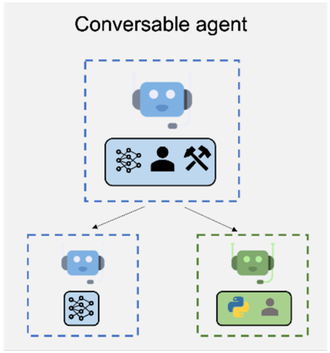
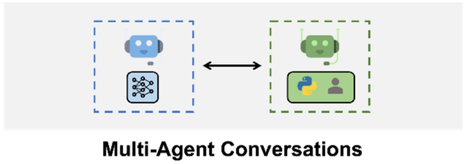
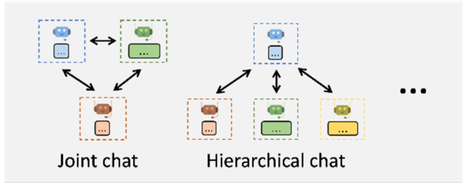
Agents, Agents Everywhere….



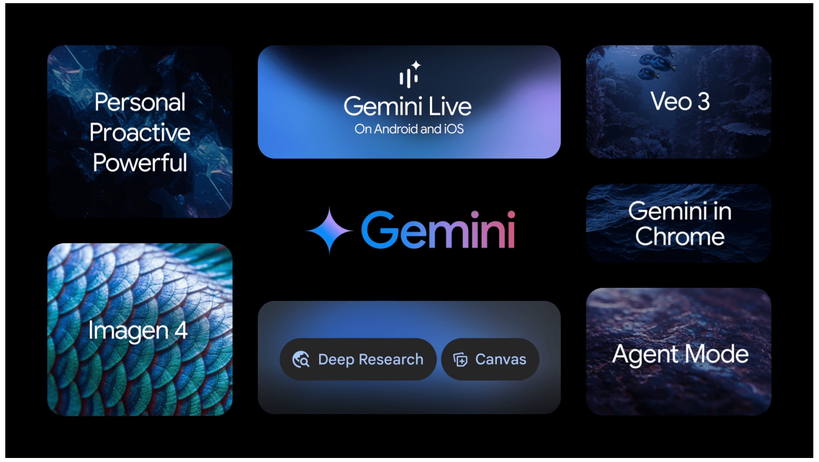
Agent Customization

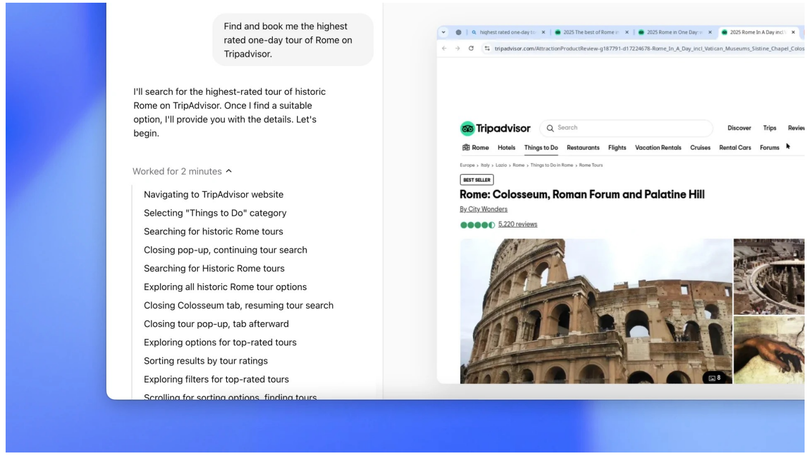


Multi-Agent Conversations



Flexible Conversation Patterns





Operator by OpenAI

"SaaS is Dead"
All software applications that we know today are just fancy interfaces sitting on databases

