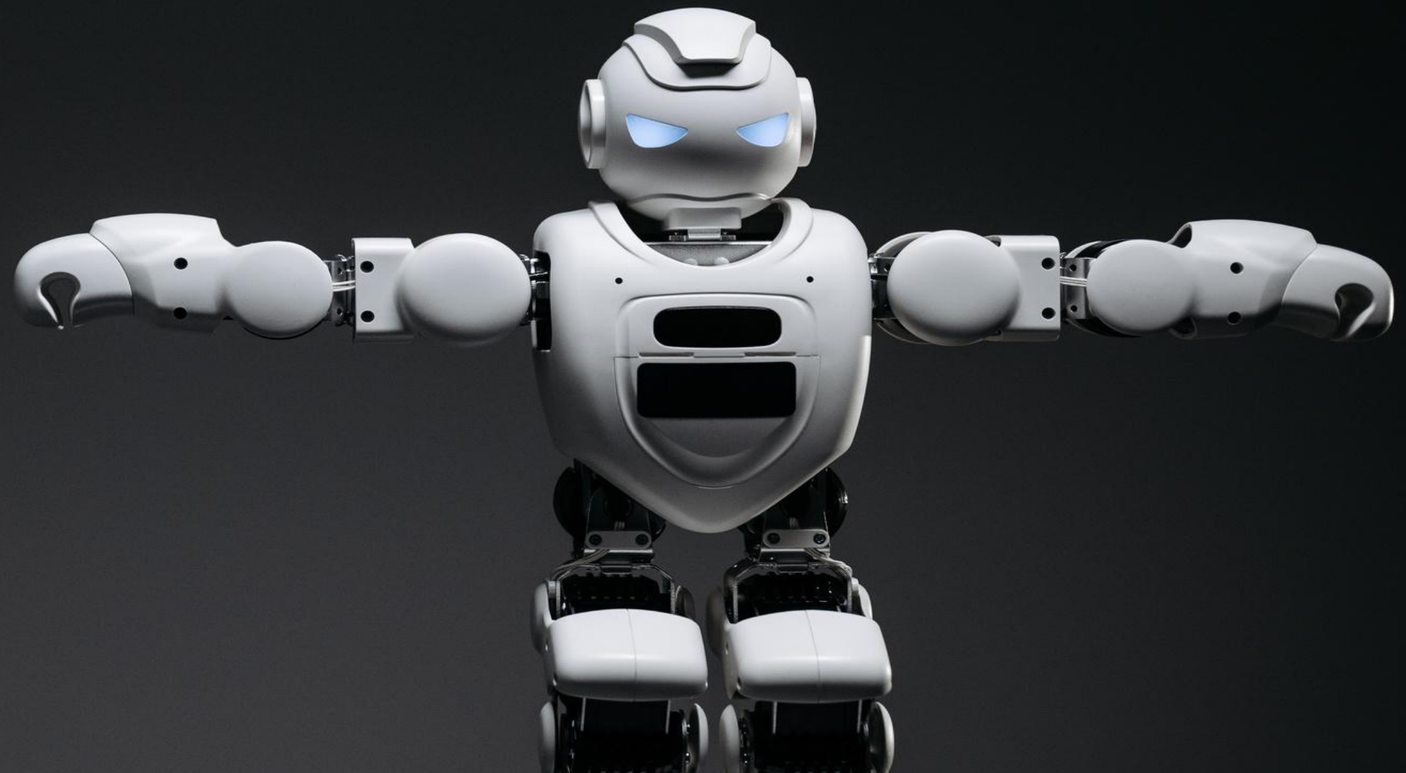
Transforming Enterprise AI with Scalable LLM Deployments

Mahmoud Fahmy
Lead AI Engineer, Mastercard

In our daily use of GPTs, are we talking to single LLM, or agent?



What we are going to cover today

1. From Excellence in Structured Data to LLMs
2. Enterprise AI transformation
3. Scaling LLM in deployment
4. The Rise of AI Agents

