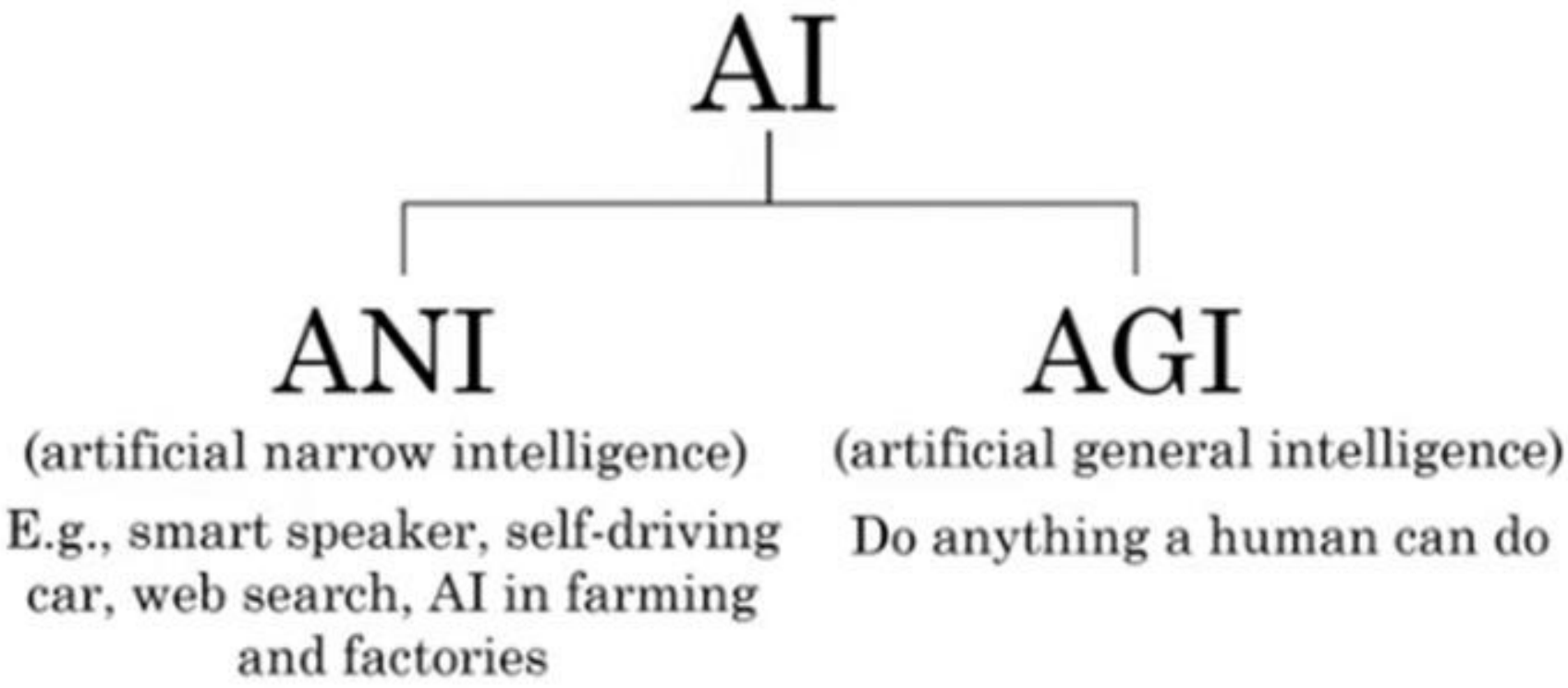


Generative AI

Naveen Kumar Bhansali
Co-Founder BlitzAI | Adjunct Faculty @ IIM Bangalore

AGI Vs. ANI


AI
ANI
(artificial narrow intelligence)
E.g., smart speaker, self-driving car, web search, AI in farming and factories
AGI
(artificial general intelligence)
Do anything a human can do

AGI Levels

AGI tier list

-- Google DeepMind

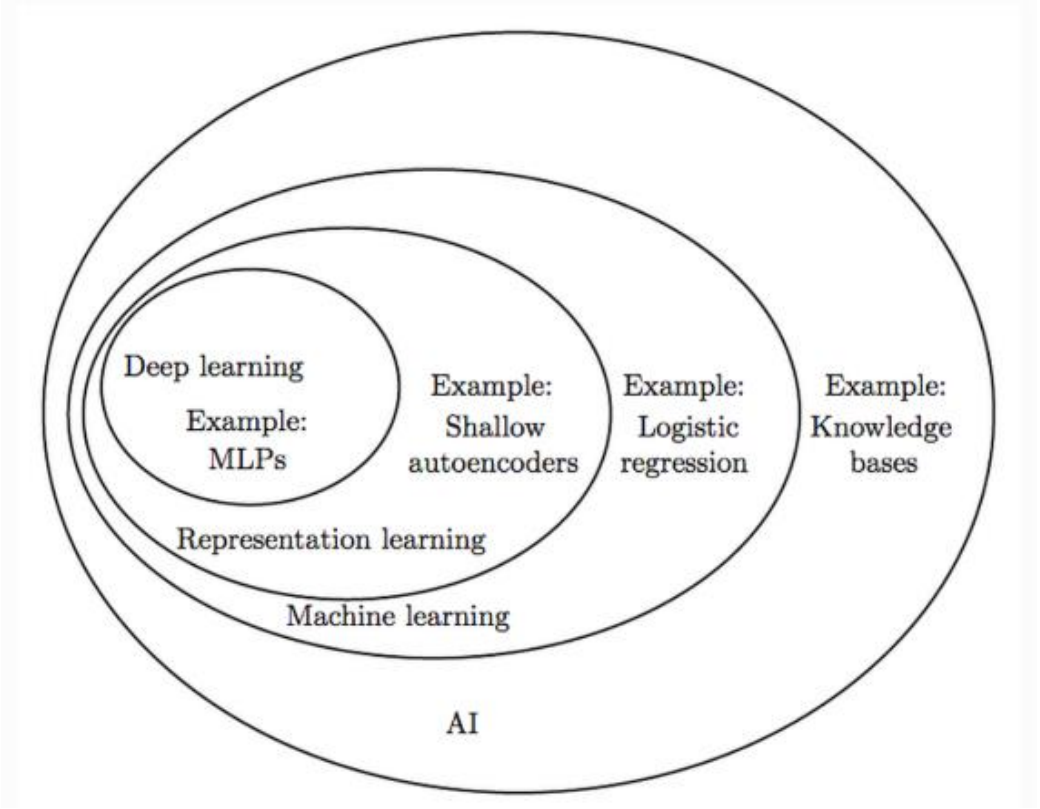
Levels of AGI

Performance (rows) x **Generality** (columns)	**Narrow** *clearly scoped task or set of tasks*	**General** *wide range of non-physical tasks, including metacognitive abilities like learning new skills*
Level 0: No AI	**Narrow Non-AI** calculator software; compiler	**General Non-AI** human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging *equal to or somewhat better than an unskilled human*	**Emerging Narrow AI** GOFAI[4]; simple rule-based systems, e.g., SHRDLU (Winograd, 1971)	**Emerging AGI** ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023)
Level 2: Competent *at least 50th percentile of skilled adults*	**Competent Narrow AI** toxicity detectors such as Jigsaw (Das et al., 2022); Smart Speakers such as Siri (Apple), Alexa (Amazon), or Google Assistant (Google); VQA systems such as PaLI (Chen et al., 2023); Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	**Competent AGI** not yet achieved
Level 3: Expert *at least 90th percentile of skilled adults*	**Expert Narrow AI** spelling & grammar checkers such as Grammarly (Grammarly, 2023); generative image models such as Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022)	**Expert AGI** not yet achieved
Level 4: Virtuoso *at least 99th percentile of skilled adults*	**Virtuoso Narrow AI** Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016, 2017)	**Virtuoso AGI** not yet achieved
Level 5: Superhuman *outperforms 100% of humans*	**Superhuman Narrow AI** AlphaFold (Jumper et al., 2021; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023)	**Artificial Superintelligence (ASI)** not yet achieved

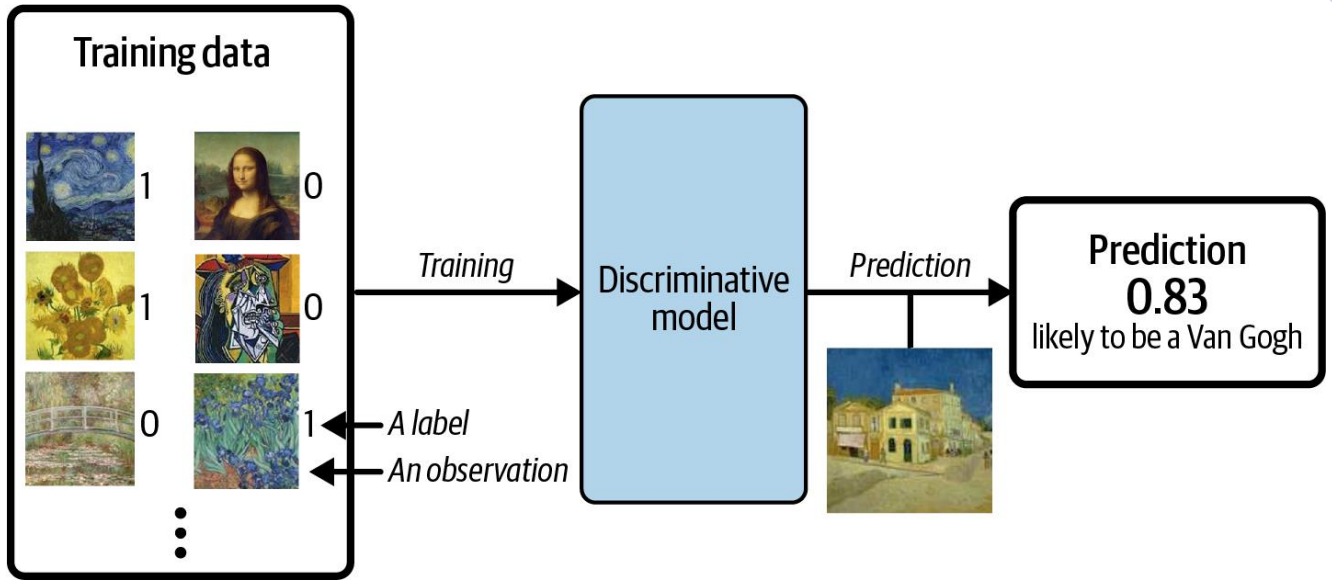
Agenda

- Generative AI - Large Language Models / Large Multimodal Models
- Applications
- Case Studies
- Generative AI project lifecycle
  - Cheat Sheet
  - Prompt Engineering
  - LLM Powered Applications

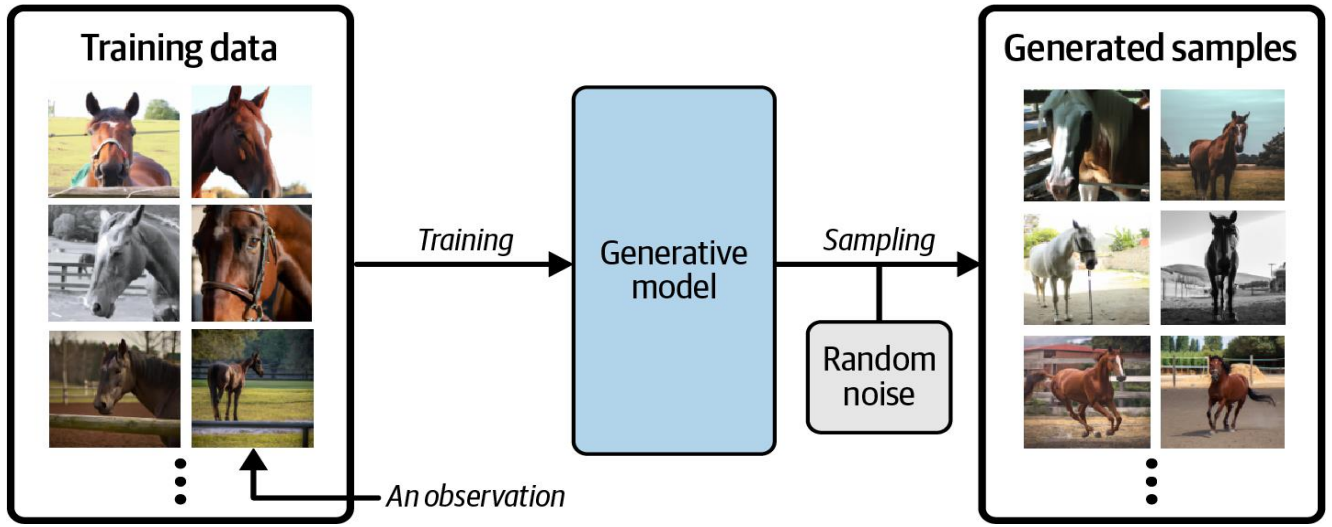
AI / ML / DL / GenAI – How are they related?