-Satya Nadella

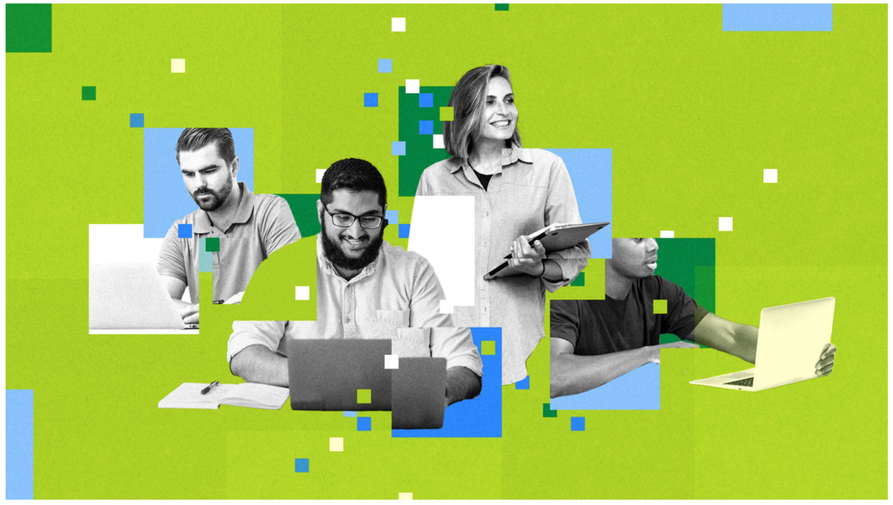
Claude Computer

Generative AI

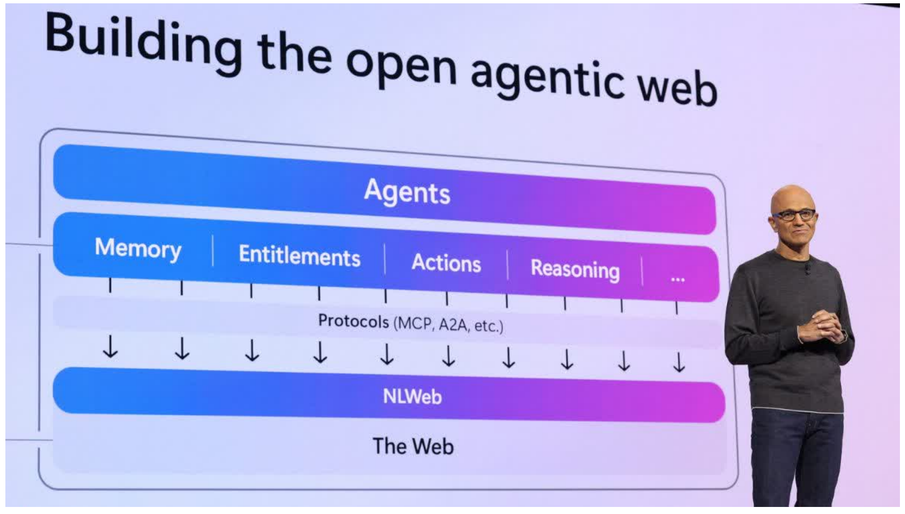
Agentic AI Is Already Changing the Workforce

by Jen Stave, Ryan Kurt and John Winsor

May 22, 2025



HBR Staff/Unsplash



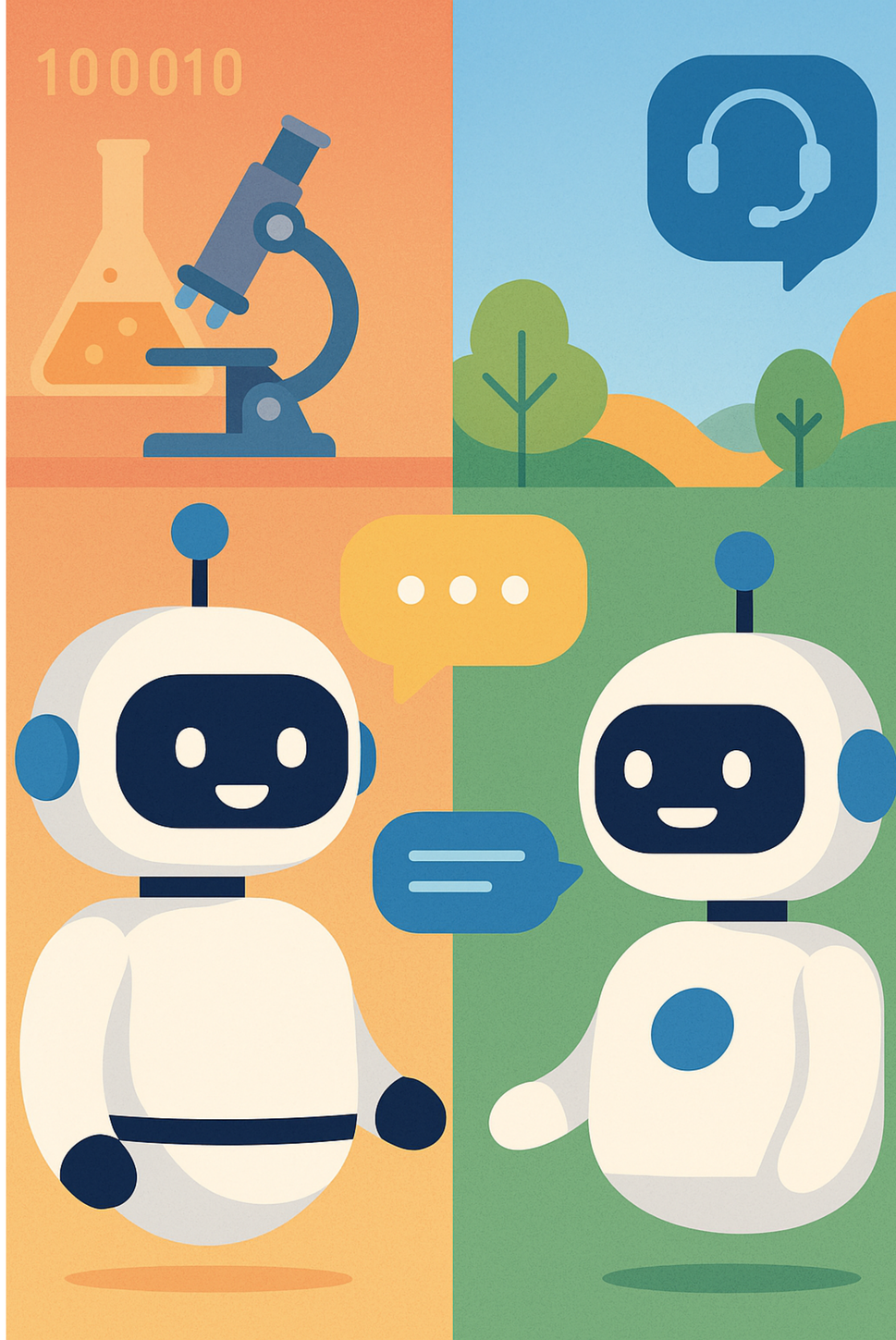
Source: openai.com, microsoft, HBR, claude, google