Artificial Intelligence

The unstructured data problem

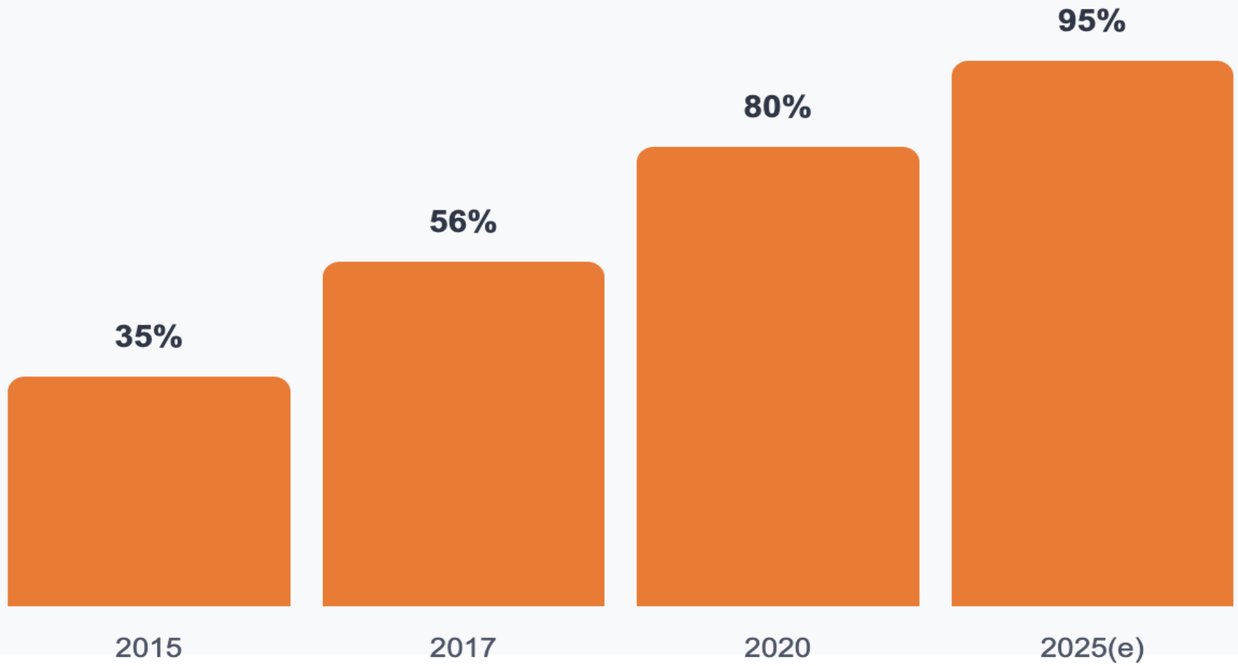
The Data Landscape

80-90%

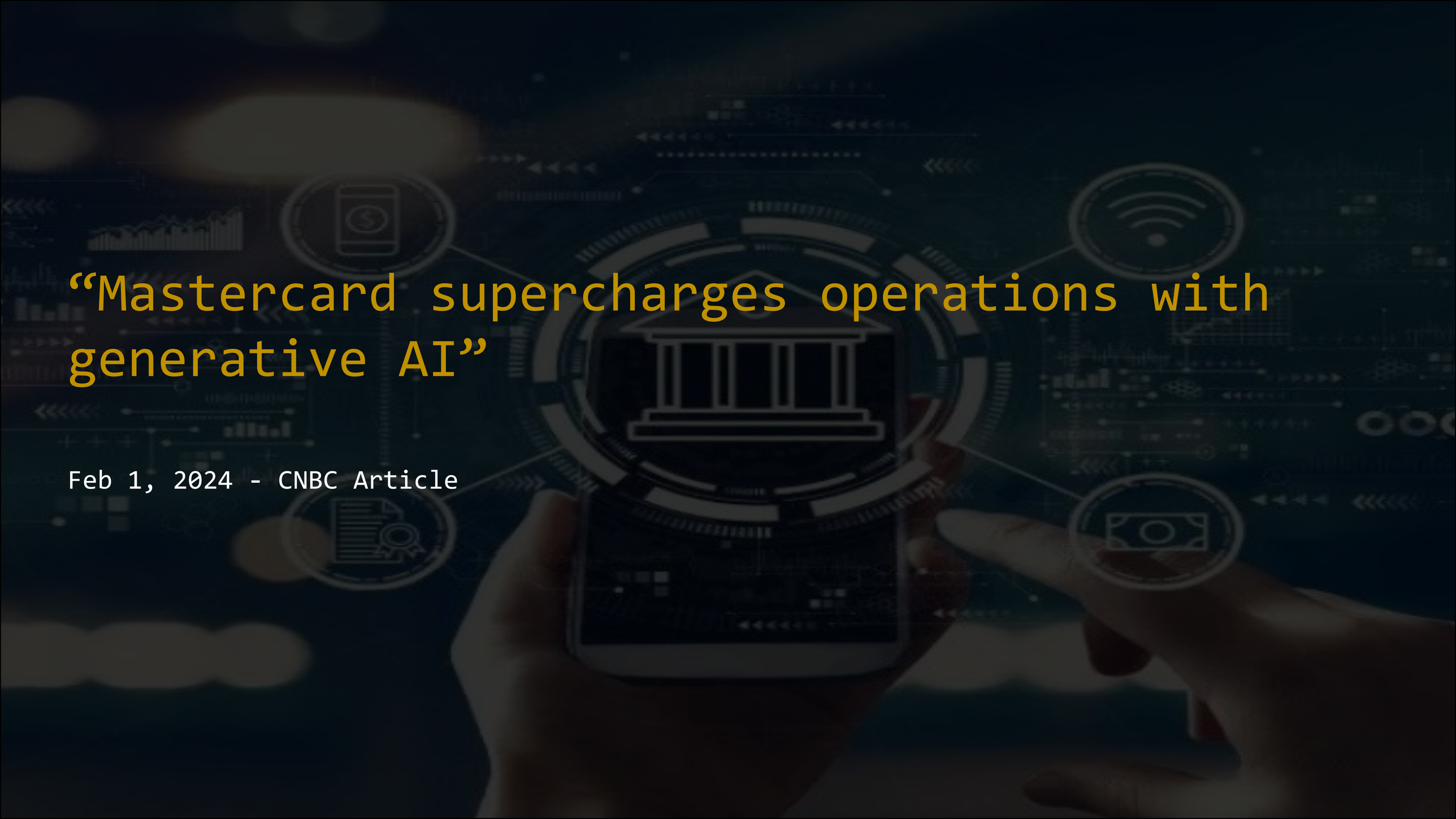
Of enterprise data is unstructured

- Documents
- Images
- Video
- Email
- Social
- Audio

Growth of Unstructured Data



- SailPoint. (2017). Market Pulse Survey: Unstructured Data Security Challenges. SailPoint Technologies.
- Gartner. (2017). Managing the Data Deluge: Best Practices for Unstructured Data Management. Gartner Research.
- Additional statistics derived from industry reports and market research (2015-2025)

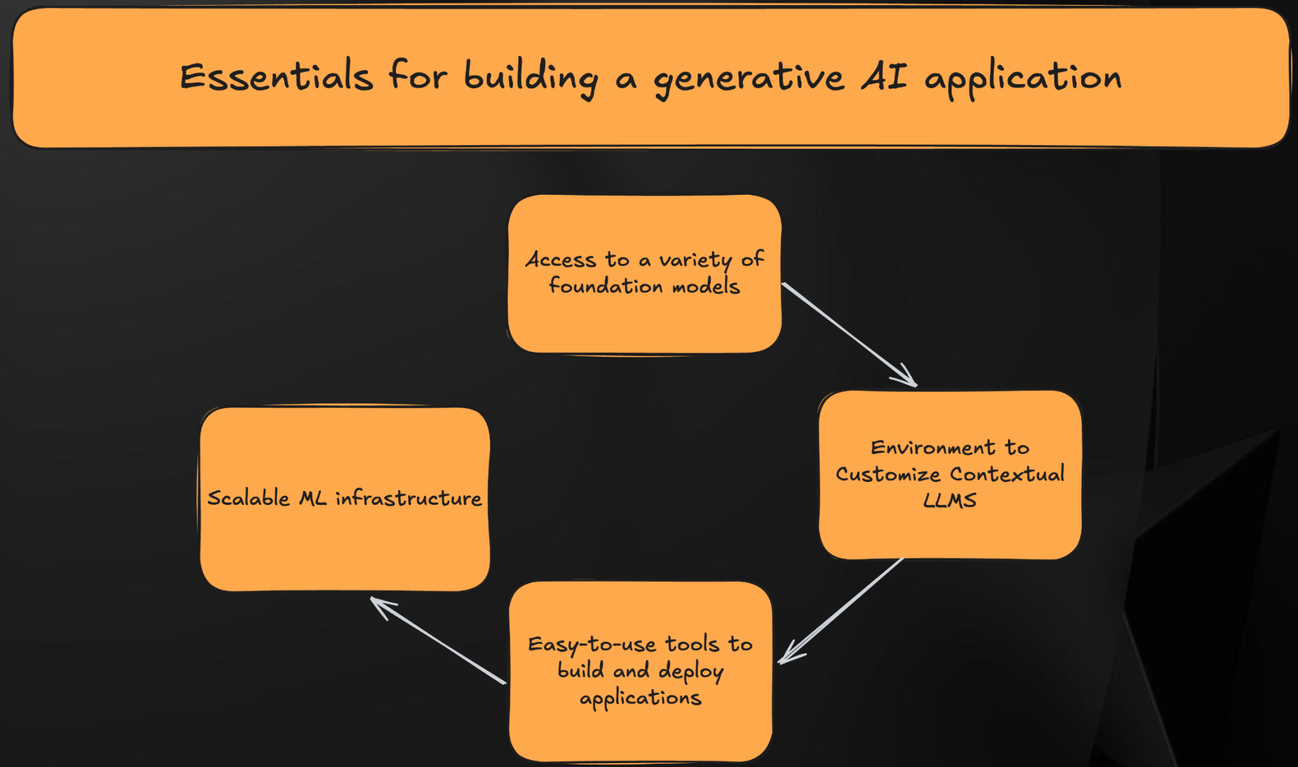


“Mastercard supercharges operations with generative AI”