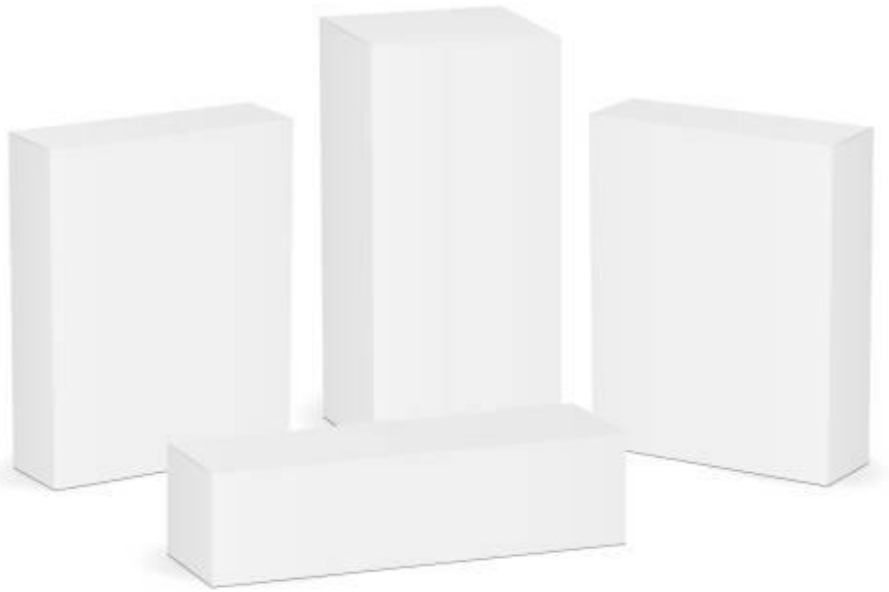


Discriminative AI



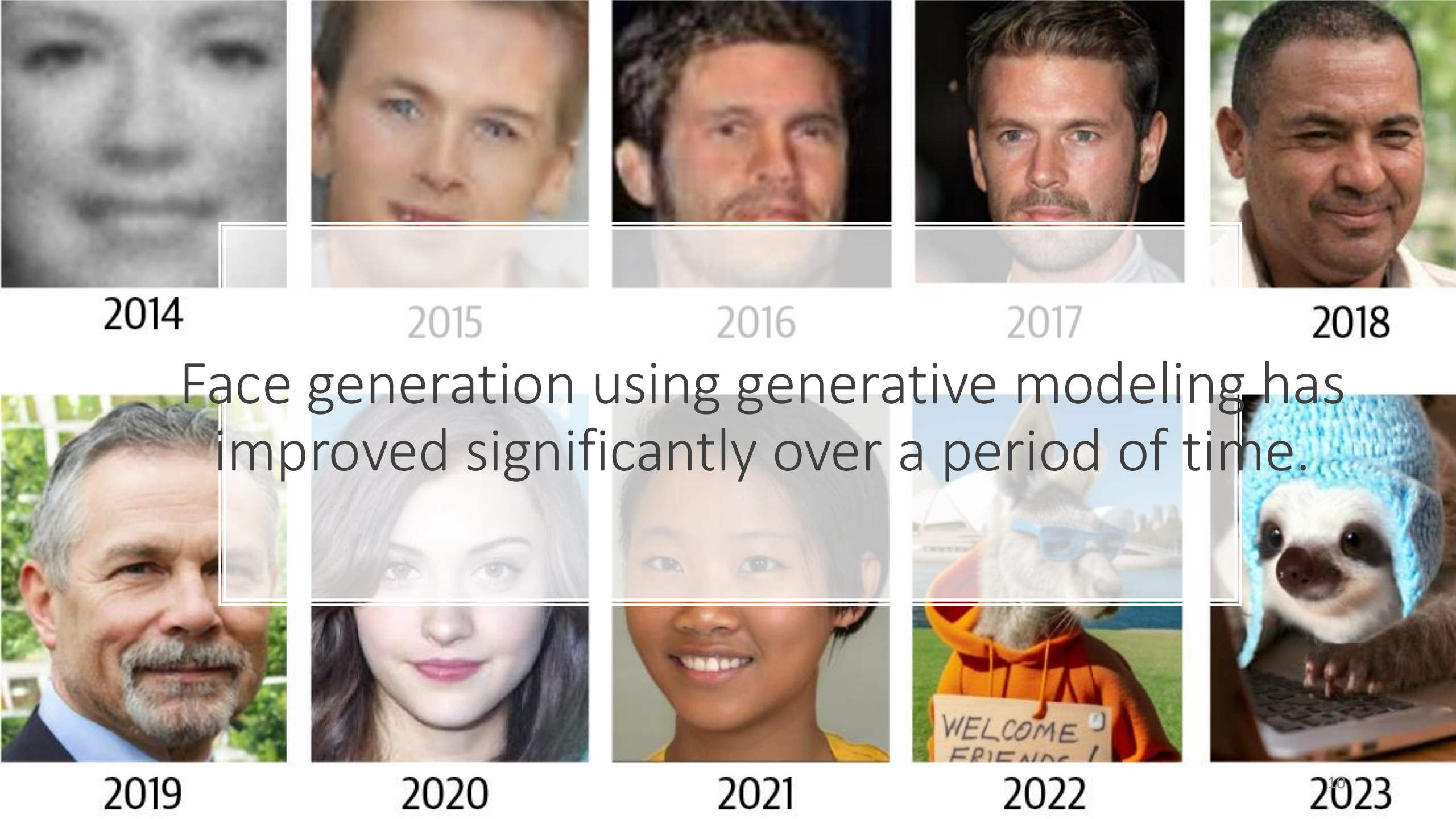
Generative AI





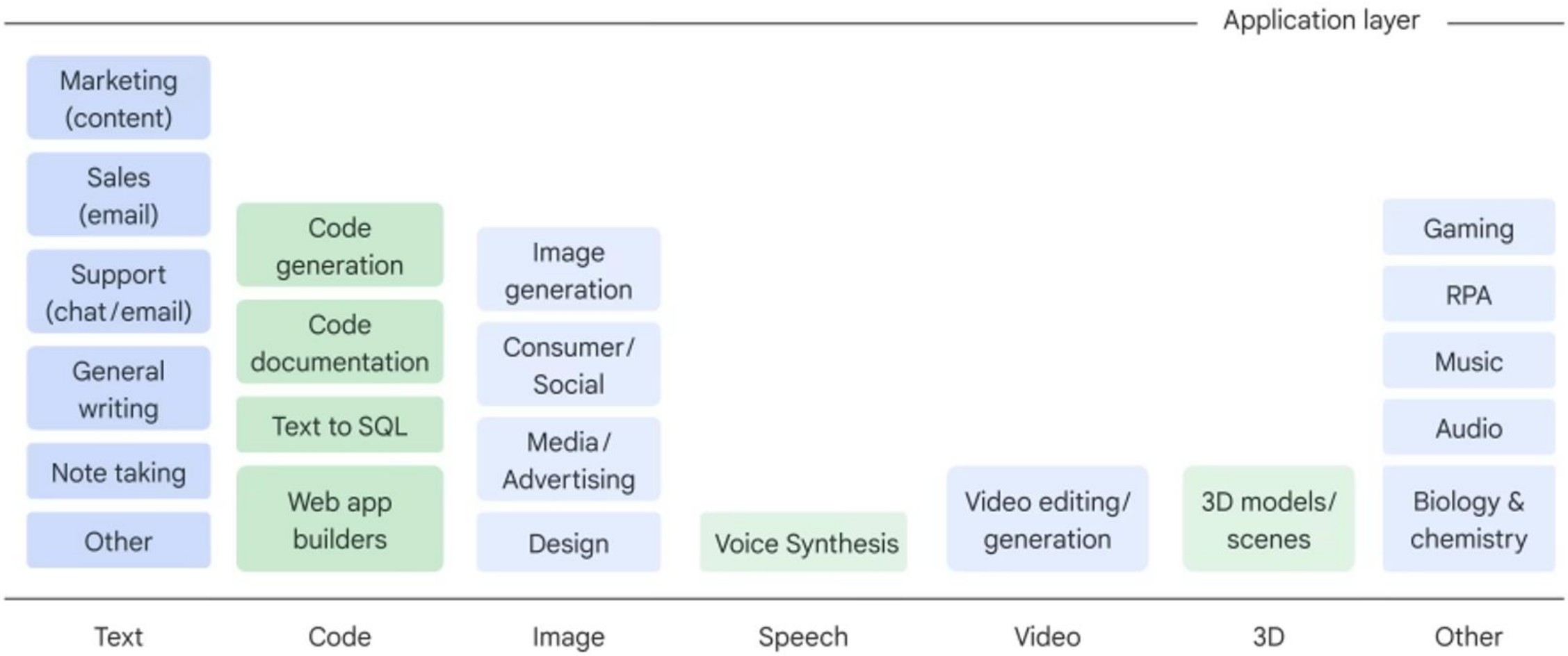
Core Principle – Representation Learning

Applications



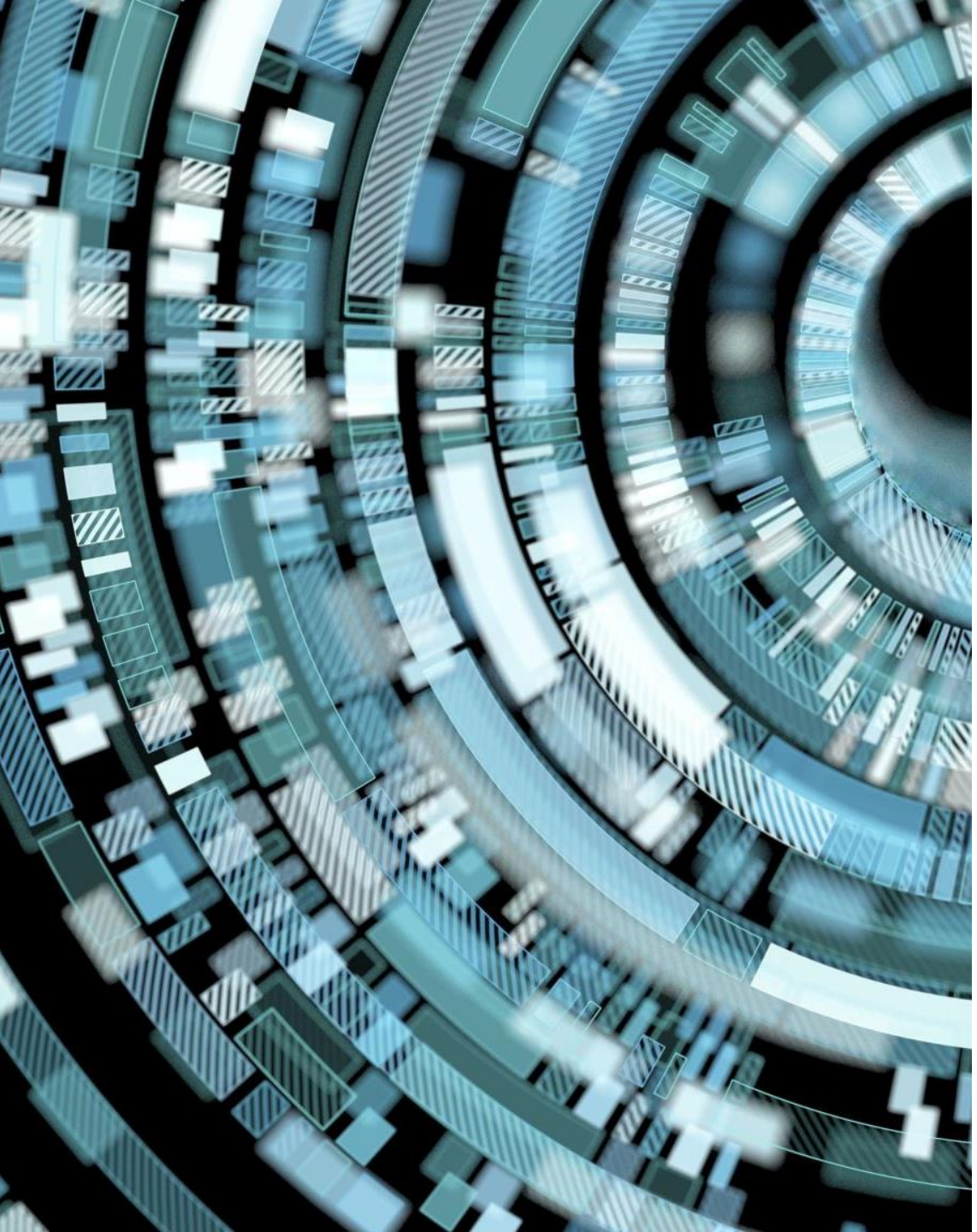
Face generation using generative modeling has improved significantly over a period of time.

Generative AI Application Landscape



Large

Language Models (LLMs)



LLMs

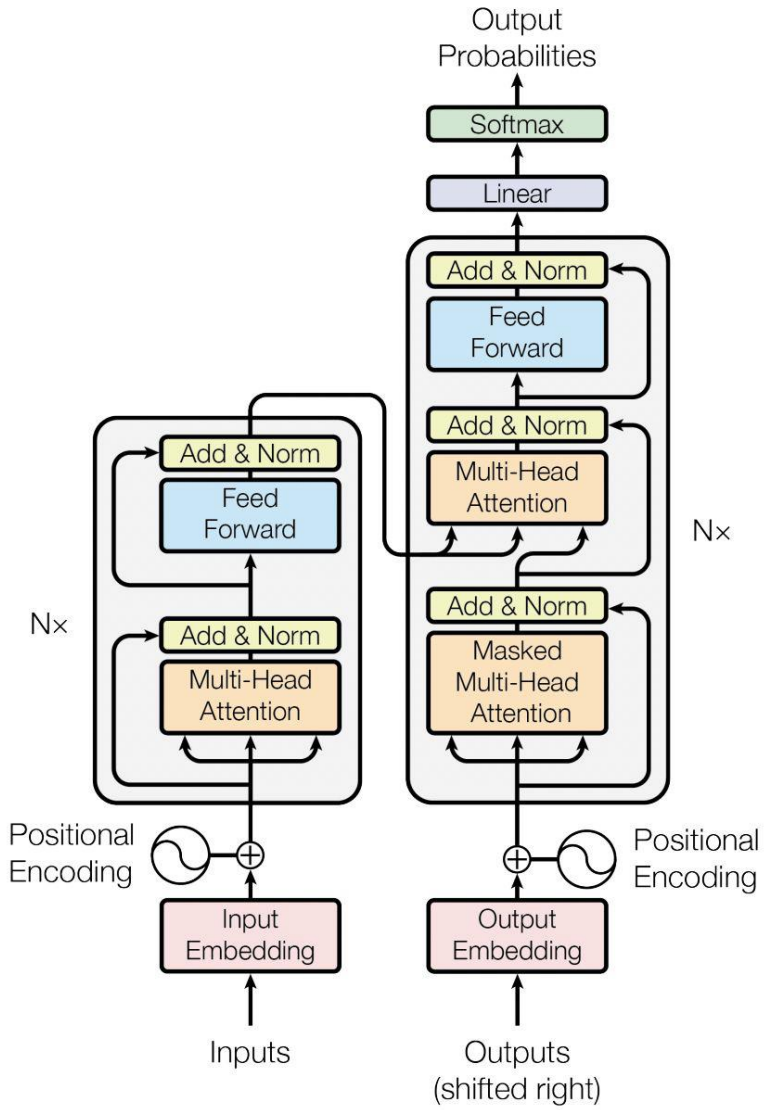
- A Language Model
- Consisting of Neural network with many parameters
  - (typically, billions to trillions of weights or more)
- Trained on large quantities (trillions of words) of unlabelled text.
  - Wikipedia, GitHub, Common Crawl, The Pile etc.
- Using self-supervised learning

LLMs – Based on?