Engineering Agentic AI Systems: From Lab to Land

Dr. Karthik Vaidhyanathan

TechForward Research Seminar Series

May 30, 2025



ABOUT ME

Logic takes you from A to B, Immagination takes you elsewhere -- Albert Einstein

Karthik Vaidhyanathan

Assistant Professor

Software Engineering Research Center and

Leadership Member, Smart City Research Center

IIIT Hyderabad, India

Research Interests

AI4SA

1. AI for Architectural Knowledge
2. AI for self-adaptation

SA4AI

1. Sustainable AI-enabled systems
2. Self-adaptive AI Systems (Edge-Cloud)

Education

Double Master Degree - Software Architecture and Machine Learning

PhD from GSSI, Italy

Postdoc, University of L'Aquila, Italy

Fun Facts!

1. Cricket fanatic!
2. Movie buff!!
3. From God's own Country!!

https://karthikvaidhyanathan.com /in/karthikv1392/ karthi_ishere karthik.vaidhyanathan@iiit.ac.in

Software Engineering Research Center (SERC)

Aims to research and develop state of art techniques, methods and tools in various areas of software engineering and programming languages.

Raghu Reddy
Associate Professor and Center Head
raghu.reddy@iiit.ac.in

Venkatesh Choppella
Associate Professor
venkatesh.choppella@iiit.ac.in

Karthik Vaidhyanathan
Assistant Professor
karthik.vaidhyanathan@iiit.ac.in

Raman Saxena
Professor of Practice
raman.saxena@iiit.ac.in

Vasudeva Varma
Professor
vv@iiit.ac.in

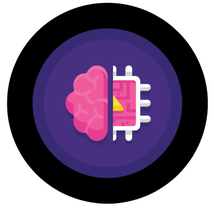
Viswanath Kasturi
Research Professor of Eminence
viswanath.iiithyd@gmail.com