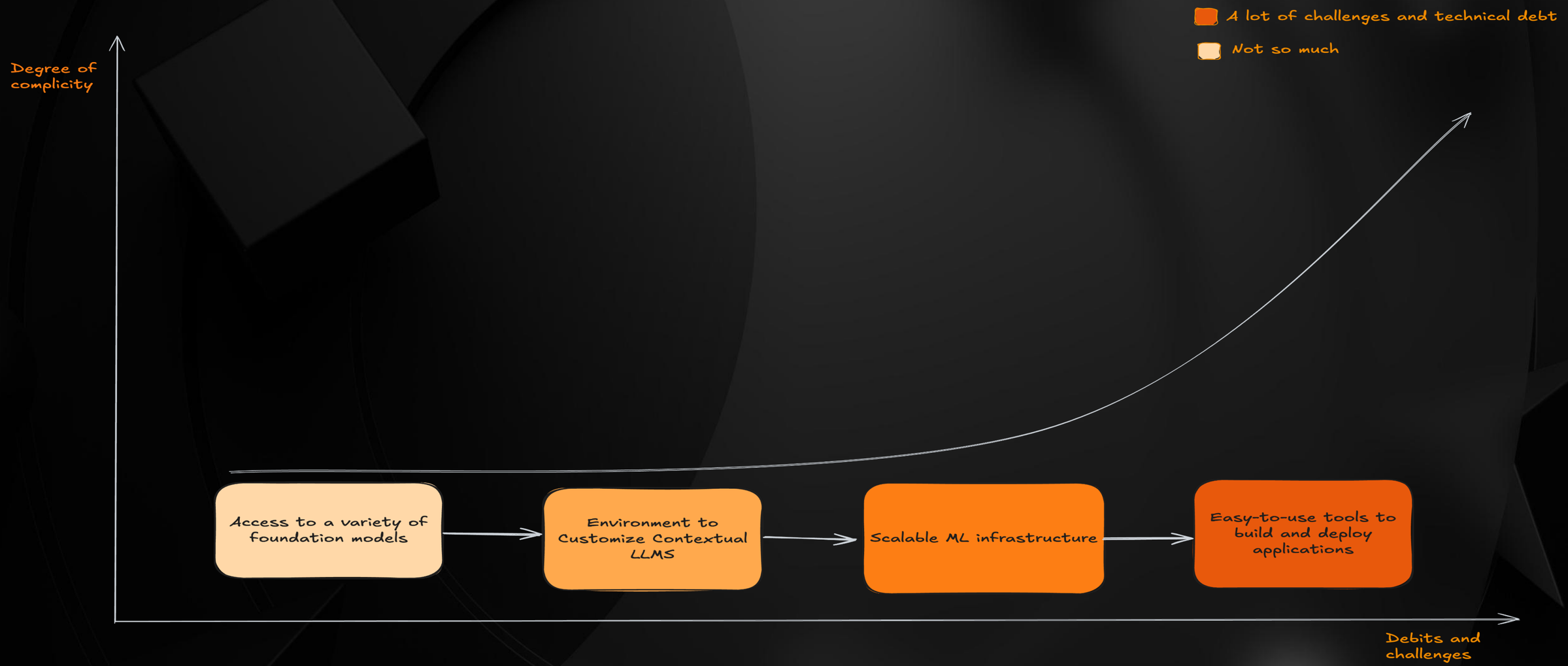
Feb 1, 2024 - CNBC Article

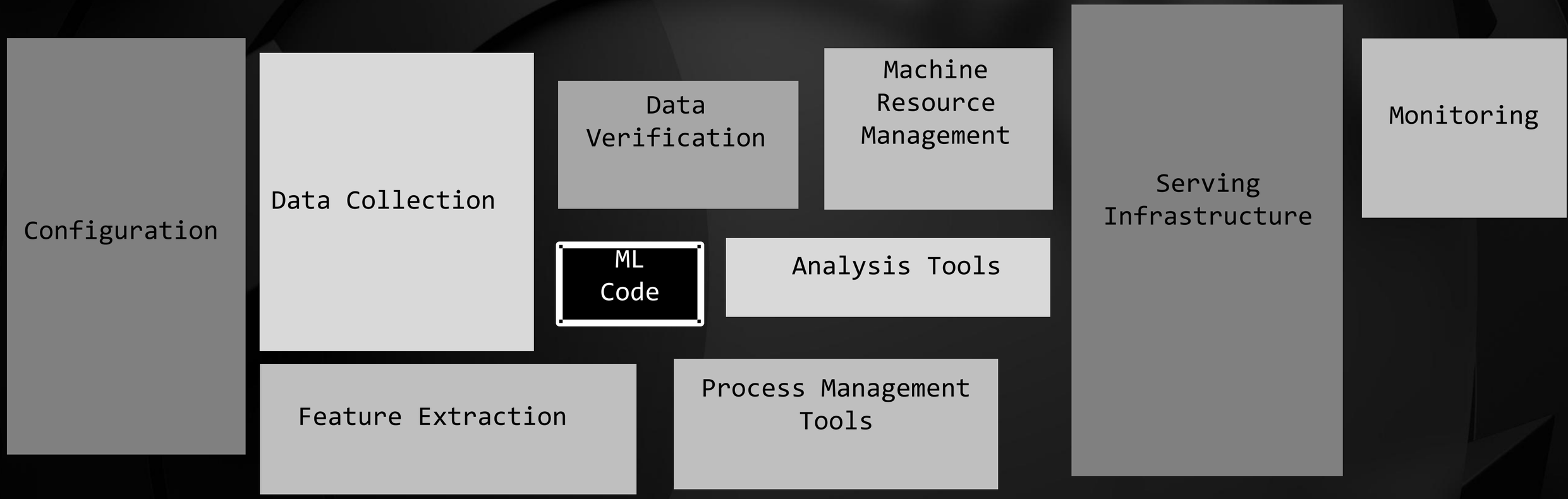
Essentials for building a **Generative AI Application**

1. Access to a variety of foundational models
2. Environment to customize contextual LLMs
3. Easy-to-use tools to build and deploy
4. Scalable ML infrastructure



Essentials for building a generative AI application

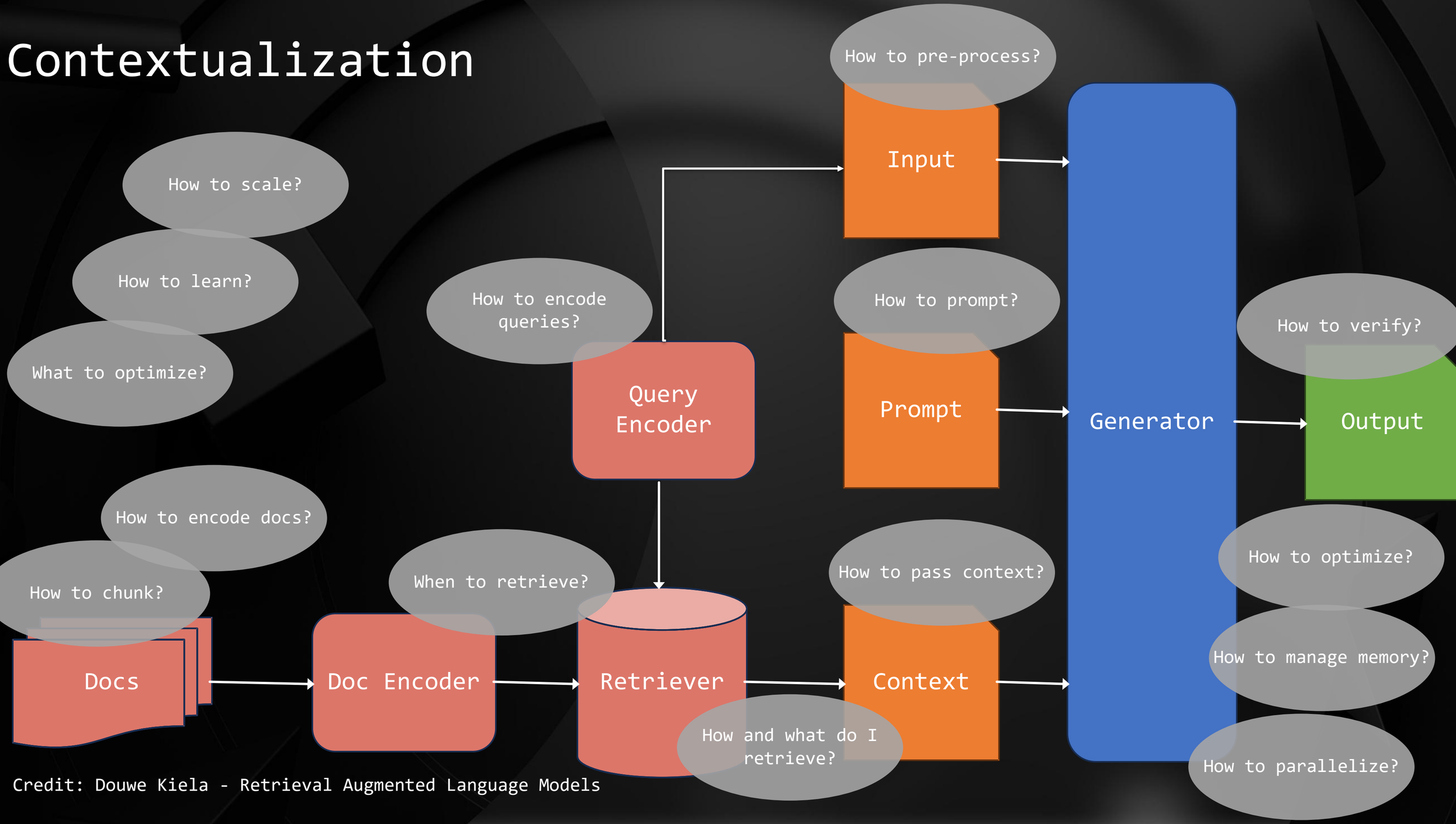




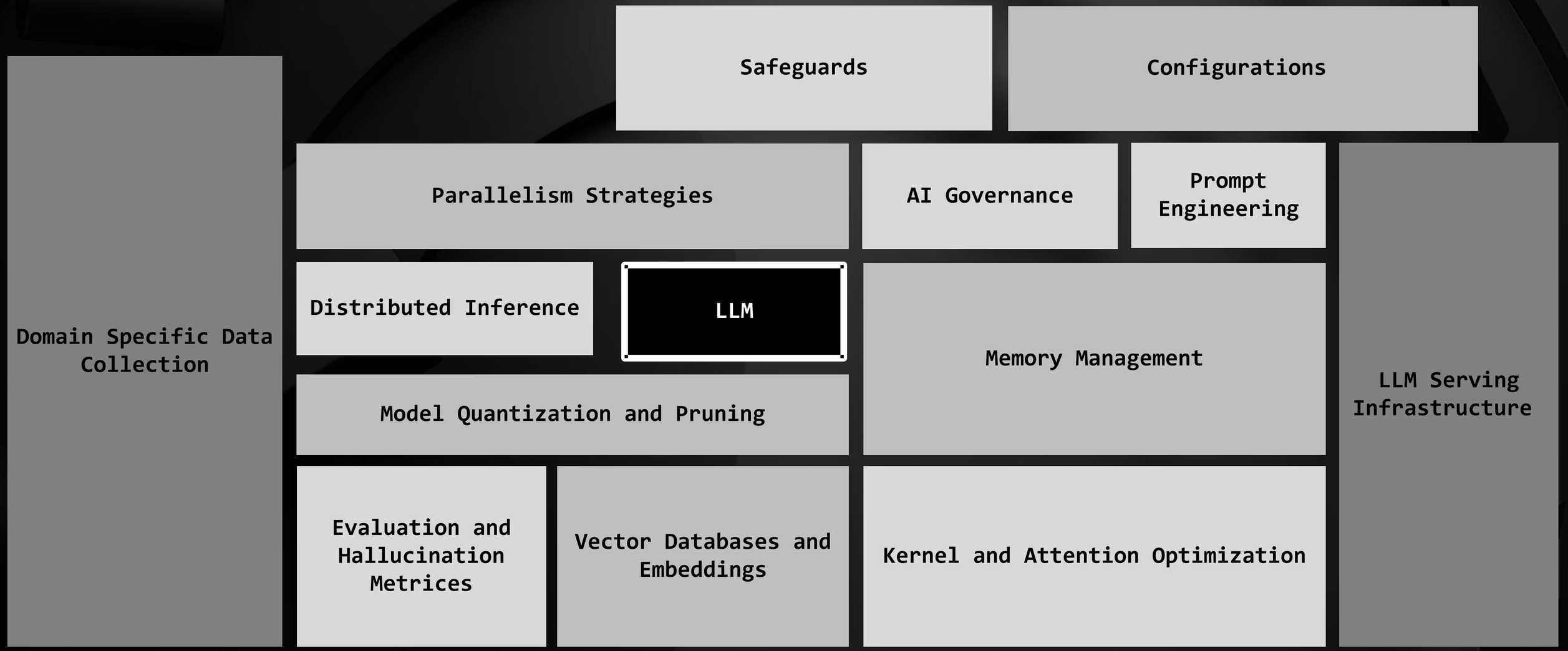
Vast and complex Surrounding challenges and technical debt around ML deployment*

* D. Sculley et al. 2015. Hidden technical debt in Machine learning systems. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'15). MIT Press, Cambridge, MA, USA, 2503–2511.