Attention is all you need

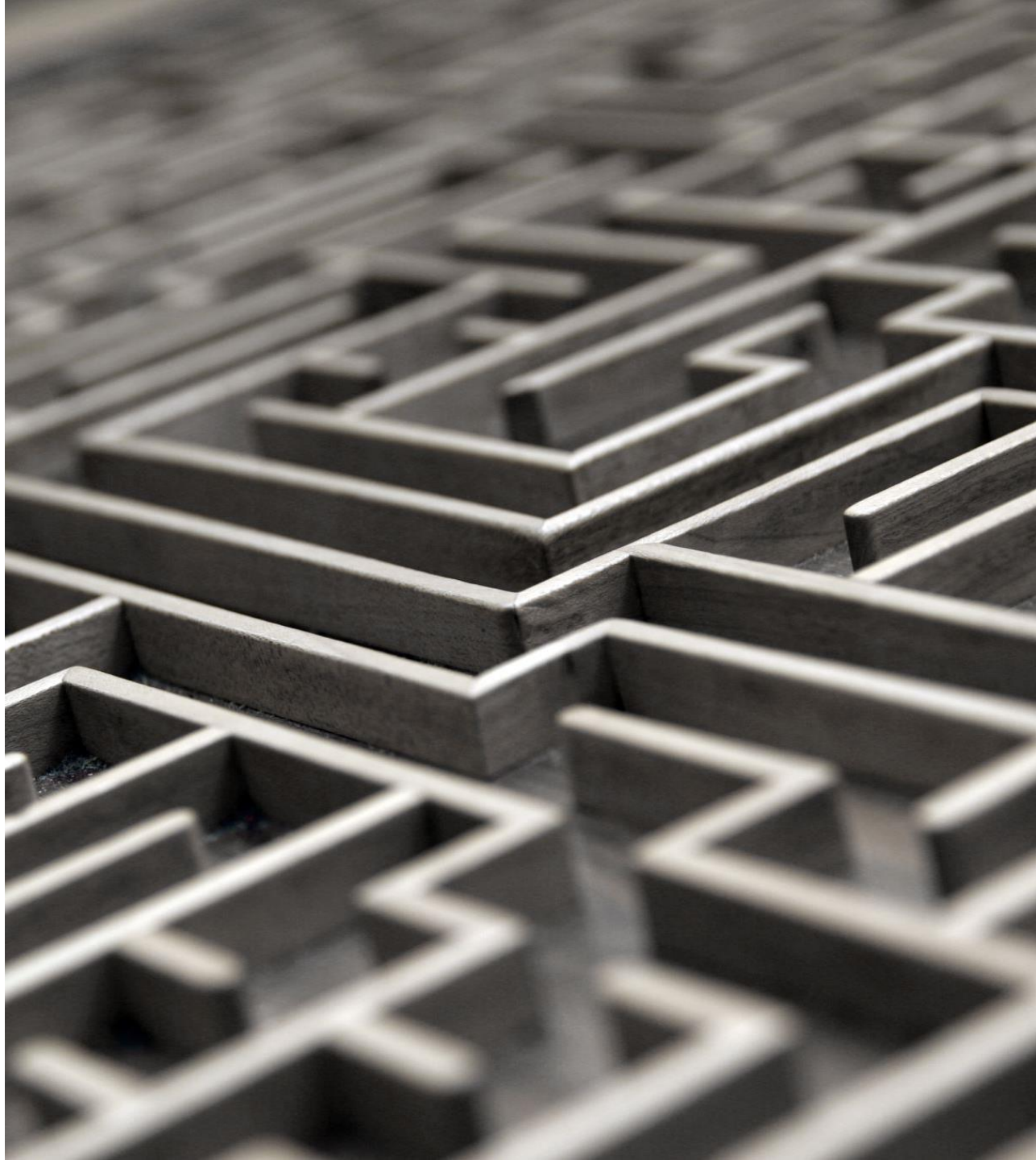
https://ig.ft.com/generative-ai/

Transformers architecture

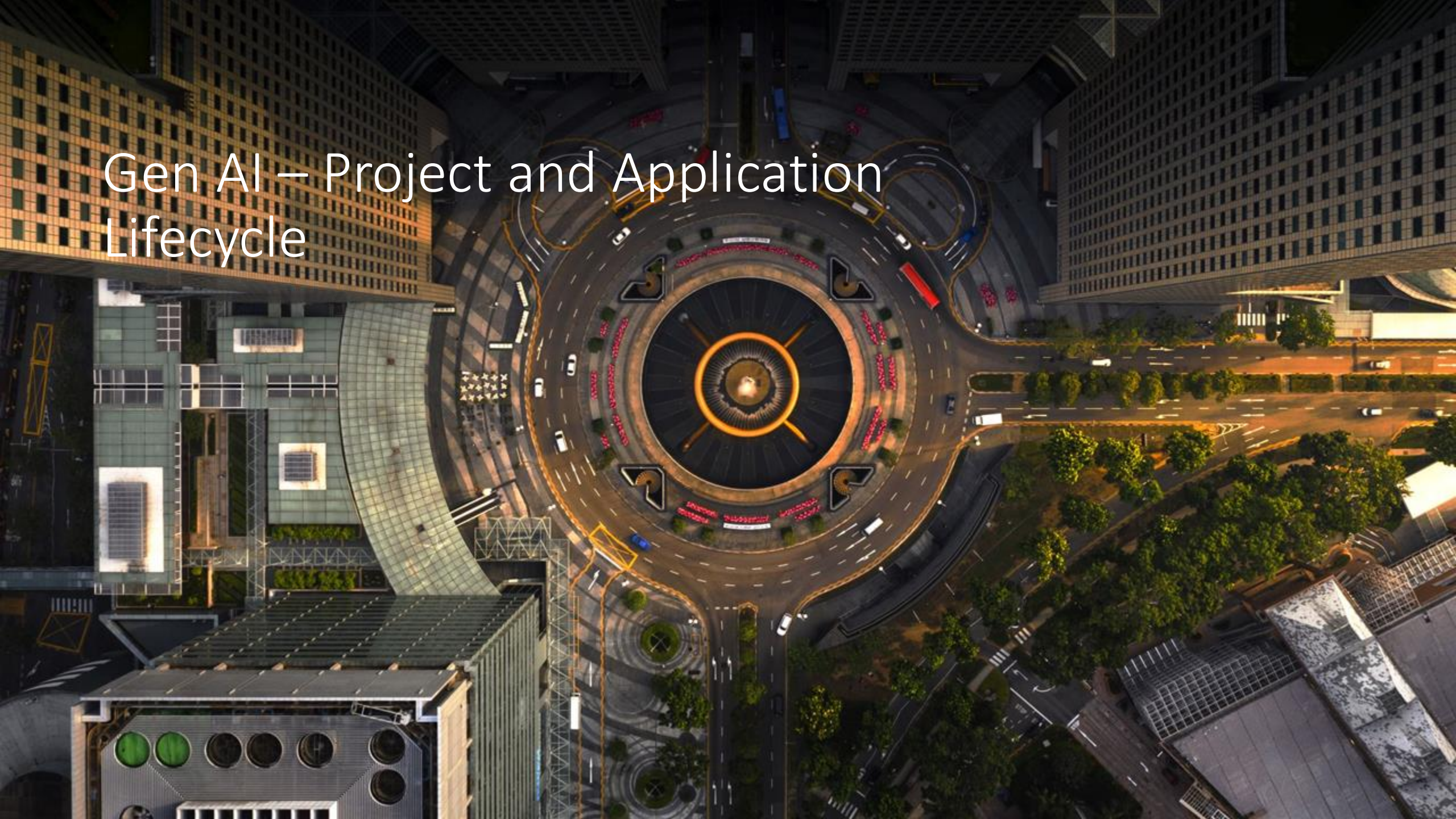


LLMs – Characteristics

- **General purpose models**
- **Stateless**
- **Stochastic**

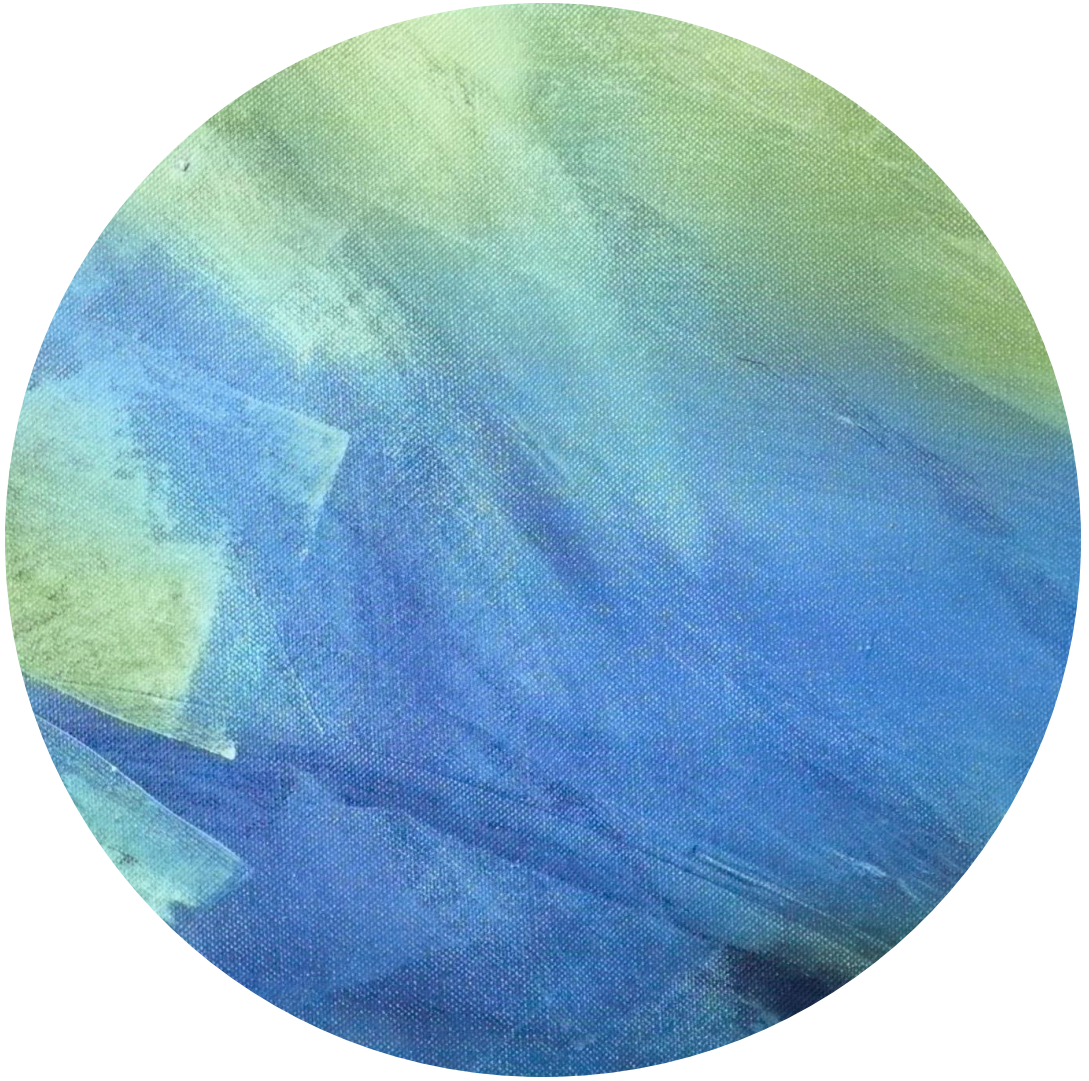


Gen AI – Project and Application Lifecycle



Gen AI Project/Product Lifecycle

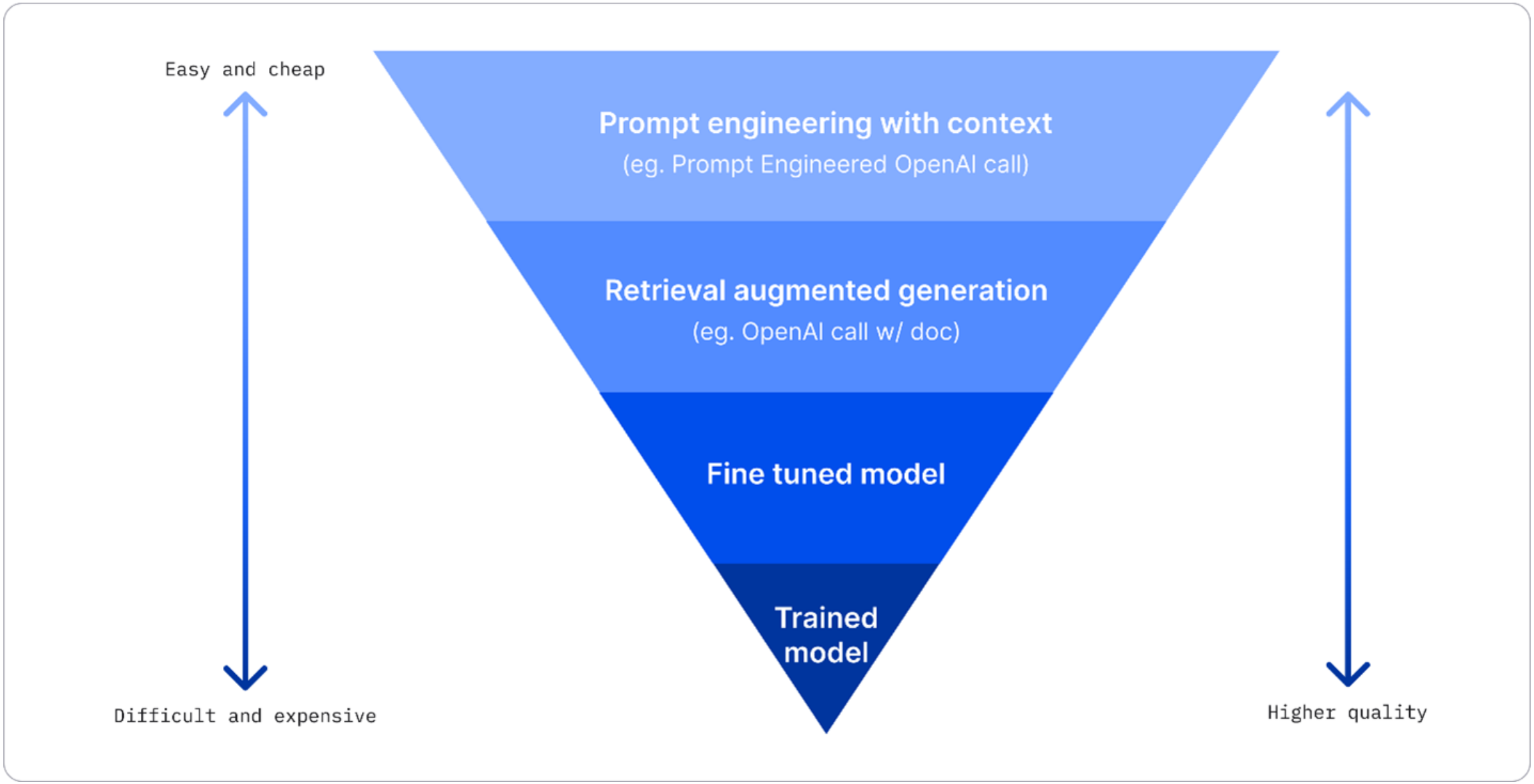
- **Identify the use case**
- **Choose an existing foundation model or use custom (transfer learning or fine-tuning) model**
- **Prompting, Tuning and Aligning**
  - **Prompt Engineering**
  - **Instruction Tuning**
  - **Fine-Tuning (Full or PEFT)**
  - **Align with human feedback (RLHF)**
- **Evaluation**
- **Deployment (Optimized model)**
- **Building LLM Powered Application**