Ramesh Loganathan
Professor of Practice
ramesh.loganathan@iiit.ac.in

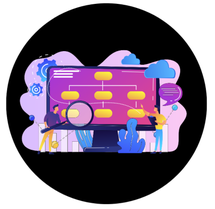
Prakash Yalla
Professor of Practice

Abhishek Kumar Singh
Assistant Professor
abhishek.singh@iiit.ac.in

VR and AR



SE and AI



Formal Methods

Software Quality

Gamification

Computing Education

HCI

Programming Languages

Self-adaptive Systems

Software Analytics

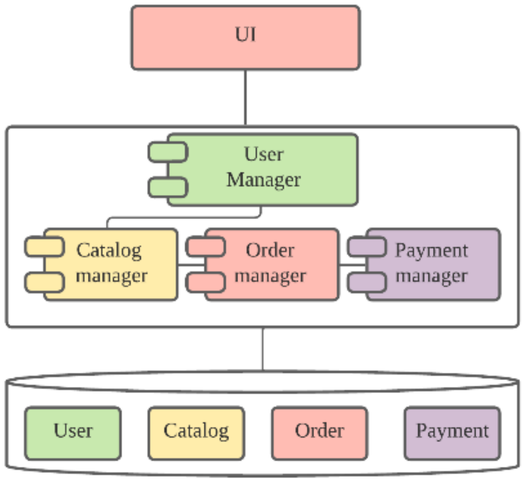
Software Sustainability

IoT

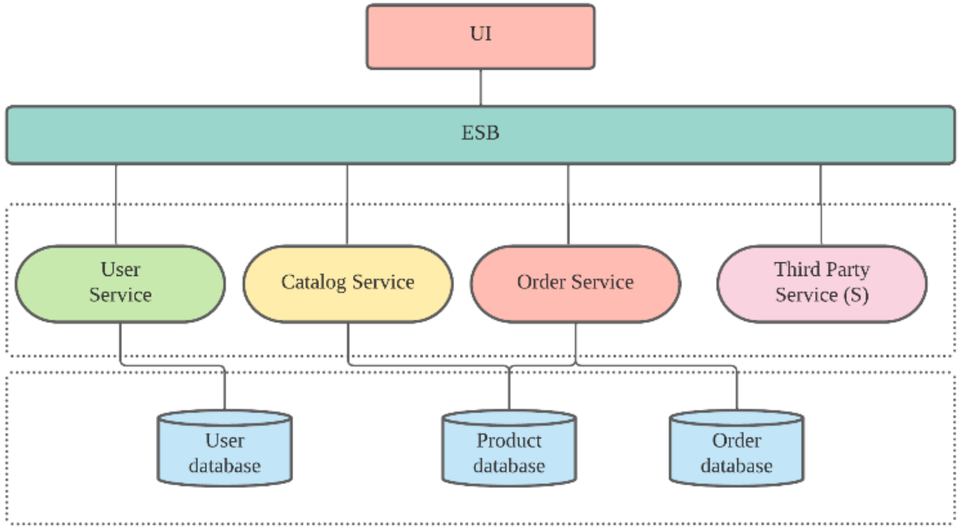
More info: https://serc.iiit.ac.in

Software Systems Evolution Over the Years

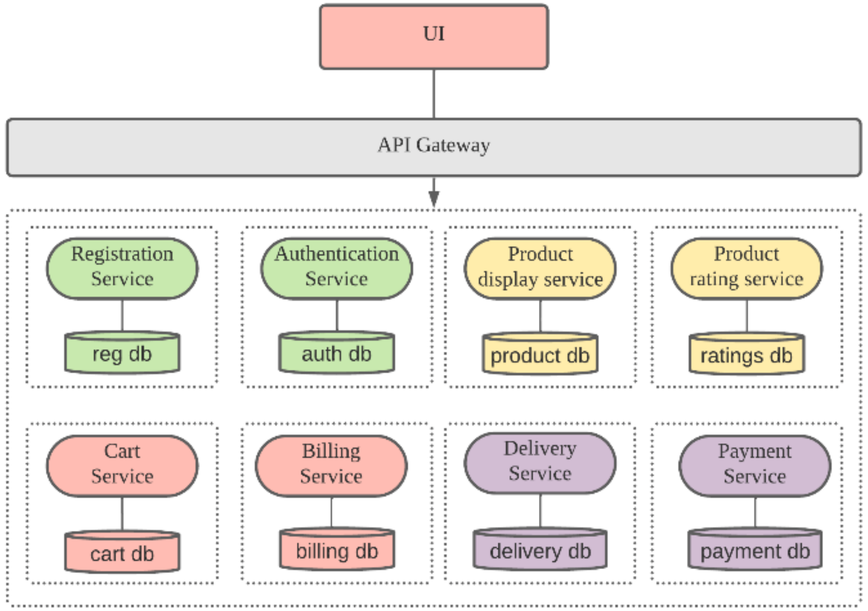
Monoliths



SOA



Microservices



1990 2000 2010 2020

Serverless, age of intelligent connected systems….

AI Over the years….