Contextualization



Credit: Douwe Kiela - Retrieval Augmented Language Models

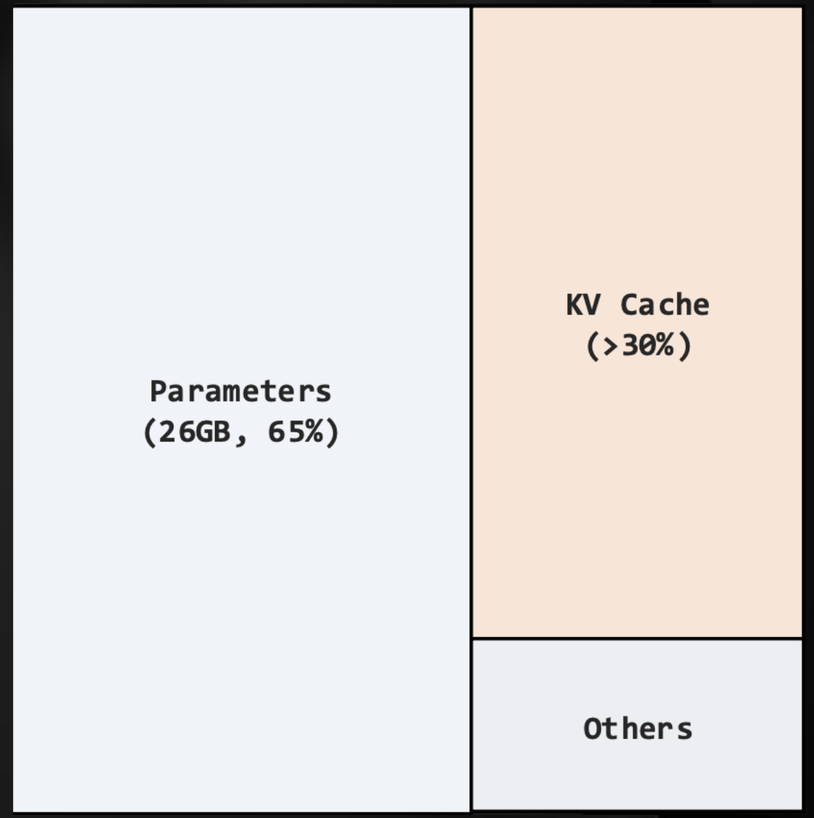


Vast and complex Surrounding challenges and technical debt around LLM deployment*

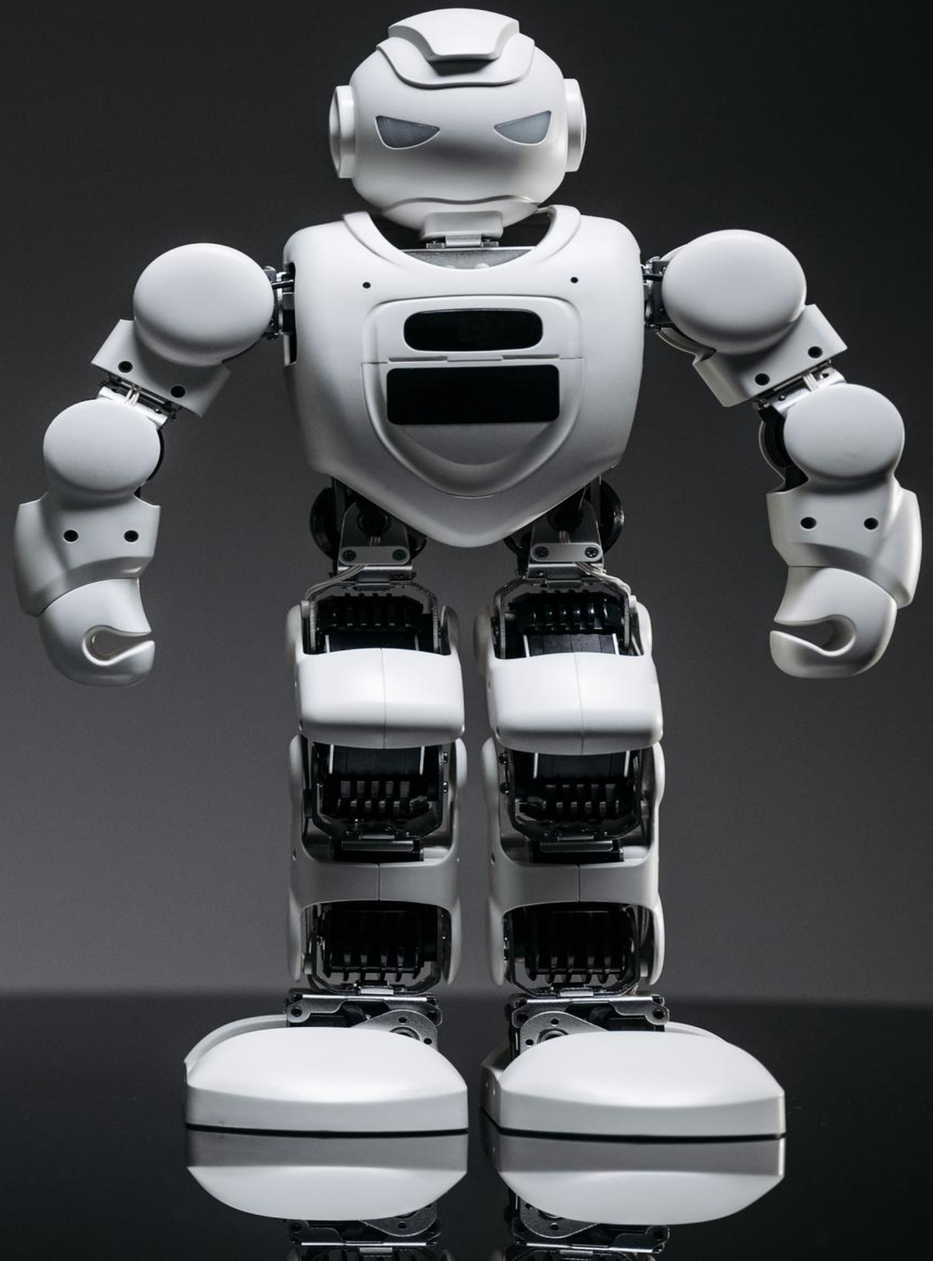
* Menshawy, A., Nawaz, Z., Fahmy, M., & Minervini, P. (2024). Navigating Challenges and Technical Debt in Large Language Models Deployment. Proceedings of the EuroMLSys '24. ACM Inc., New York, NY

Key Factors for Scalable and Modular GenAI Production

- Fully utilizes GPU's capabilities
- Efficiently scale with data and user demands
- Reducing latency and boosting throughput with optimized serving and batching strategies
- Backed by active communities
- Modular designs for easy updates and integration