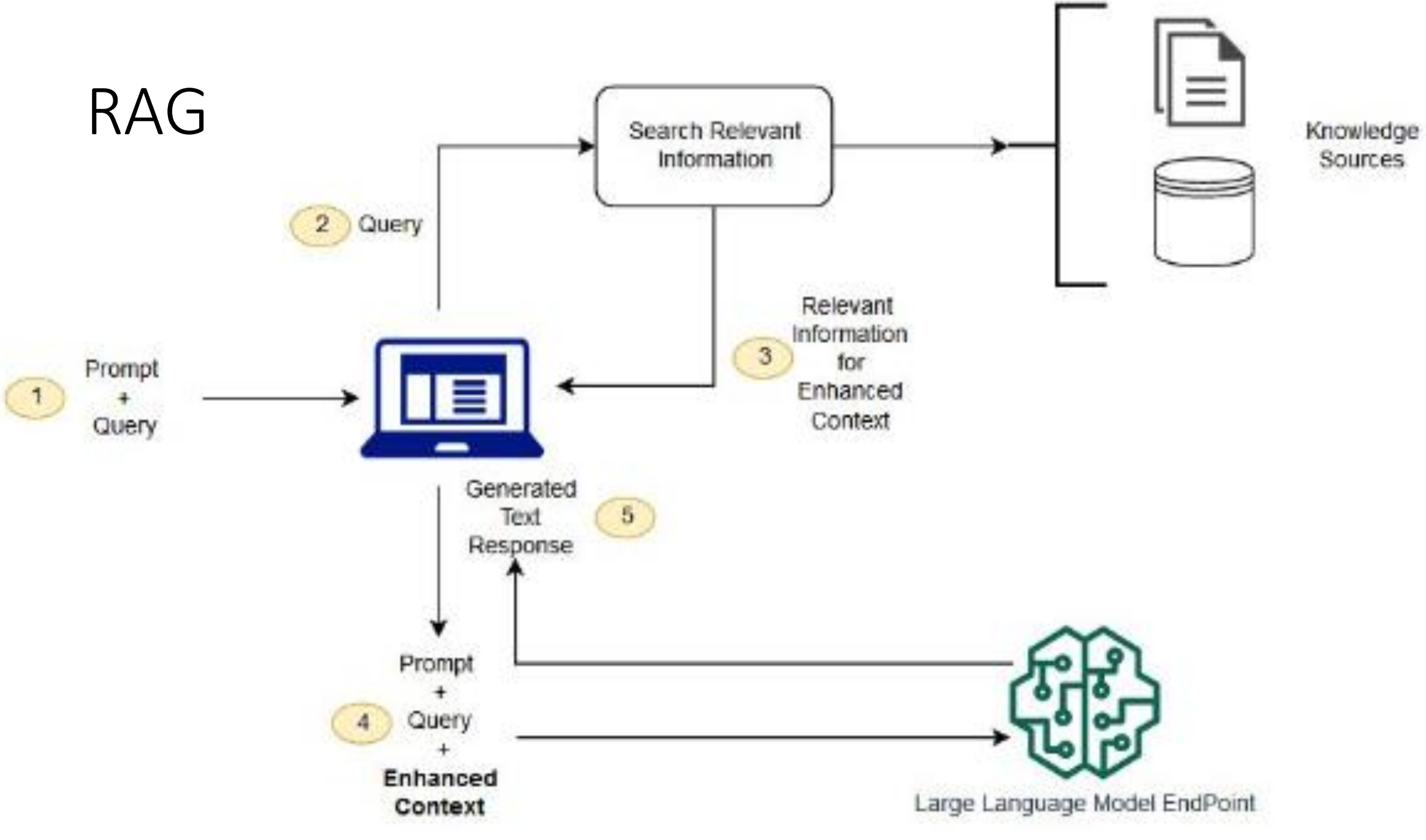
Iterative in nature

Cheat Sheet - Time and effort in the lifecycle

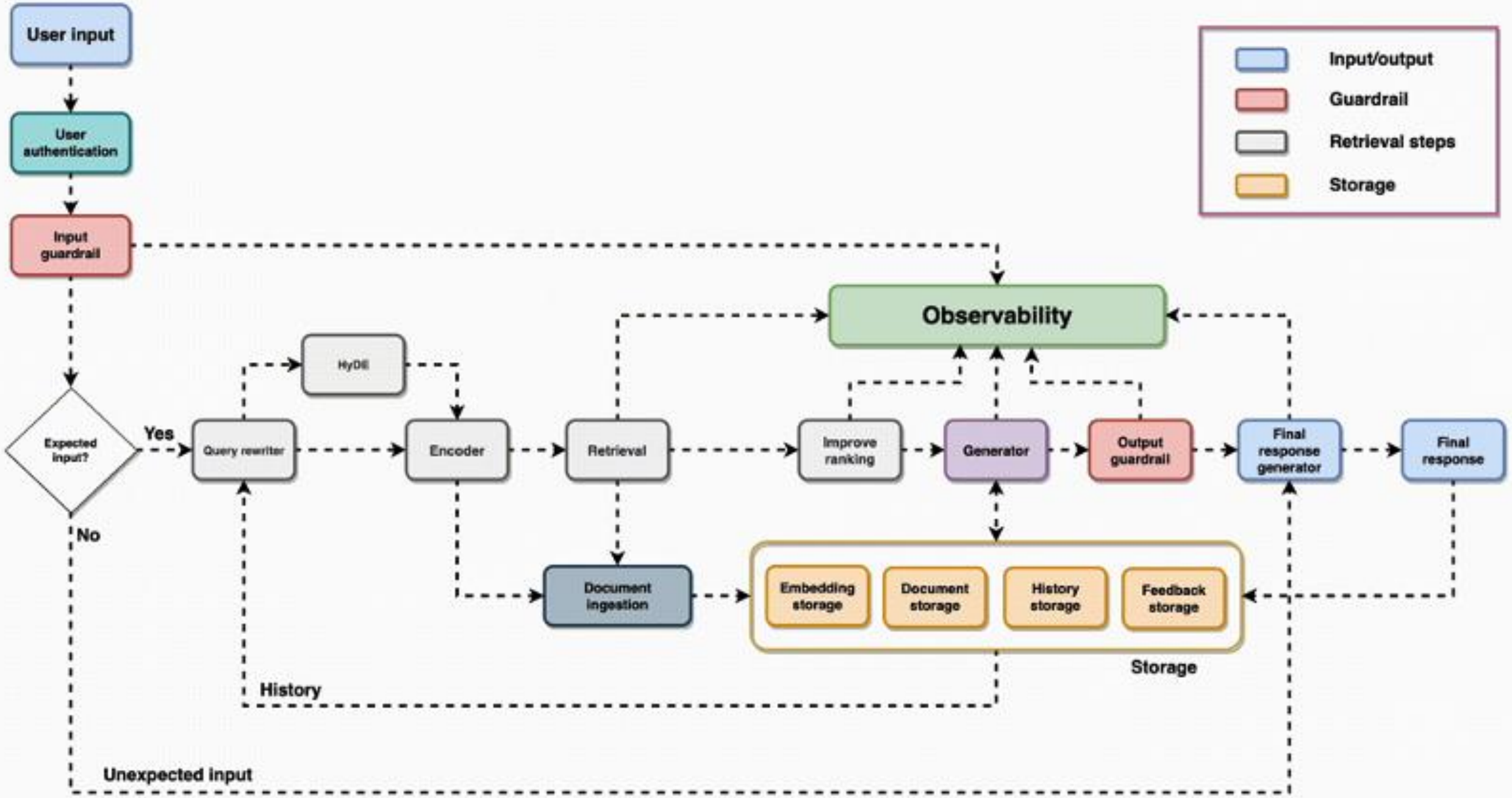
	Pre-training	Prompt engineering	Prompt tuning and fine-tuning	Reinforcement learning/human feedback	Compression/ optimization/ deployment
Training duration	Days to weeks to months	Not required	Minutes to hours	Minutes to hours similar to fine-tuning	Minutes to hours
Customization	Determine model architecture, size and tokenizer. Choose vocabulary size and # of tokens for input/context Large amount of domain training data	No model weights Only prompt customization	Tune for specific tasks Add domain-specific data Update LLM model or adapter weights	Need separate reward model to align with human goals (helpful, honest, harmless) Update LLM model or adapter weights	Reduce model size through model pruning, weight quantization, distillation Smaller size, faster inference
Objective	Next-token prediction	Increase task performance	Increase task performance	Increase alignment with human preferences	Increase inference performance
Expertise	High	Low	Medium	Medium-High	Medium


Easy and cheap
Prompt engineering with context
(eg. Prompt Engineered OpenAI call)
Retrieval augmented generation
(eg. OpenAI call w/ doc)
Fine tuned model
Trained model
Difficult and expensive
Higher quality

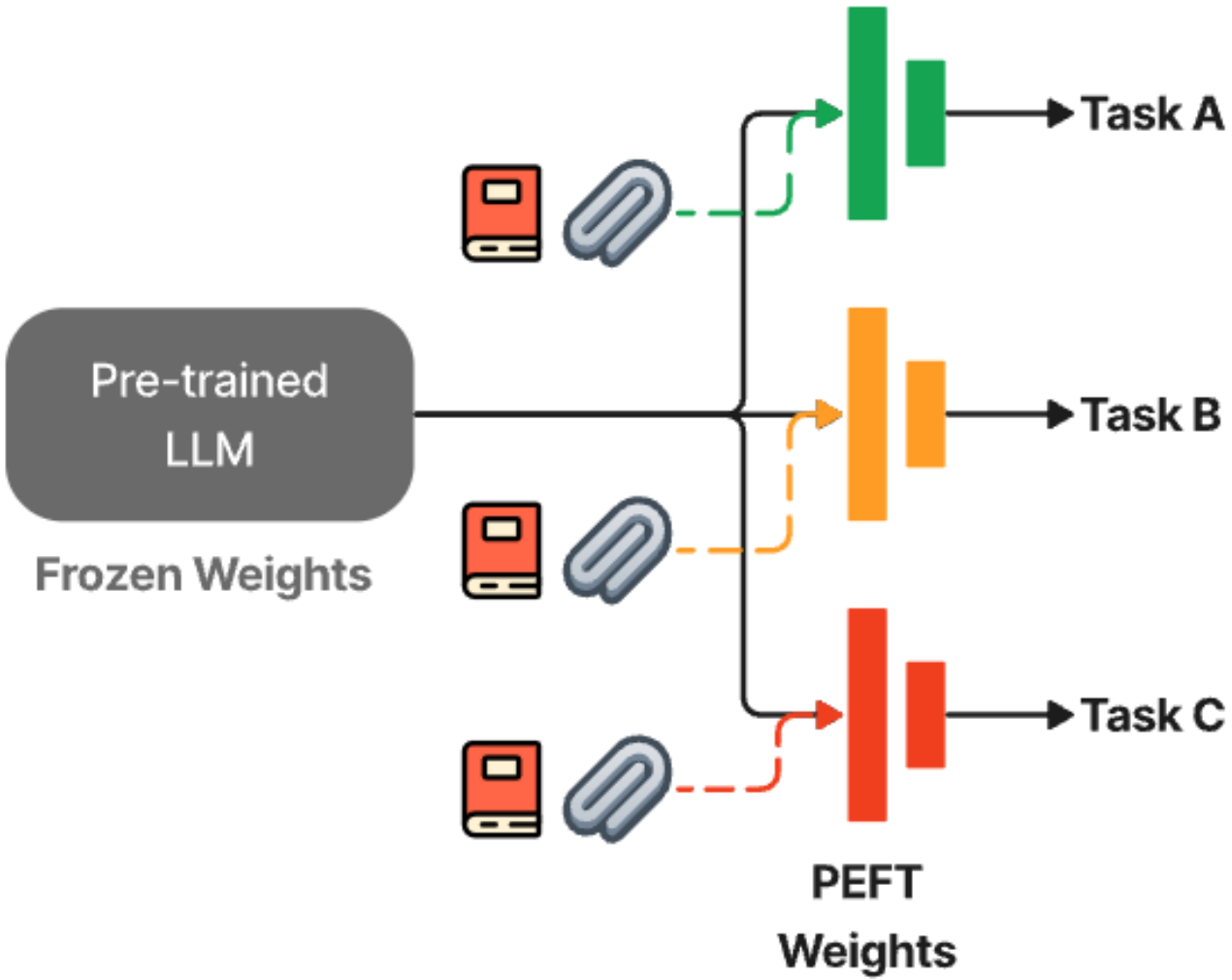
RAG


Search Relevant Information
Knowledge Sources
2 Query
1 Prompt + Query
3 Relevant Information for Enhanced Context
Generated Text Response
5
Prompt + 4 Query + Enhanced Context
Large Language Model EndPoint