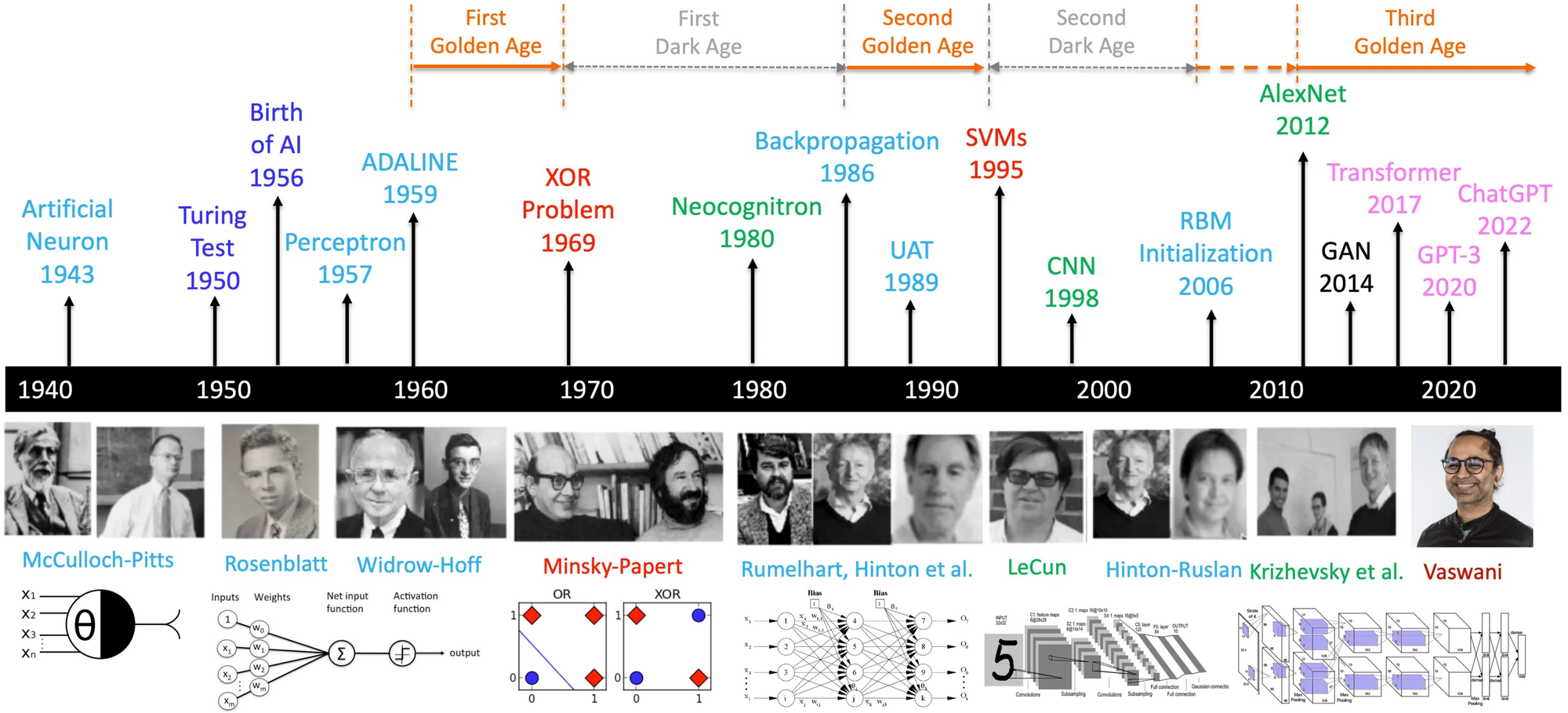
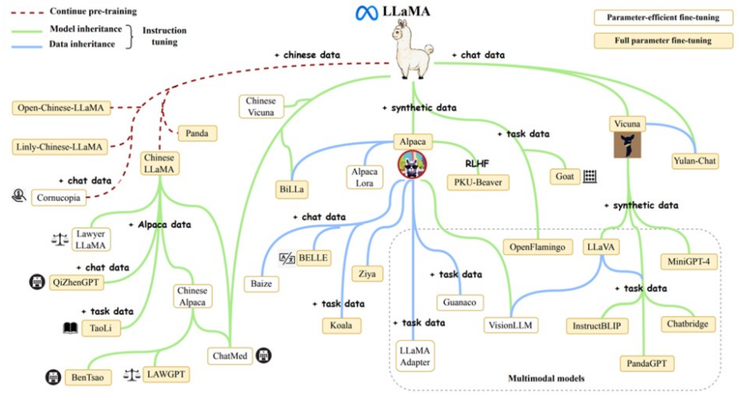


Image source: https://medium.com/@lmpo/a-brief-history-of-ai-with-deep-learning-26f7948bc87b

Age of LLMs and Agentic AI

Proprietary

Open Source



Framework for AI Agents

Leaderboard Overview

See how leading models stack up across text, image, vision, and beyond. This page gives you a snapshot of each Arena, you can explore deeper insights in their dedicated tabs. Learn more about it here.

View Blog

Text — 5 days ago

Rank (UB)	Model	Score	Votes
1	gemini-2.5-pro-preview-05-06	1446	6,115
1	o3-2025-04-16	1435	7,921
2	chatgpt-4o-latest-20250326	1422	10,280
3	gpt-4.5-preview-2025-02-27	1417	15,276
3	gemini-2.5-flash-preview-05-…	1415	3,892

WebDev — 5 days ago

Rank (UB)	Model	Score	Votes
1	Gemini-2.5-Pro-Preview-05-06	1415	3,464
2	Claude 3.7 Sonnet (20250219)	1357	7,481
3	Gemini-2.5-Flash-Preview-05-…	1310	981
4	GPT-4.1-2025-04-14	1257	4,880
5	Claude 3.5 Sonnet (20241022)	1238	26,338

Lite | Verified | Full | Multimodal

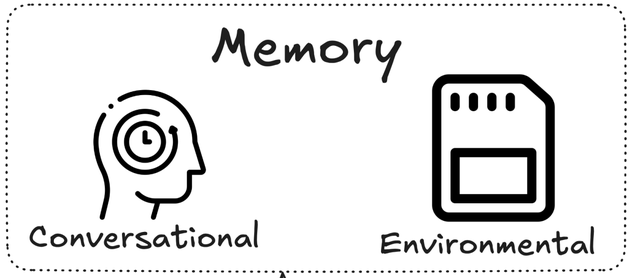
Open Weight Model | Open Source System | Checked | (All Tags Selected)