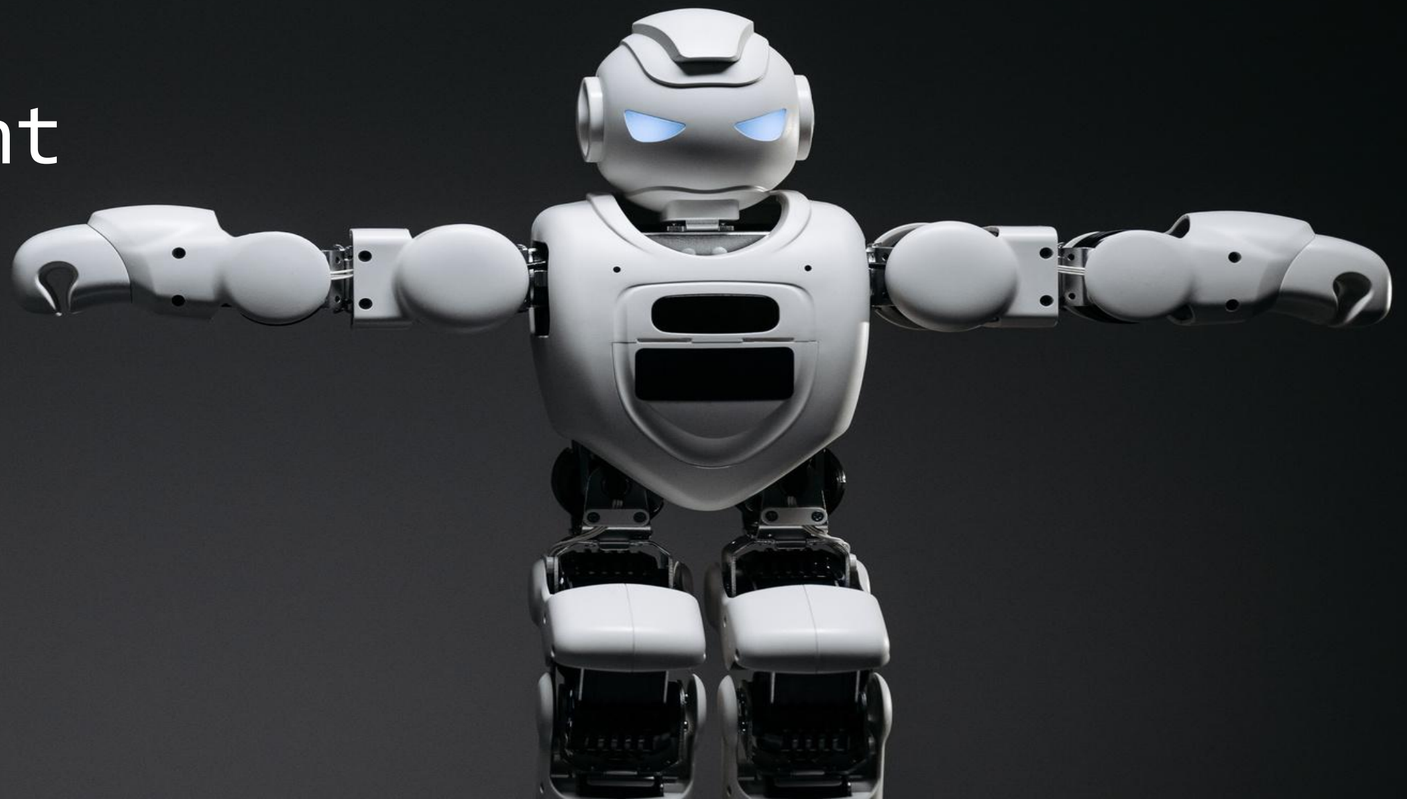


The Rise of AI Agents



Core Ideas of building your agent

- Don’t build agents for everything
- Keep it simple
- Think like your agent

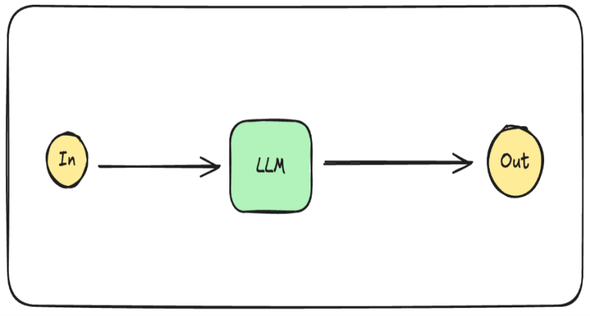


Don’t build agents for everything

Non-Agentic Workflows (ZERO-SHOT)

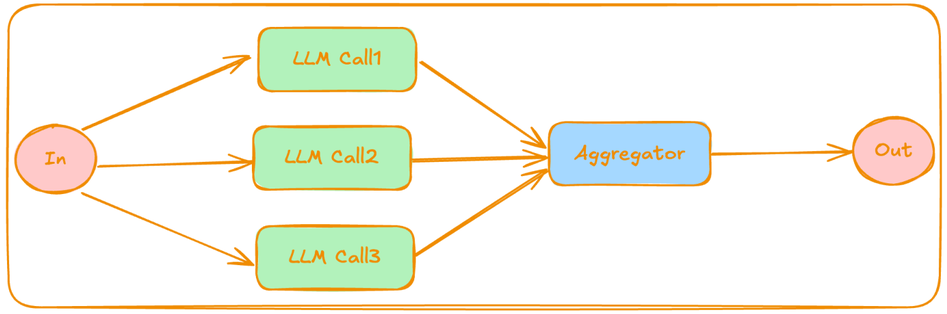
Single LLM

Text generations, classifications, extractions



Workflows

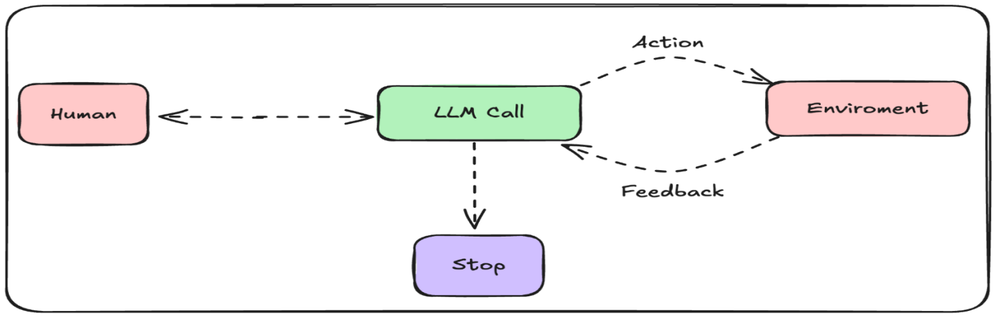
LLMs orchestrations (by code)



Agents

Text generations, classifications, extractions

Agentic
Agentic protocols
Capability
Cost/latency/Consequences



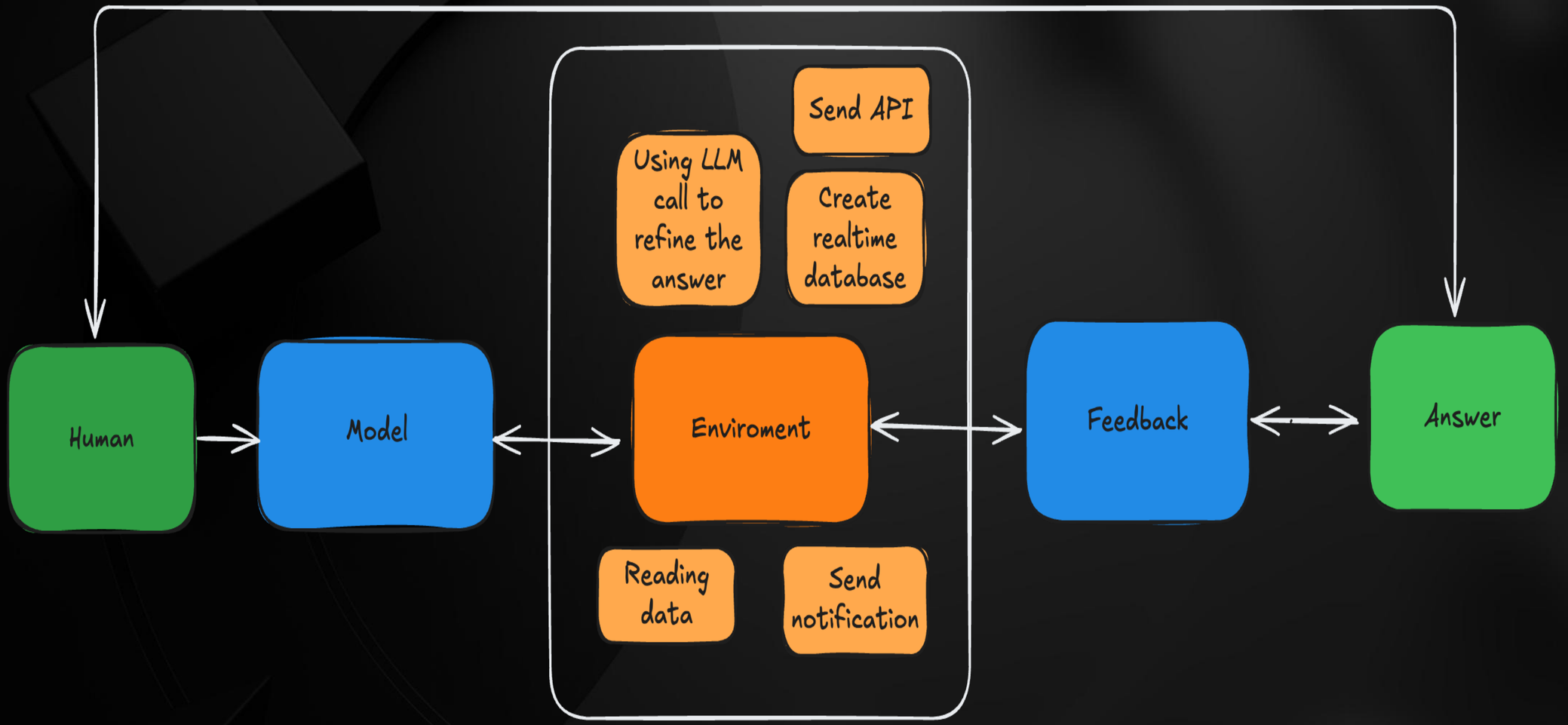
Physical AI

Robotics, Fully autonomous Vehicles

Slide source: https://www.anthropic.com/engineering/building-effective-agents

Keep it simple

“Agents are just models using tools in a loop”



Think like your agent!!