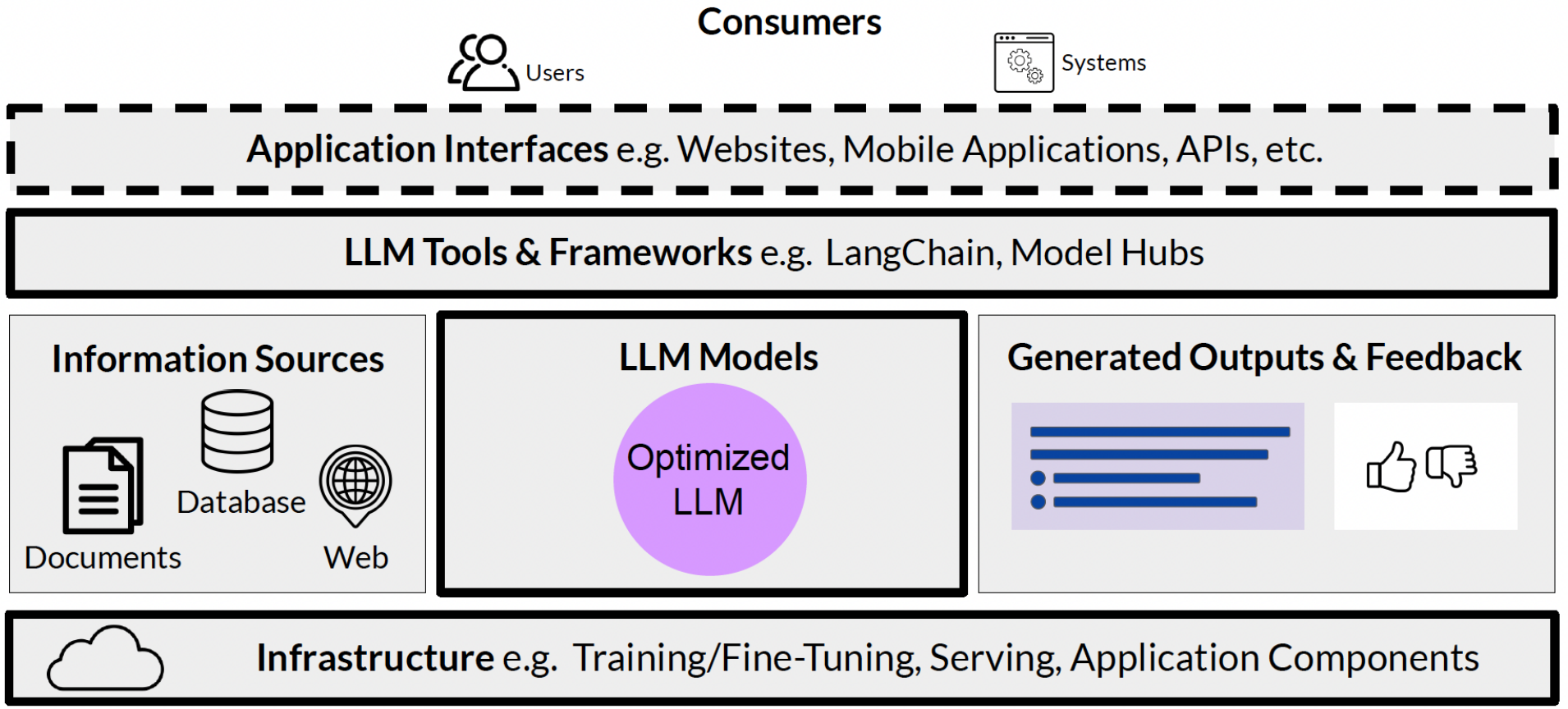
Architecture For Enterprise RAG

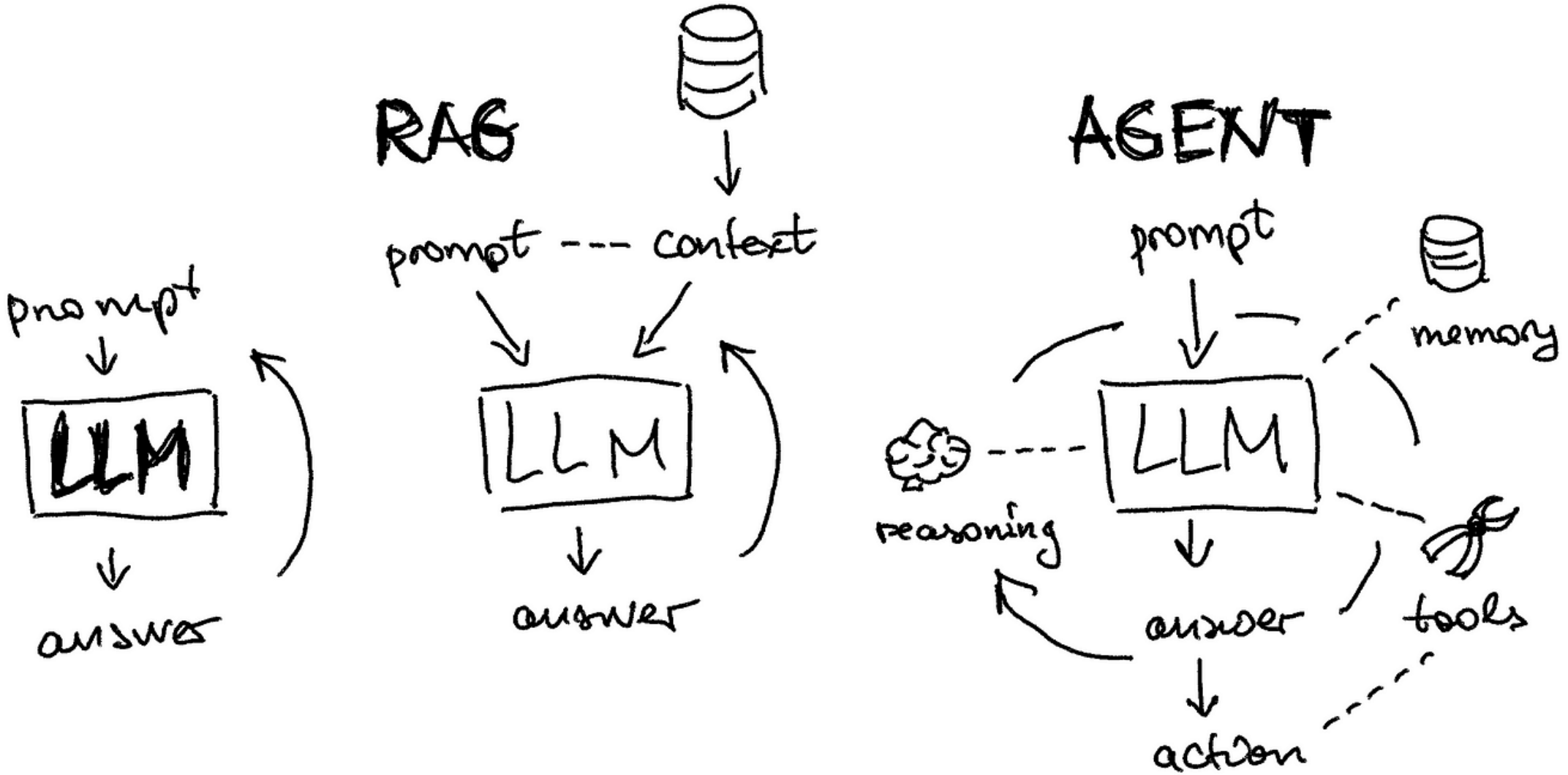


Parameter Efficient Fine-Tuning (PEFT)

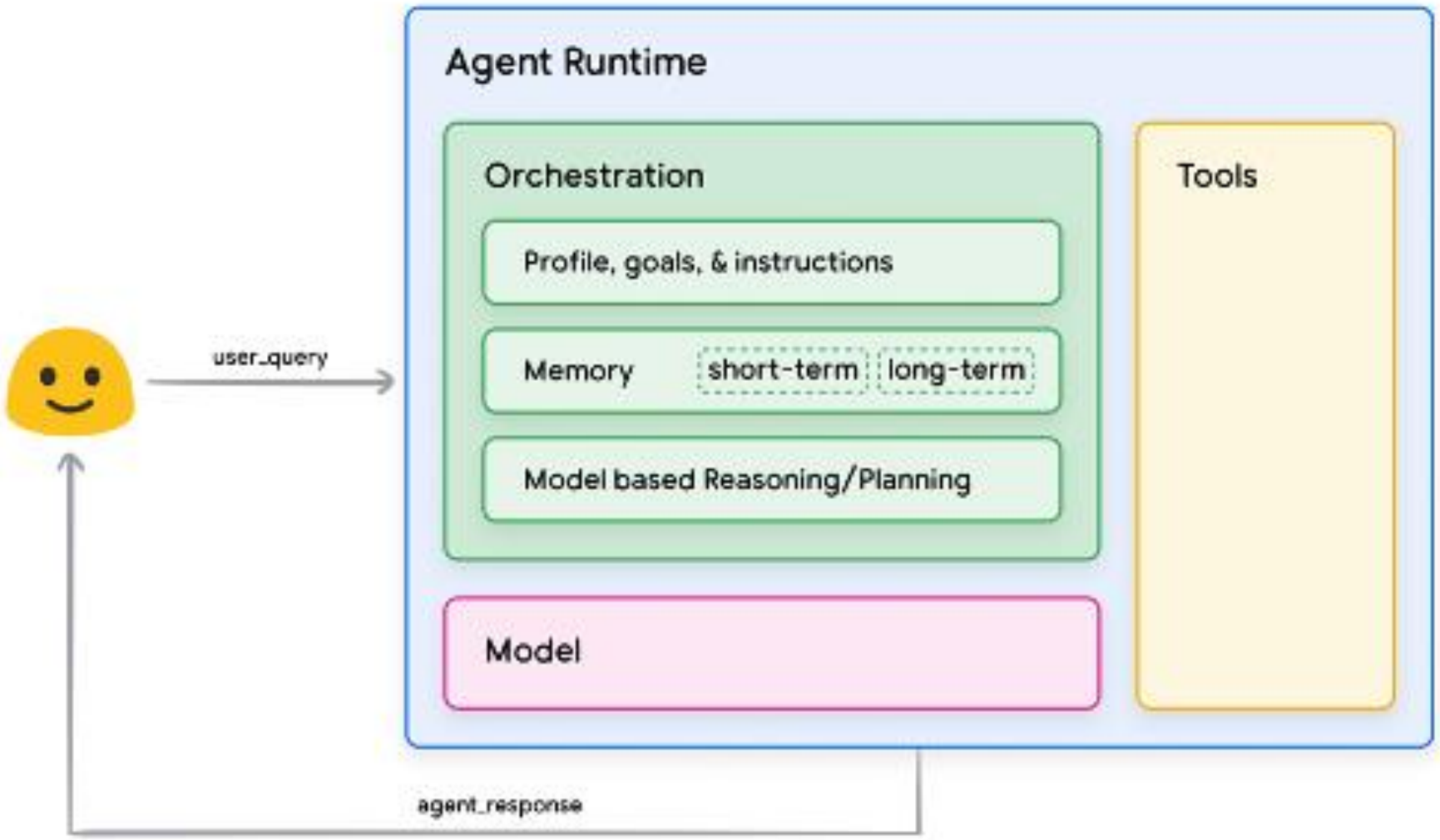


Building generative applications



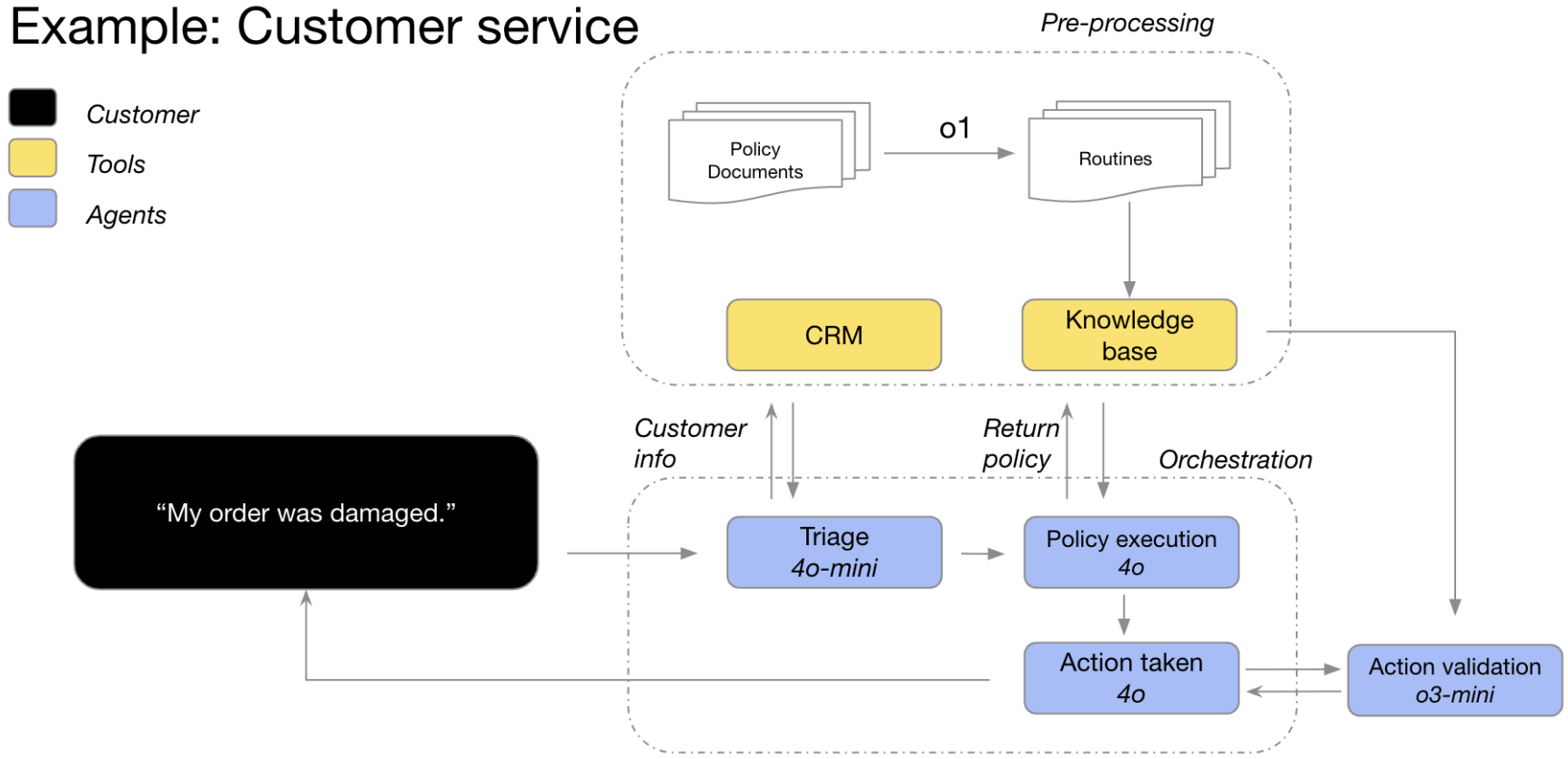

prompt
LLM
answer
RAG
prompt --- context
LLM
answer
AGENT
prompt
memory
reasoning
LLM
answer
tools
action

General agent architecture and components



How to choose?