Model	% Resolved	Org	Date	Logs	Trajs	Site
SWE-agent + Claude 3.7 Sonnet	48.00		2025-02-26	✓	✓	
DARS Agent	47.00		2025-02-05	✓	✓	
Lingxi	42.67		2025-05-09	✓	✓	
OpenHands + CodeAct v2.1 (claude-3-5-sonnet-20241022)	41.67		2024-10-25	✓	✓	
PatchKitty-0.9 + Claude-3.5 Sonnet (20241022)	41.33		2024-12-20	✓	✓	-
Composio SWE-Kit (2024-10-30)	41.00		2024-10-30	✓	✓	

Source: lmarena, swe-bench

Agents: What are they made up of?

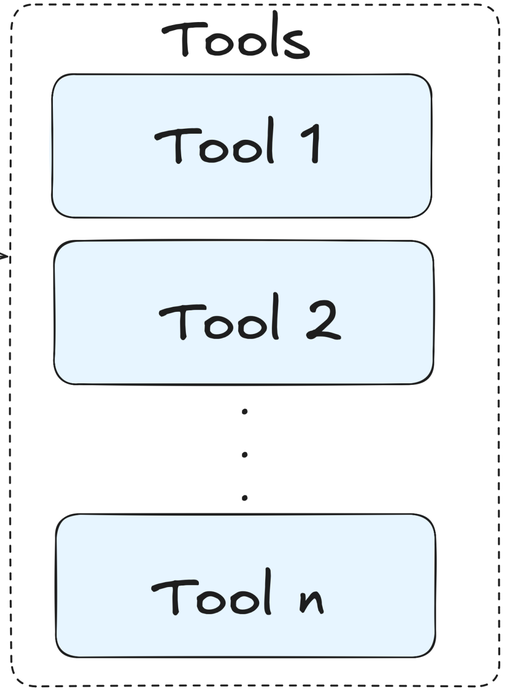
eg: What kind of movie user likes?



requires

LLM Agent ←Uses→

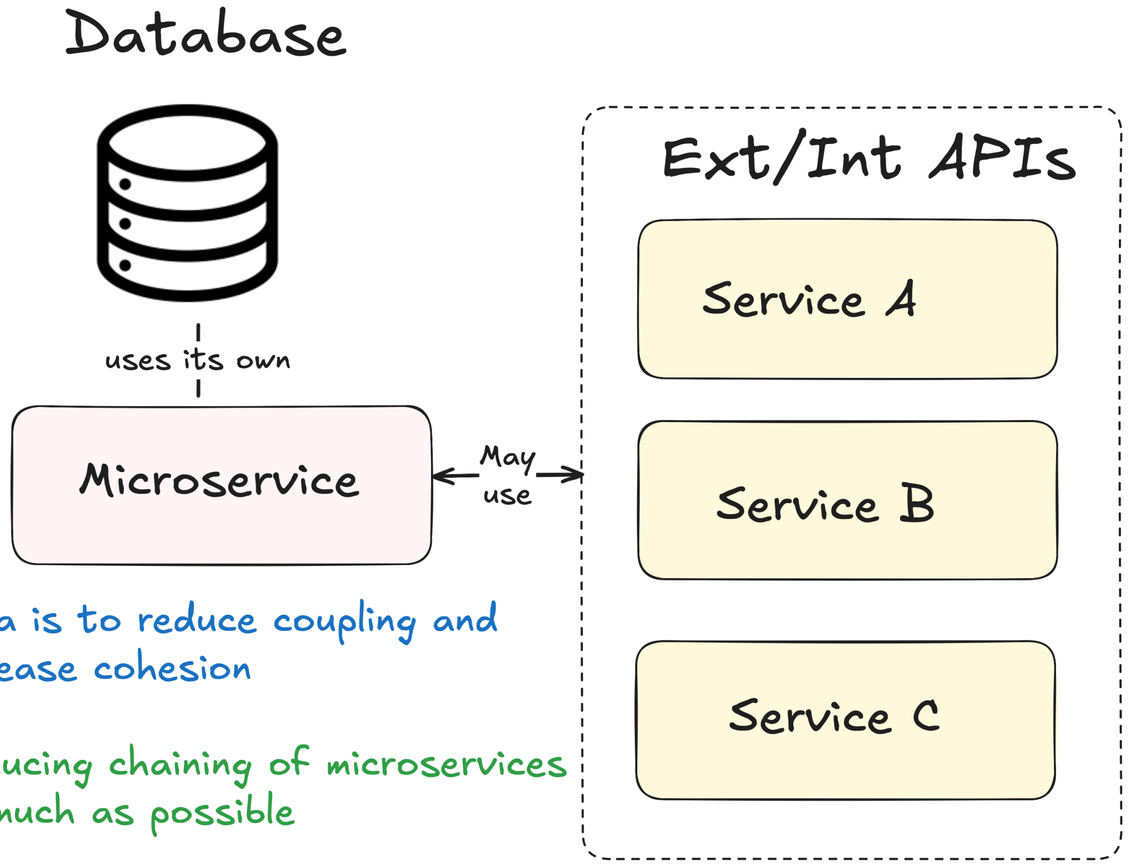
eg: Web scrapper



Not necessary that every agent needs an LLM. It could be any AI in magnetic AI setup.