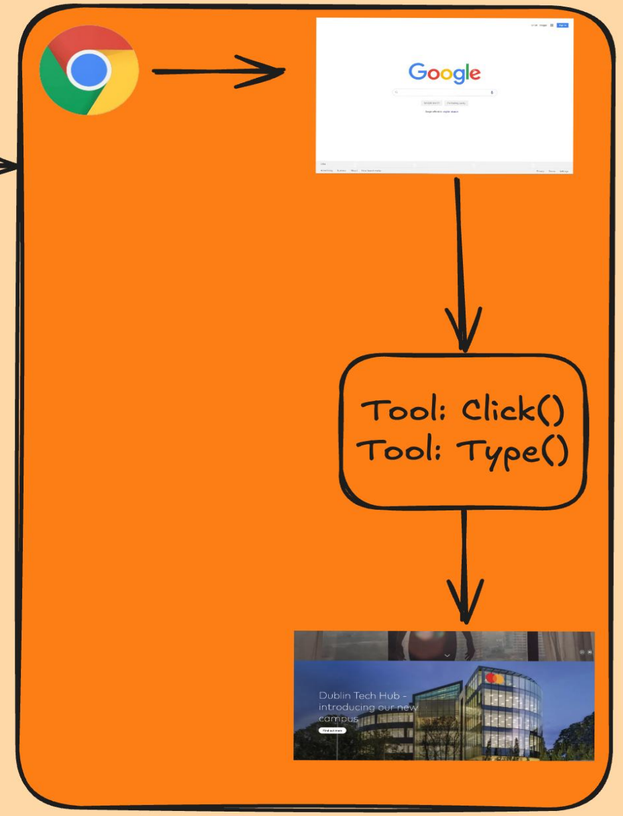
Search about Mastercard global

you are a computer use agent:

you can move your mouse, click, type and use key.
your task is search on google about Mastercard global

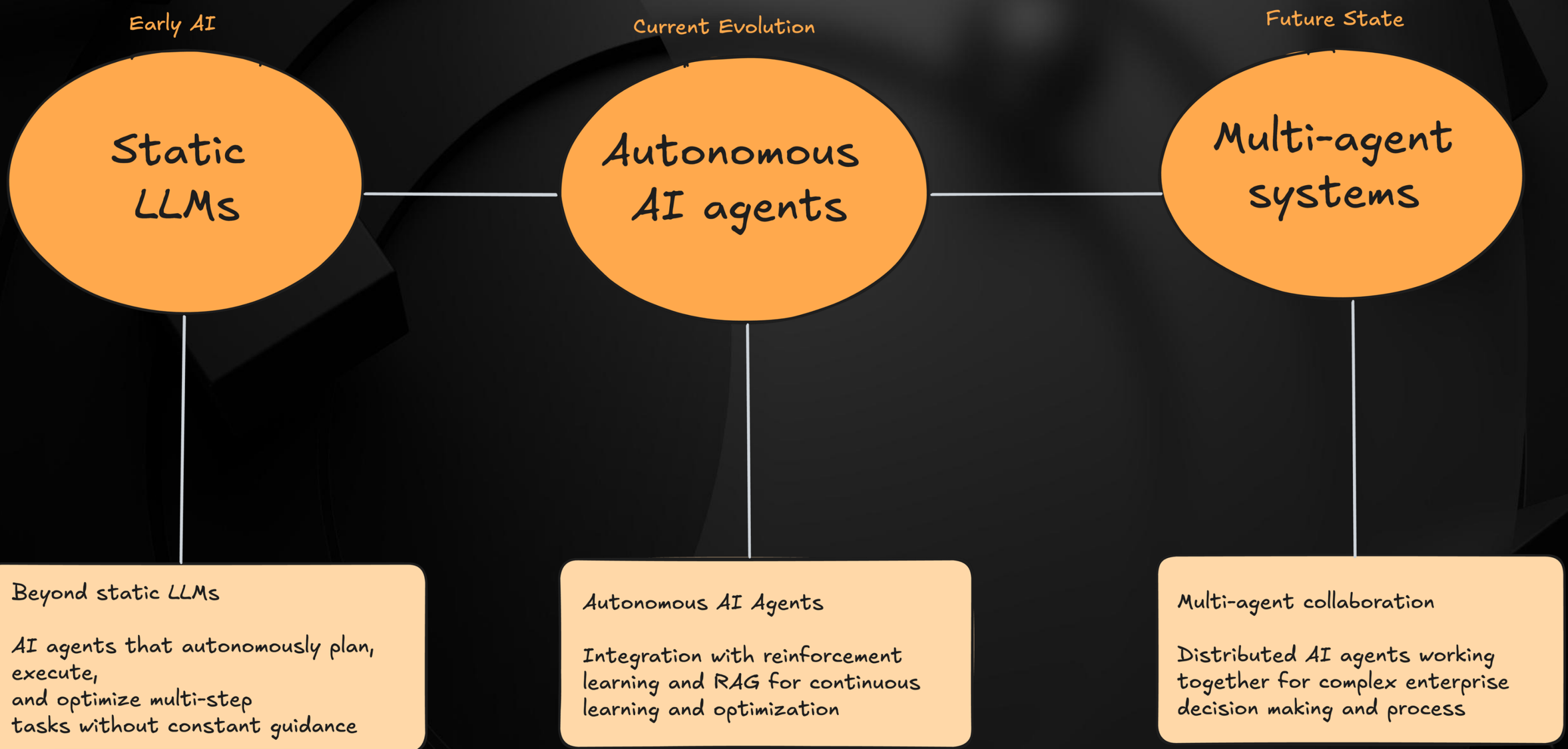
Tool: move()
Tool: Click()

close your eyes
3:6 secs



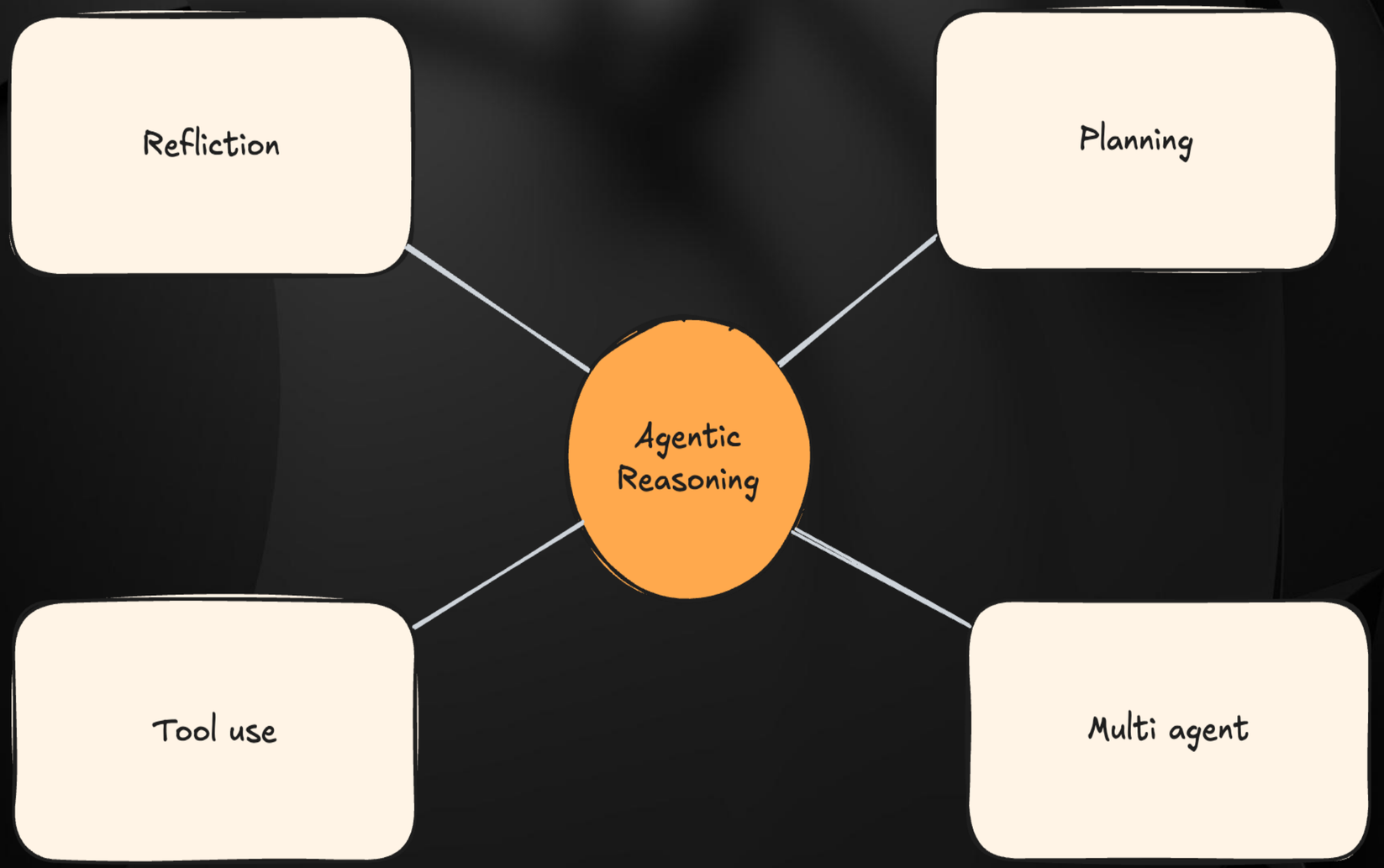
"Think like your agent"