GPT-4o and GPT-4o mini models triage order details with customer information, identify the order issues and the return policy, and then feed all of these data points into o3-mini to make the final decision about the viability of the return based on policy.



Agentic AI is picking up where GenAI and RPA left off

Low code?

RPA

"I follow instructions exactly"

RPA is the **task automation** that replaces manual effort in routine, rule-based processes.

Key Characteristics:
- Executes structured, rule-based processes
- Performs repetitive digital tasks with precision
- Operates within defined system boundaries
- Follows exact step-by-step procedures

GenAI

"I can create based on prompts"

GenAI is a **productivity amplifier** that supports and enhances human work, transforming workflows without fully replacing human decision-making.

Key Characteristics:
- Assists with specific tasks (writing, analysis, coding)
- Requires human direction and oversight
- Improves individual productivity
- Works within existing job roles

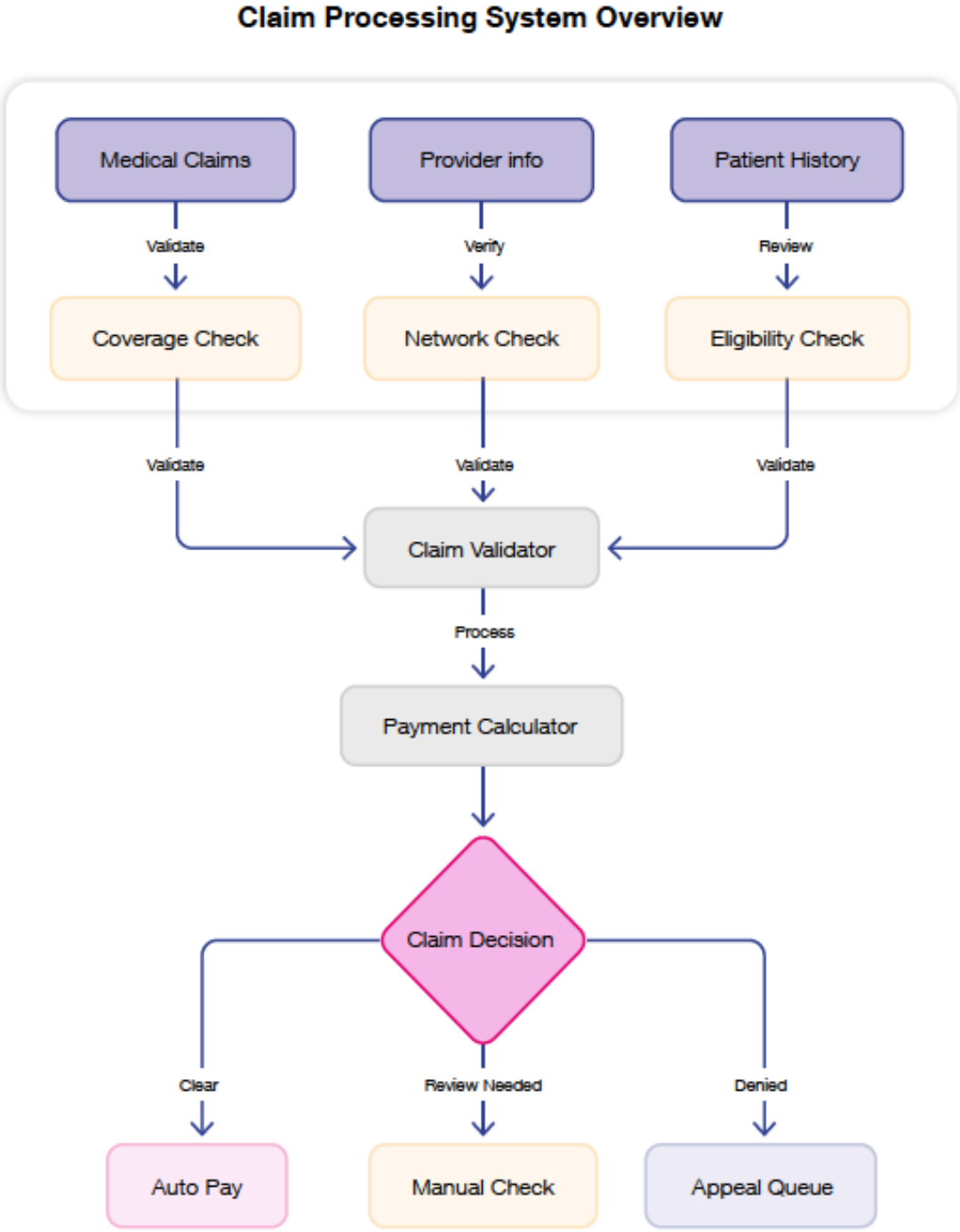
Agentic AI

"I can understand goals and figure out how to achieve them"

Agentic AI is a **collaborative actor** that autonomously executes and coordinates complex tasks.

Key Characteristics:
- Acts as virtual coworker completing end-to-end processes
- Self-directs and coordinates multiple tasks
- Transforms entire workflows
- Creates new organizational paradigms

Advancing the Claims Processing Agent



AI Agent Challenges and Solutions

DEVELOPMENT ISSUES

Poorly Defined Prompts
- Define Clear Objectives
- Craft Detailed Personas
- Use Effective Prompting

Evaluation Challenges
- Continuous Evaluation
- Use Real-World Scenarios
- Incorporate Feedback Loops

LLM ISSUES

Difficult to Steer
- Specialized Prompts
- Hierarchical Design
- Fine-Tuning Models

High Cost of Running
- Reduce Context Size
- Use Smaller Models
- Cloud-Based Solutions

Planning Failures
- Task Decomposition
- Multi-Plan Selection
- Reflection and Refinement

Reasoning Failures
- Enhance Reasoning Capabilities
- Fine-Tune LLMs with Feedback
- Use Specialized Agents

Tool Calling Failures
- Define Clear Parameters
- Validate Tool Outputs
- Tool Selectio VerificationLoops

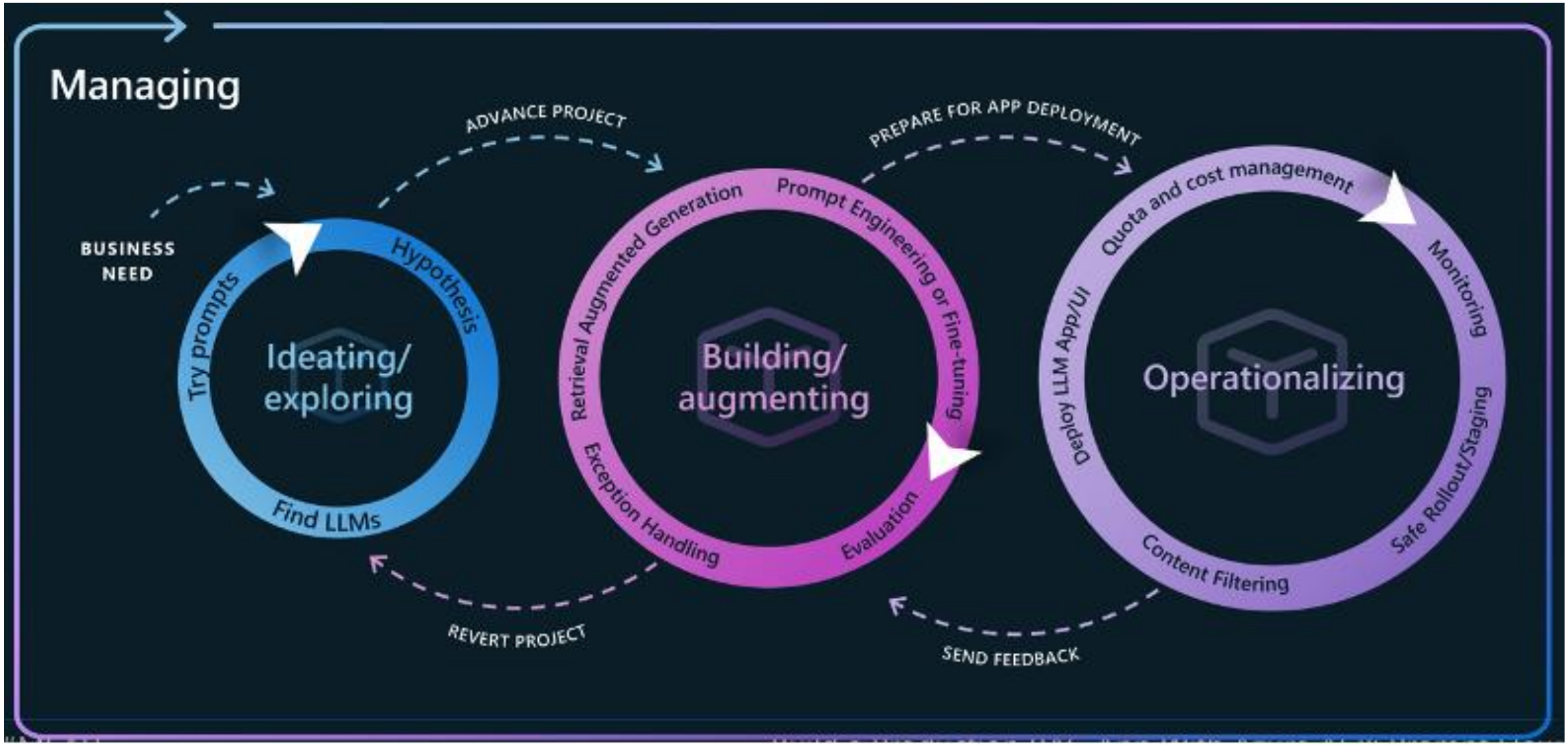
PRODUCTION ISSUES

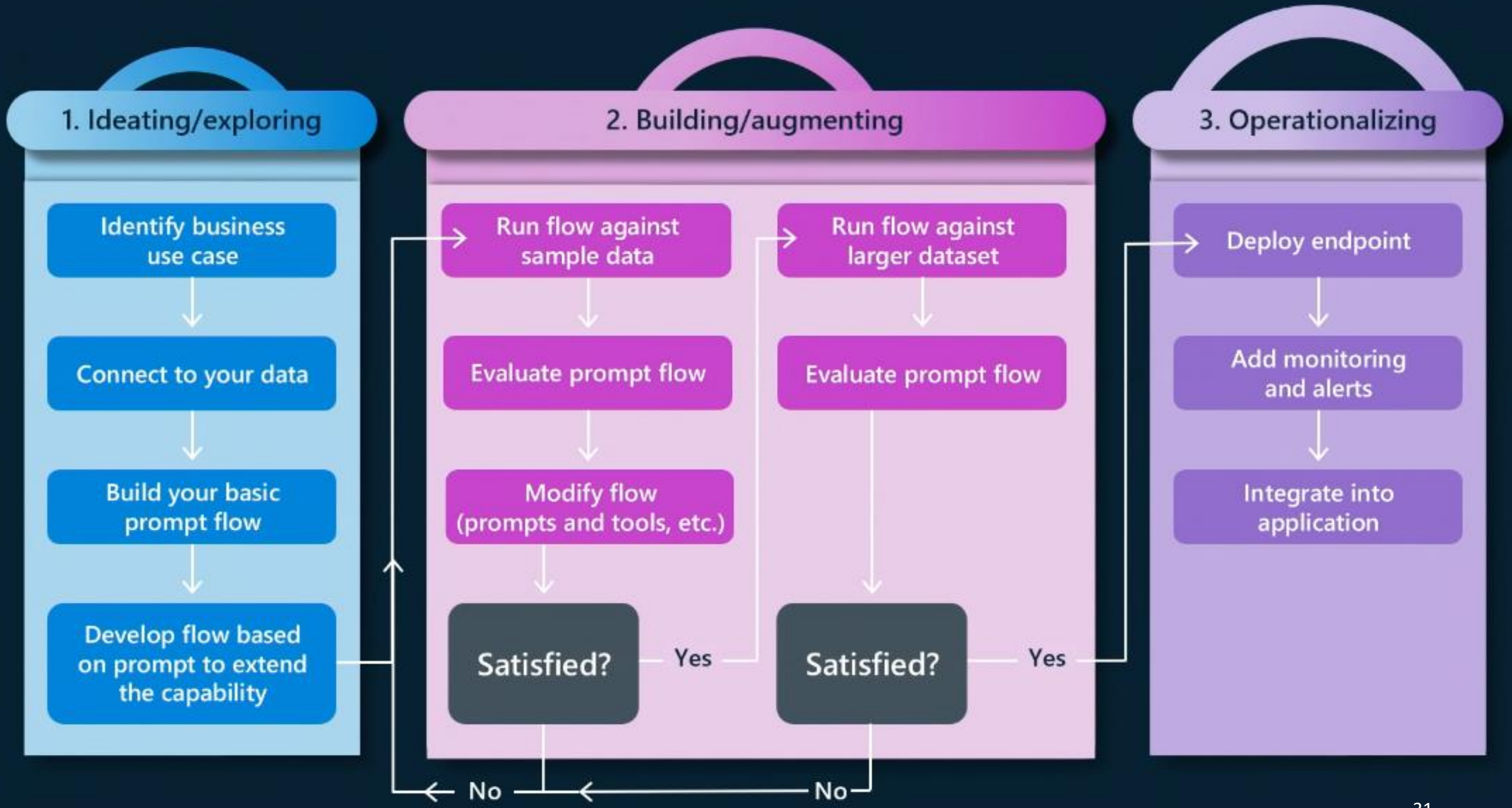
Guardrails
- Rule-Based Filters & Validation
- Human-in-the-Loop Oversight
- Ethical & Compliance Frameworks

Agent Scaling
- Scalable Architectures
- Resource Management
- Monitor Performance