One of the tool itself can be another LLM call. Tool can be anything. It could be a service offering a functionality

Autonomous entity that senses and responds to its environment and take actions to achieve its goals

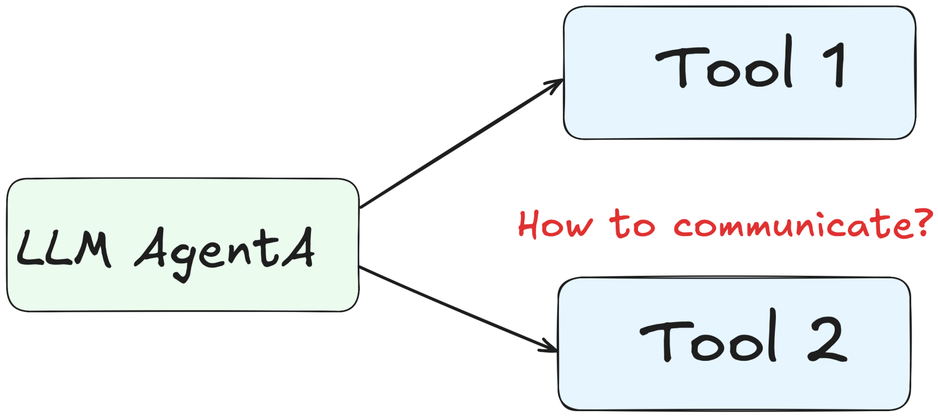


Idea is to reduce coupling and increase cohesion

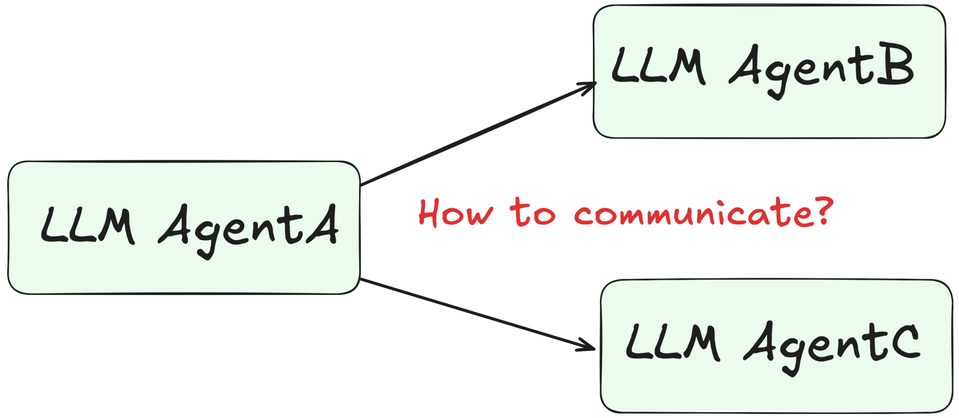
Reducing chaining of microservices as much as possible

Suite of small autonomous services that communicate with each other using light weight protocols

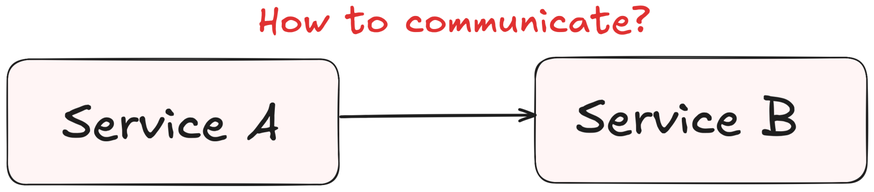
Communication between Agents and/or Tools/Agents



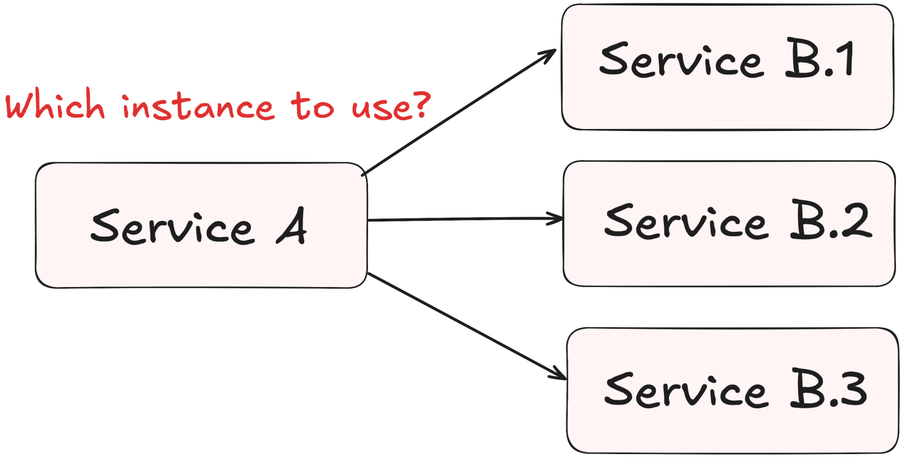
1. Tools can be added dynamically
2. Each tool may have different way of invocation
3. Use protocols like MCP (Model Context Protocol)