Agentic AI: The Next Evolution in Enterprise AI



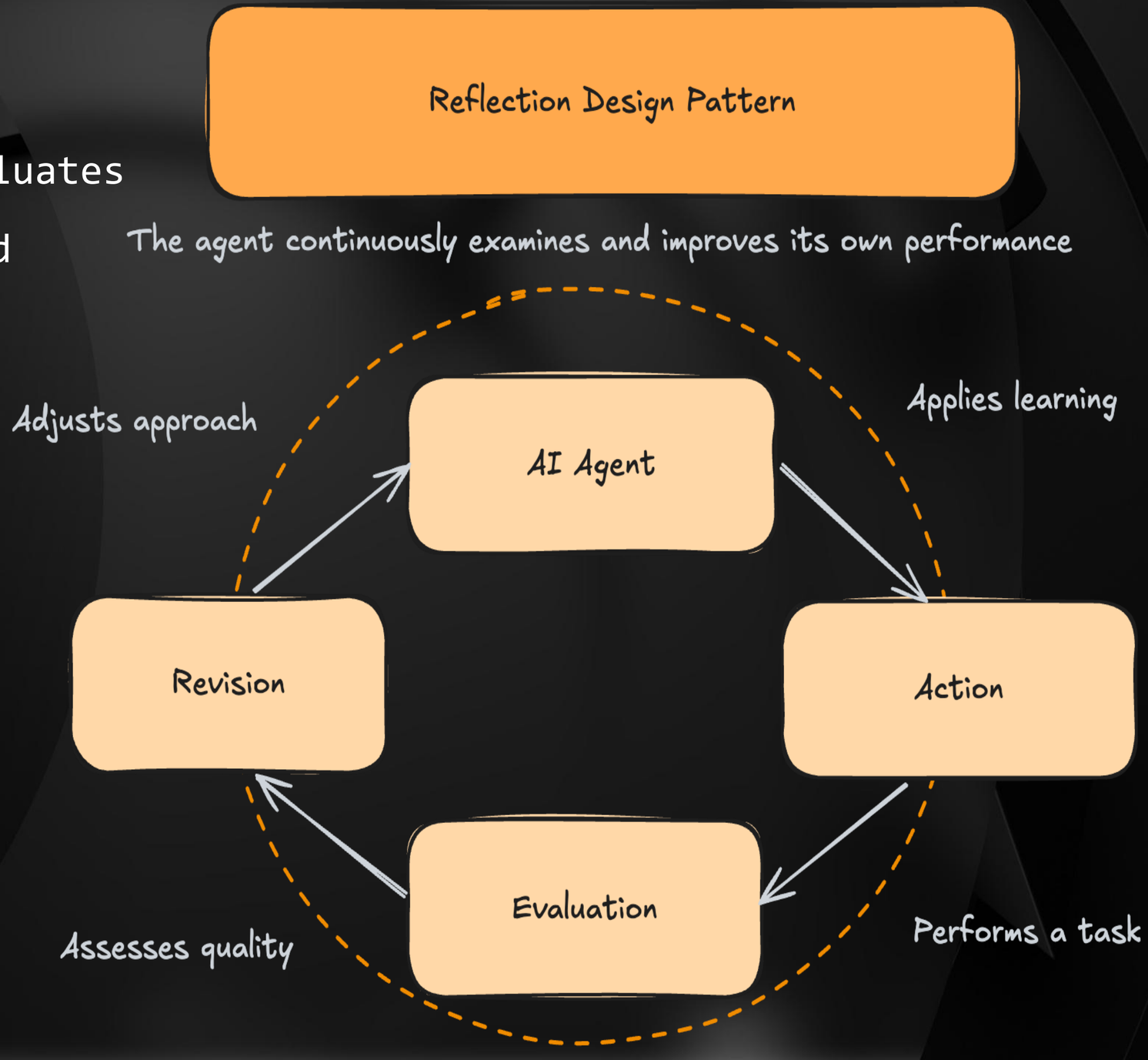
Agentic Reasoning Design Patterns

- Reflection
- Tool use
- Planning
- Multi-agents



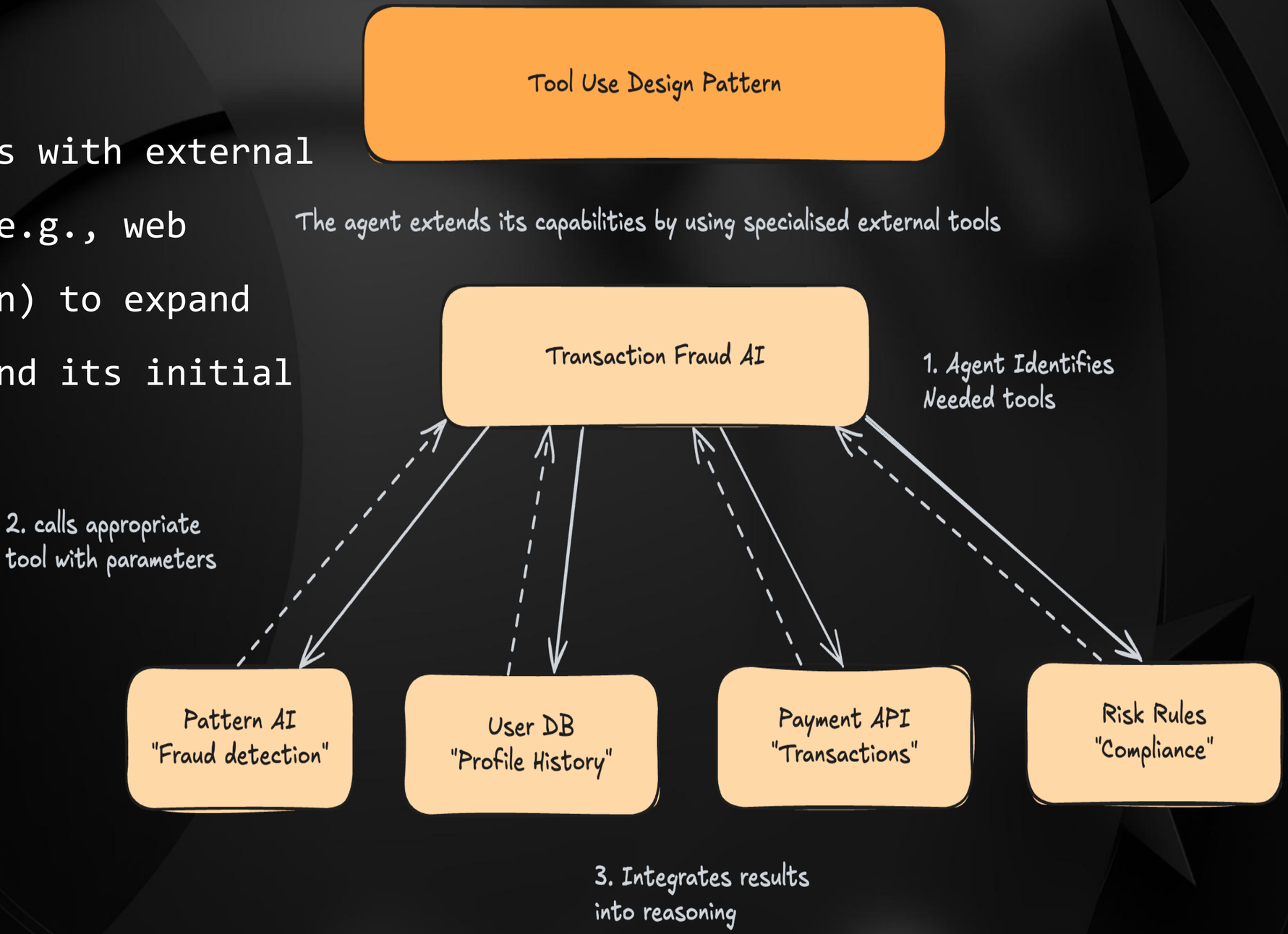
Reflection

The AI agent self-monitors and evaluates its actions, identifying errors and refining strategies for improved performance.