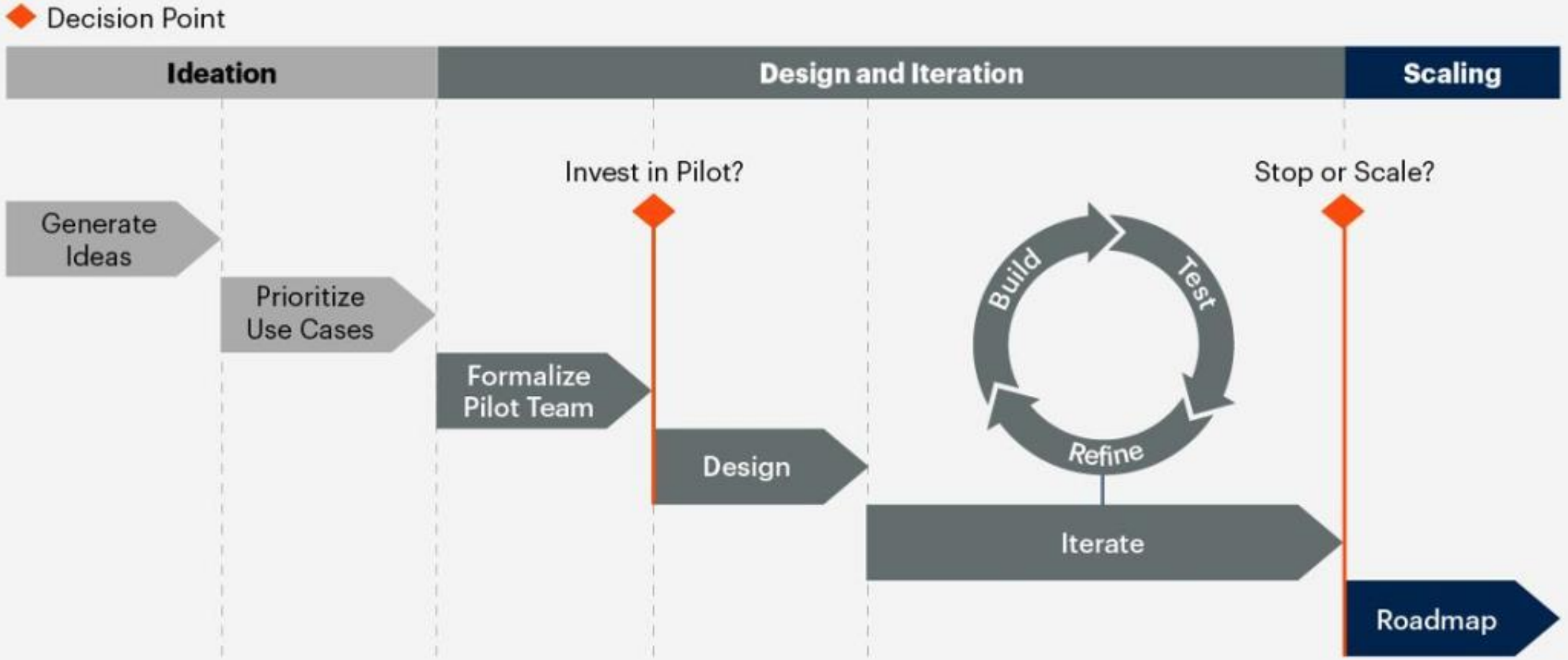
Fault Tolerance
- Redundancy
- Automated Recovery
- Stateful Recovery

Infinite Looping
- Clear Termination Conditions
- Enhance Reasoning & Planning
- Monitor Agent Behavior


Managing
ADVANCE PROJECT
PREPARE FOR APP DEPLOYMENT
BUSINESS NEED
Ideating/ exploring
Try prompts
Hypothesis
Find LLMs
Building/ augmenting
Retrieval Augmented Generation
Prompt Engineering or Fine-tuning
Exception Handling
Evaluation
Operationalizing
Quota and cost management
Monitoring
Safe Rollout/Staging
Content Filtering
Deploy LLM App/UI
REVERT PROJECT
SEND FEEDBACK


1. Ideating/exploring
Identify business use case
Connect to your data
Build your basic prompt flow
Develop flow based on prompt to extend the capability
2. Building/augmenting
Run flow against sample data
Evaluate prompt flow
Modify flow (prompts and tools, etc.)
Satisfied?
Yes
No
Run flow against larger dataset
Evaluate prompt flow
Satisfied?
Yes
No
3. Operationalizing
Deploy endpoint
Add monitoring and alerts
Integrate into application

Generative AI Pilot Phases and Decision Points

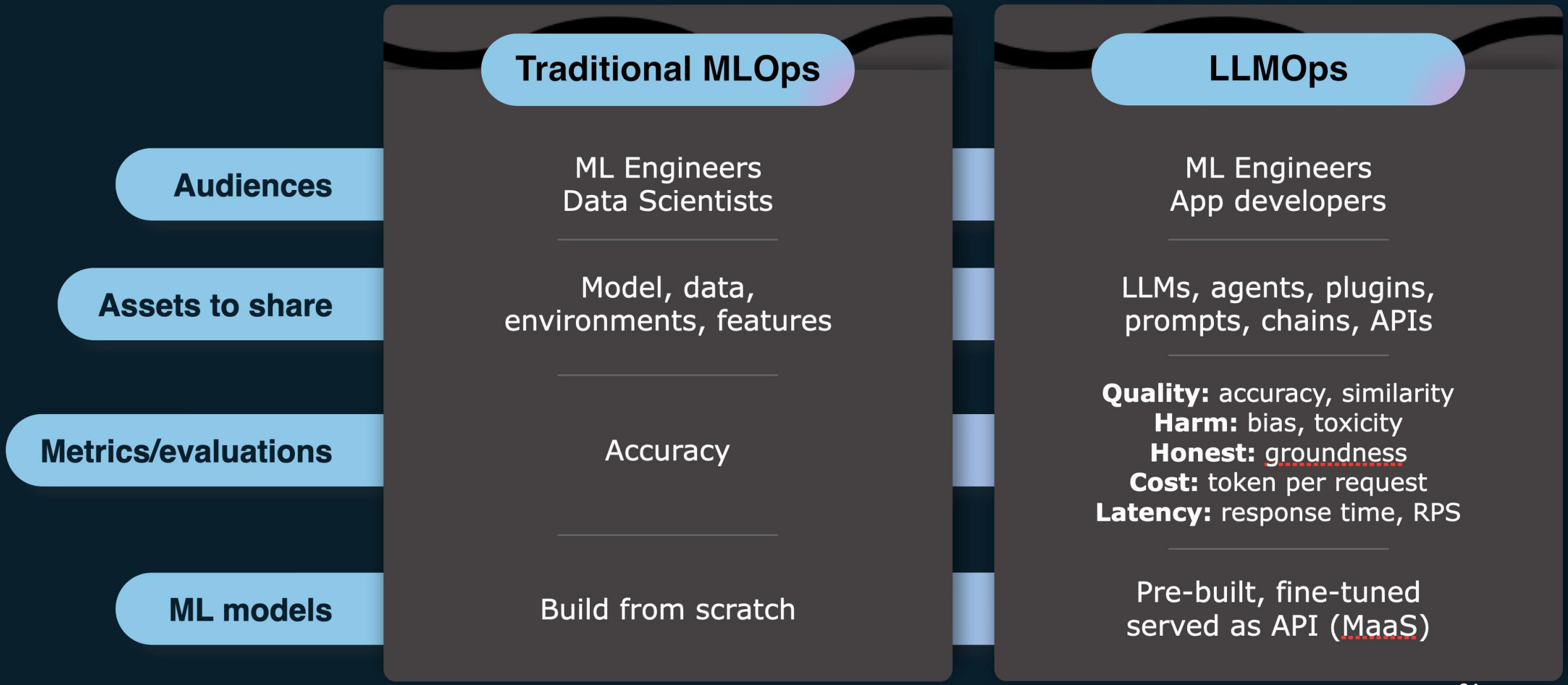


GenAI - Paradigm Shift

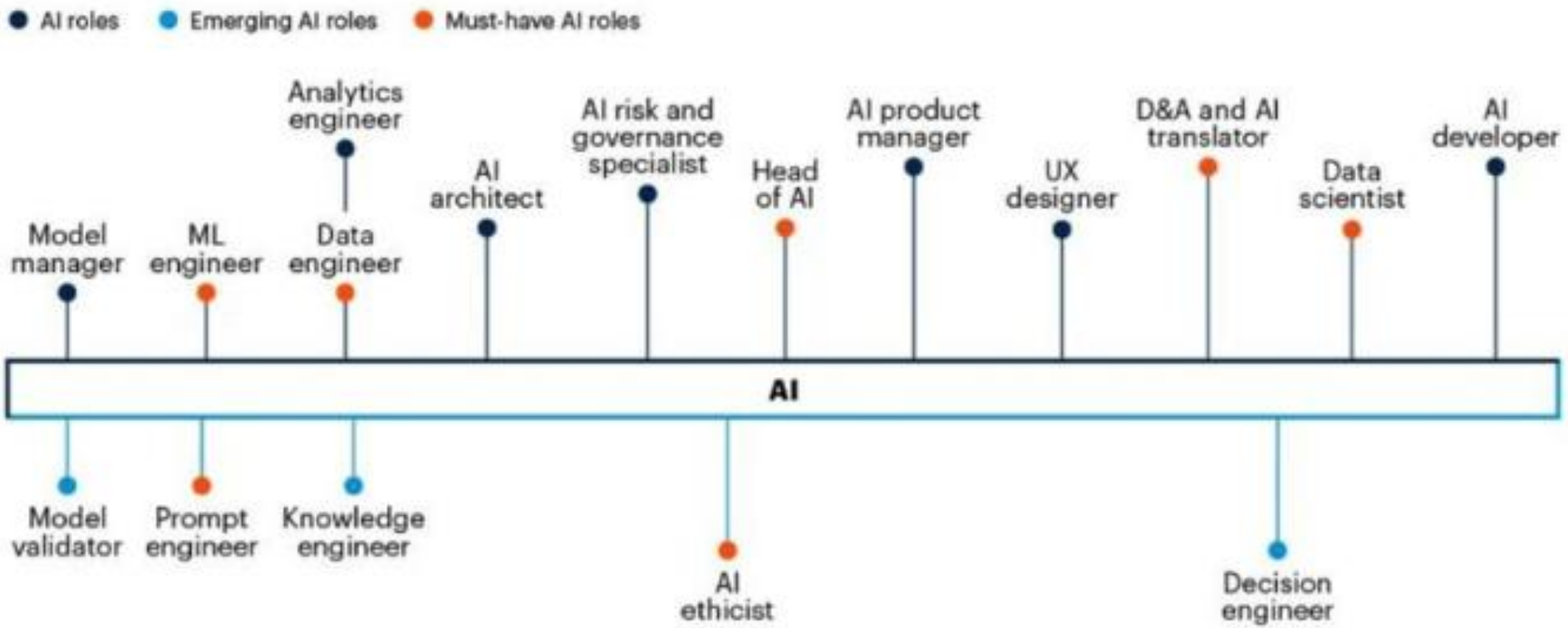
Traditional software vs LLMs

	Traditional Software	LLM-based Applications
Behavior	Predefined rules	Probability + prediction
Output	Deterministic same input -> same output	Non-deterministic Same input -> many possible outputs
Testing	testing: 1 input, 1 correct output	1 input, many correct (and incorrect) outputs
Criteria	Evaluate as "right" or "wrong"	Evaluate on: • Accuracy • Quality • Consistency • Bias • Toxicity • and more...

The paradigm shift—from MLOps to LLMOps

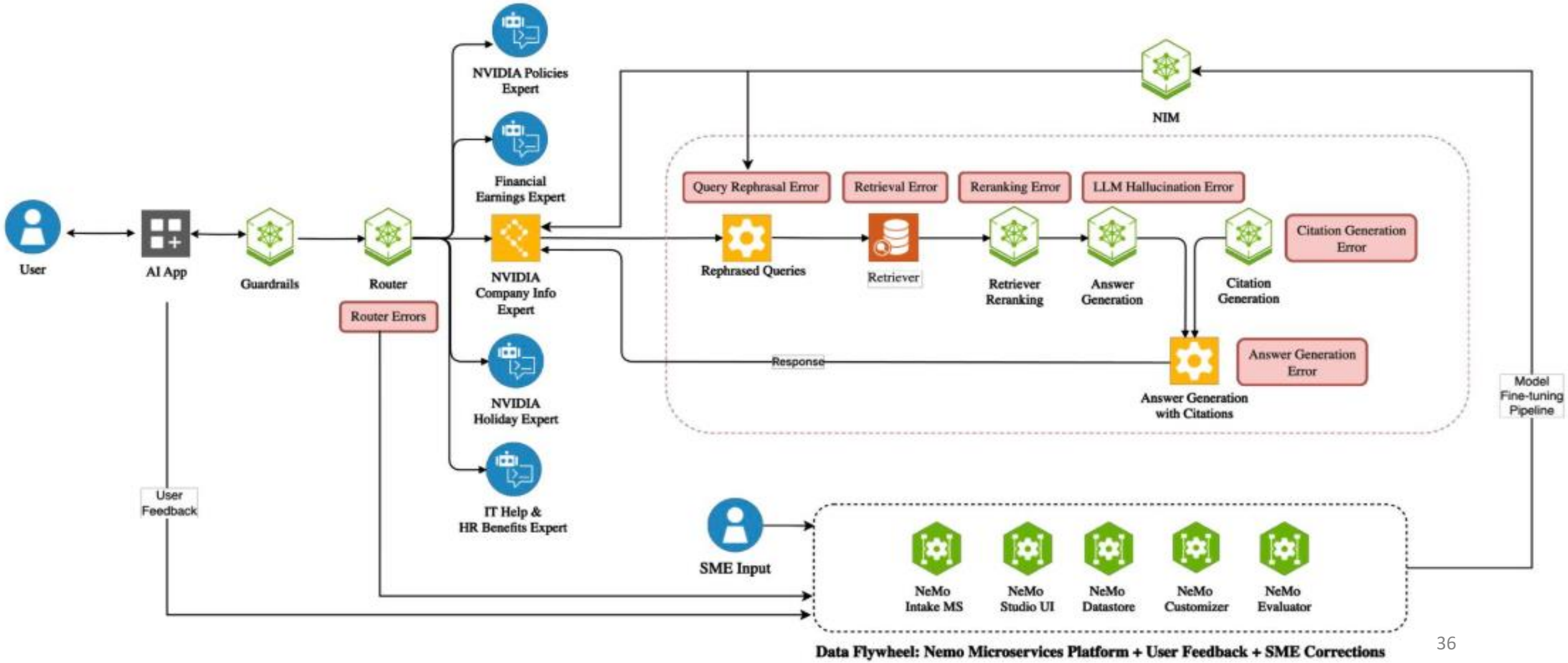


NEW ROLES FOR AI



Rise of Problem Solvers & Business _____

Architecture



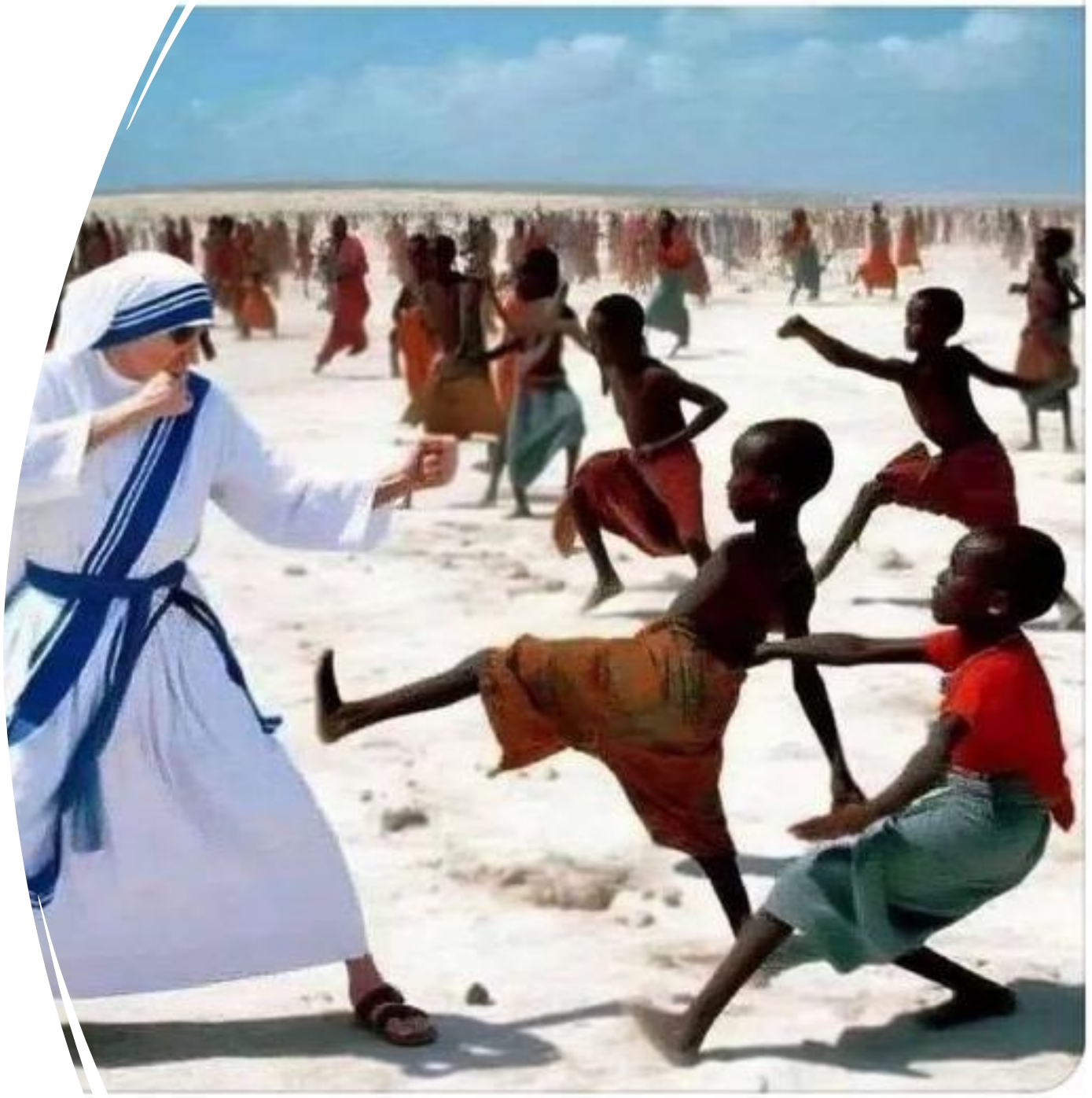
RAG going wrong

- **Air Canada Loses Court Case After Its Chatbot Hallucinated Fake Policies To a Customer**
- The airline argued that the chatbot itself was liable. The court disagreed.

AIR CANAD

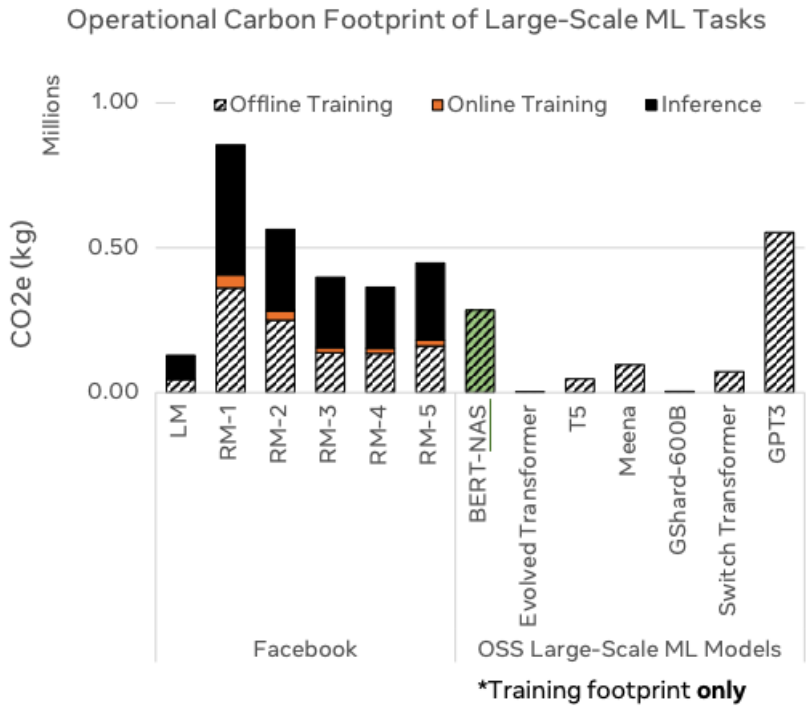
Bias

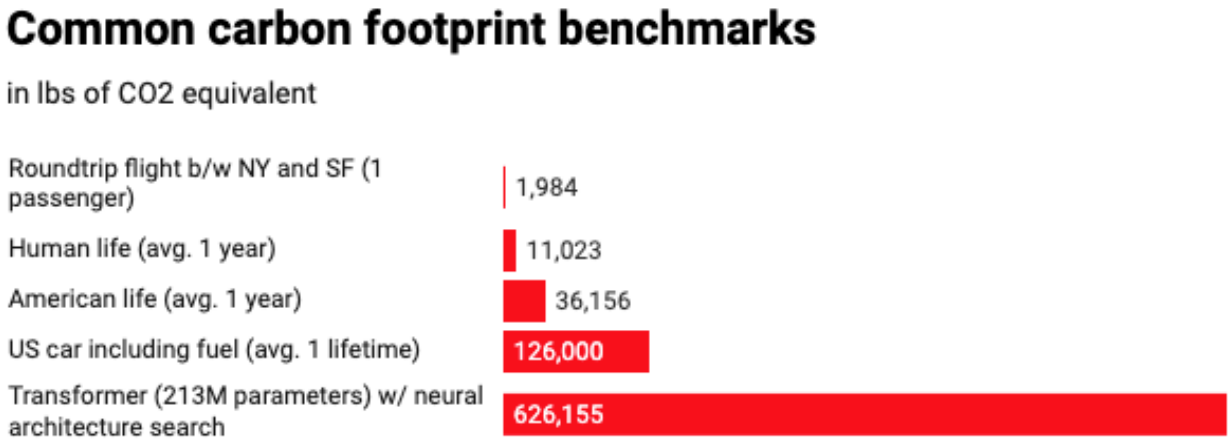
AI was asked to create a picture of Mother Theresa fighting against poverty



https://www.facebook.com/story.php/?story_fbid=812591374246846&id=100064879710238&paipv=0&eav=AfYEVE_1Oc3AFs8_ADyhlSrdd0dxR-oMQZx3d-p9I1ODQqZr67-5JtOMtiFplDaCdHQ&_rdr

Sustainable AI: Carbon Footprint



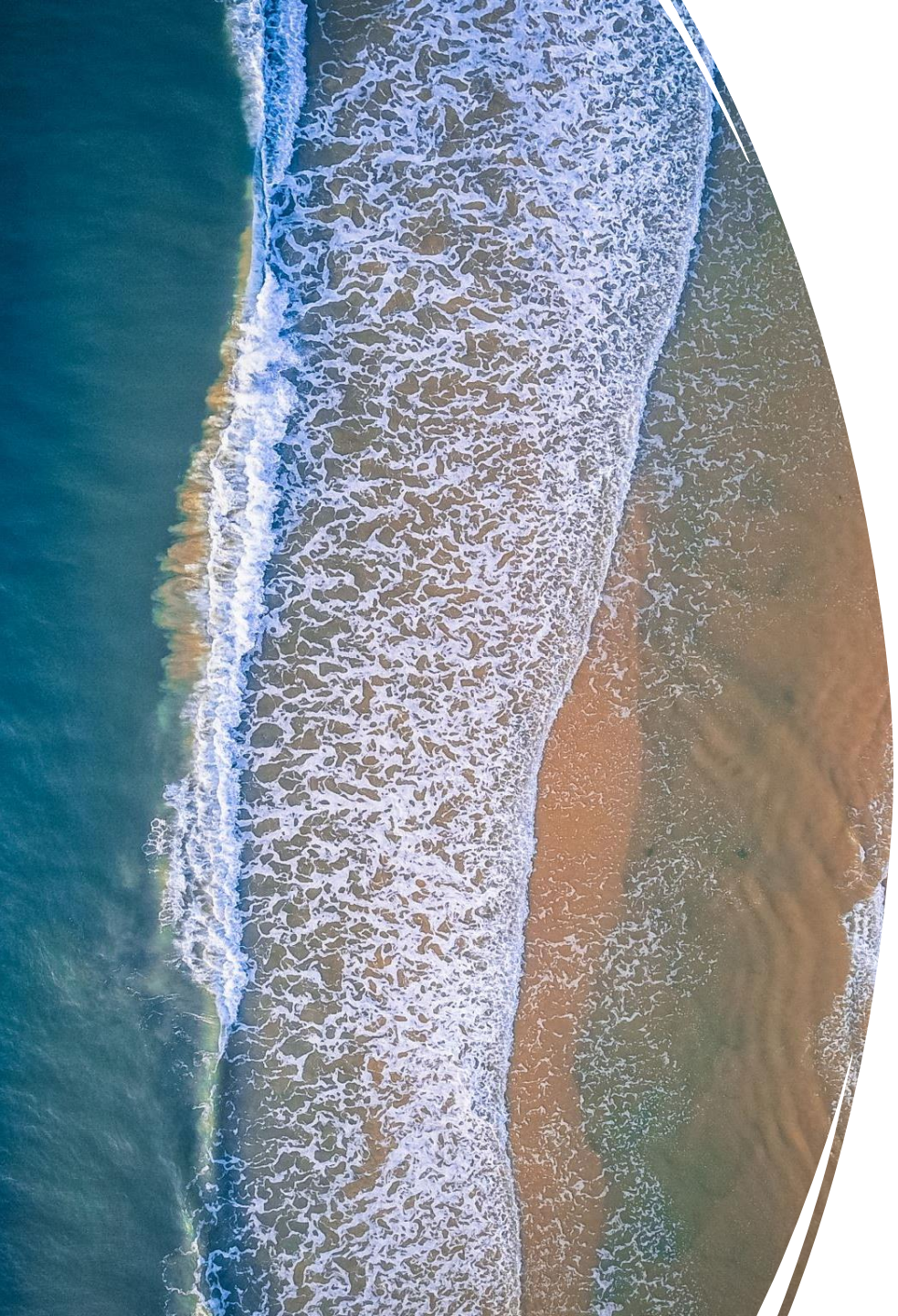


MICROSOFT IS DRAINING AN ARIZONA TOWN'S WATER SUPPLY FOR ITS AI

COULD WE JUST... NOT DO THAT?

Estimates commissioned by Microsoft itself, the 279-acre campus would approx. consume an annual **56 million gallons** of drinking water (approximately the amount that a total of **670 Goodyear (Arizona) families** would consume in a year combined)

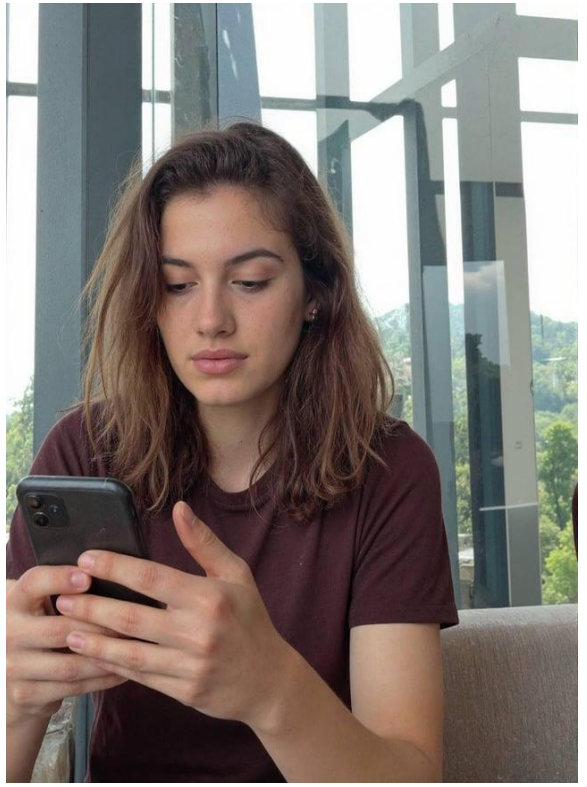
https://futurism.com/the-byte/microsoft-arizona-water-ai



Sustainable AI: Beyond Carbon Footprint

- Building and operating a leading-edge semiconductor fabrication plant, or “fab”, to produce chips like in 5nm can require up to **four million gallons of pure water each day**.
- This water usage approaches what a city of **half a million people** would require for all needs.
- Sourcing this consistently places immense strain on local water tables and reservoirs, especially in already water-stressed regions which host many high-tech manufacturing hubs.






Who is she?

https://www.reddit.com/r/ChatGPT/comments/1hu7i57/we_are_doomed/?rdt=40200