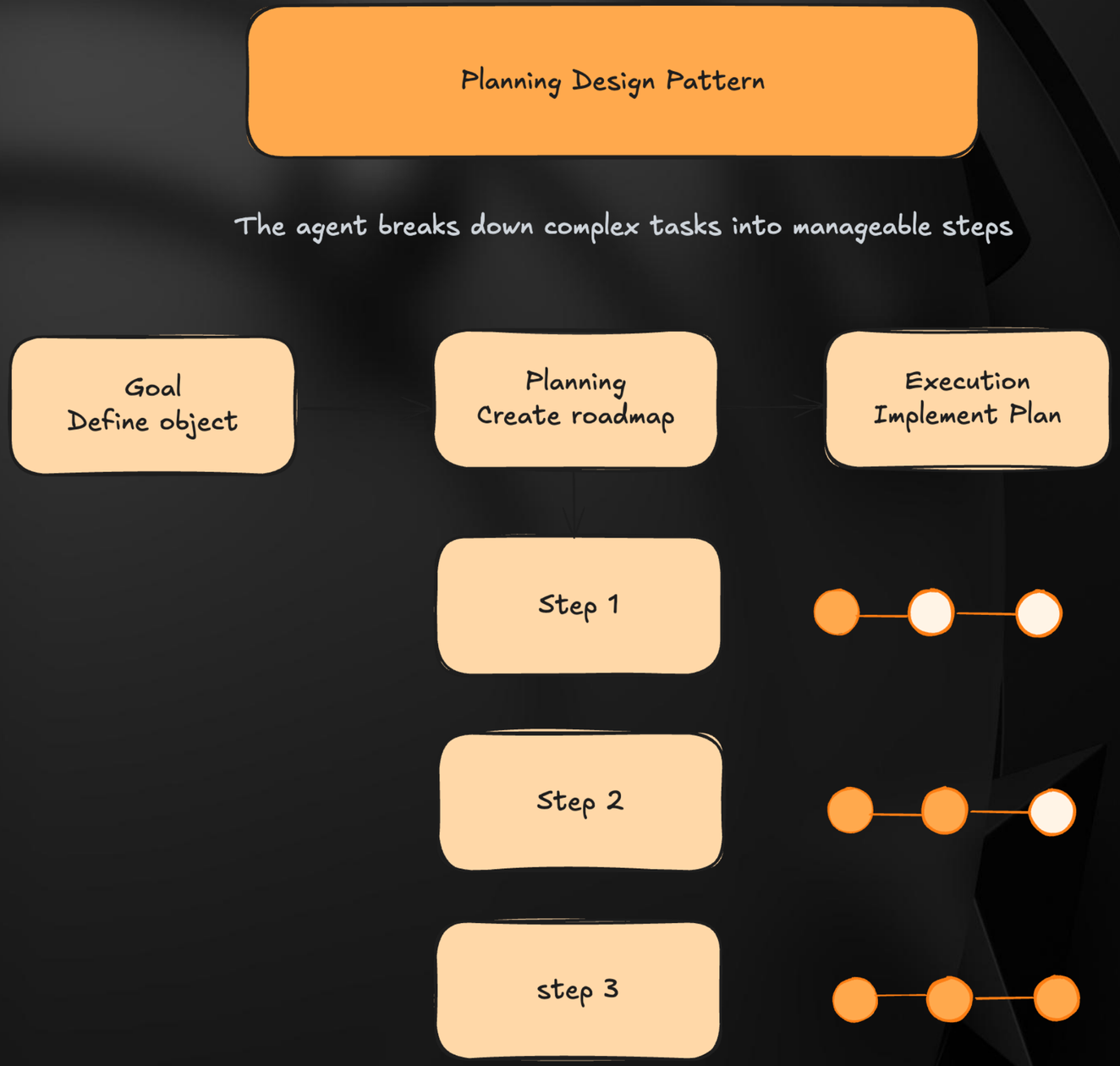
Tool Use

The AI agent interacts with external tools and resources (e.g., web search, code execution) to expand its capabilities beyond its initial training.



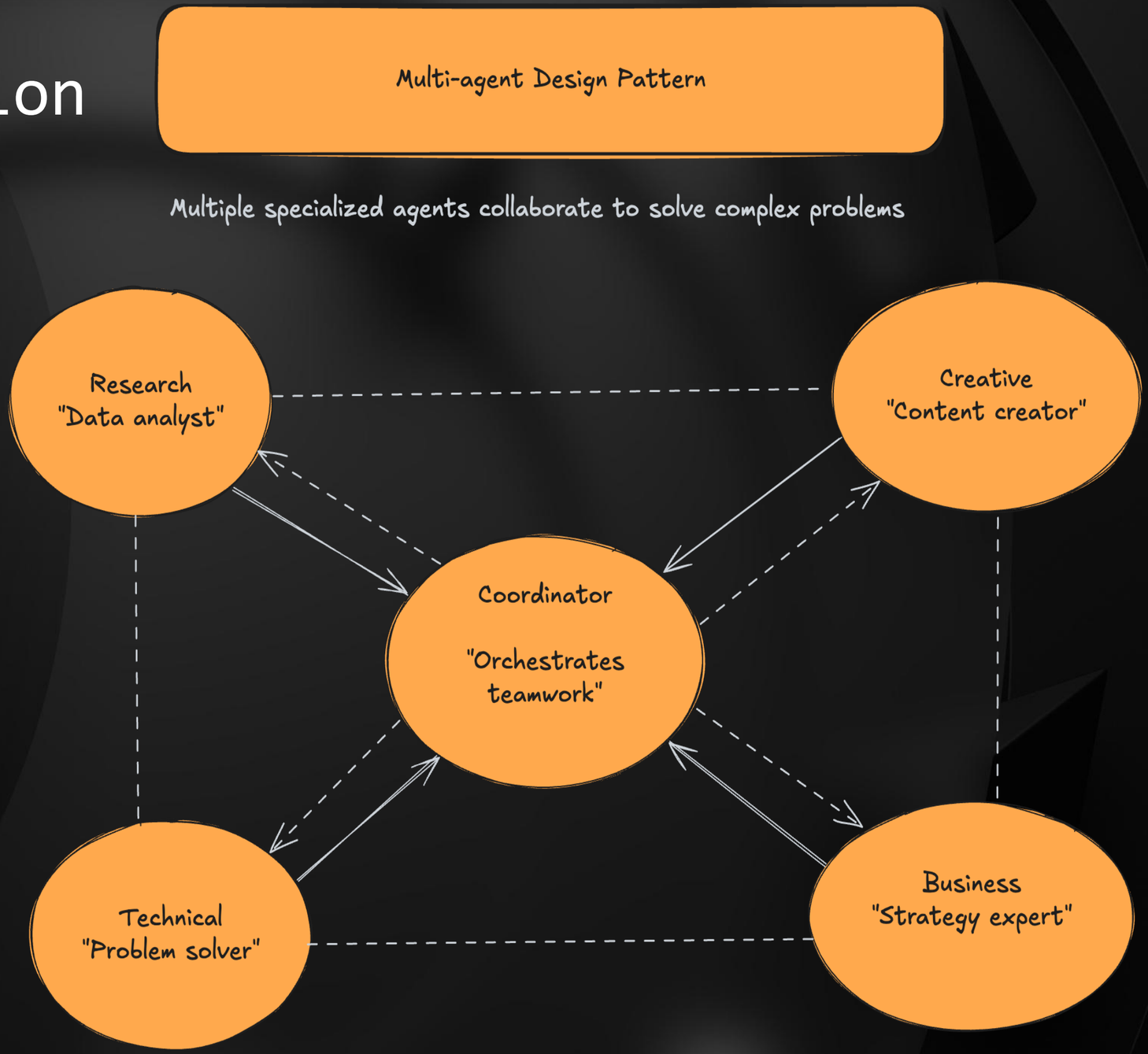
Planning

Business Example: Product Launch Strategy The AI divides the launch into sequential phases: market research, product positioning, marketing campaign, and distribution logistics.



Multi-agent collaboration

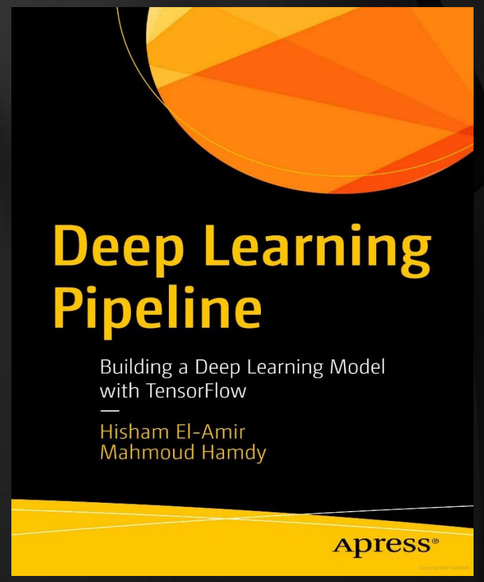
Multiple AI agents with specialized roles communicate and coordinate to solve complex problems more efficiently.

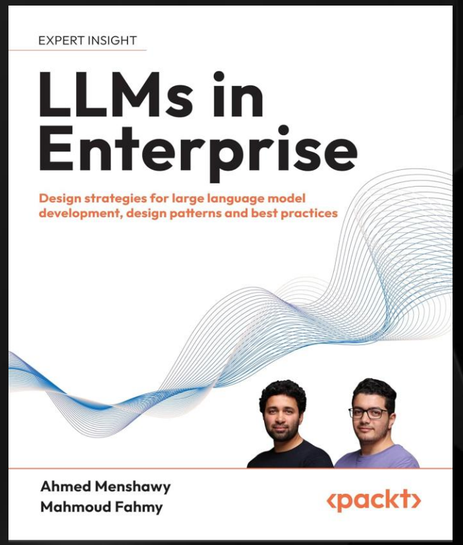


Let’s combine all of them in one equation

Thank you!

Mahmoud Fahmy
Lead - AI Engineering, Mastercard


Deep Learning Pipeline
Building a Deep Learning Model with TensorFlow
Hisham El-Amir
Mahmoud Hamdy
Apress


EXPERT INSIGHT
LLMs in Enterprise
Design strategies for large language model development, design patterns and best practices
Ahmed Menshawy
Mahmoud Fahmy
packt

Formulating your GenAI strategy and crafting the AI capability roadmap in Fintech

Vision alignment	• What fintech goals do you aim to achieve? Fraud detection, risk management, customer experience enhancement, or cost efficiency? • Ensure AI projects support strategic fintech objectives such as real-time fraud prevention
Assess Capabilities	• Ensure AI integrates seamlessly with core banking systems, digital wallets, and payment networks. • Access to high-quality, real-time financial transaction data is critical.
Meticulous Execution	• Test GenAI applications in limited fintech operations, such as automated claims processing or fraud monitoring.
Scaling Up	• Deploy AI-driven solutions across more fintech functions, from risk management to customer onboarding automation
Risk Management	• Protect customer financial data using robust AI-driven cybersecurity measures.
Organizational Change	• Promote AI-driven experimentation and adoption within financial services.