1. An agent can be added dynamically
2. Each agent may have different way of invocation
3. Use protocols like A2A (Agent to Agent)

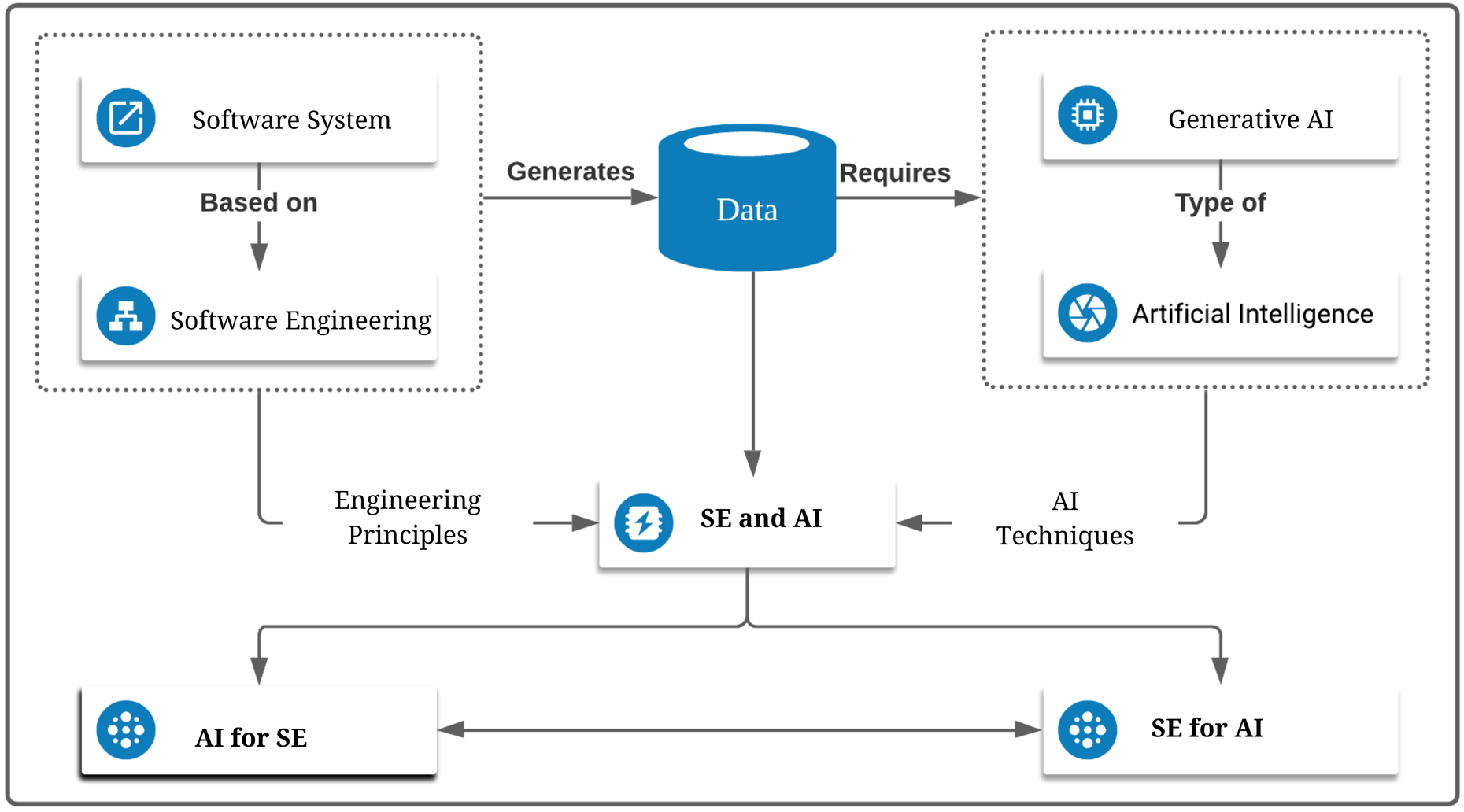


1. Use lightweight protocols like HTTP. API defined
2. Sync vs Async, Orchestration Vs Choreography
3. JSON/Protobuf as the data format

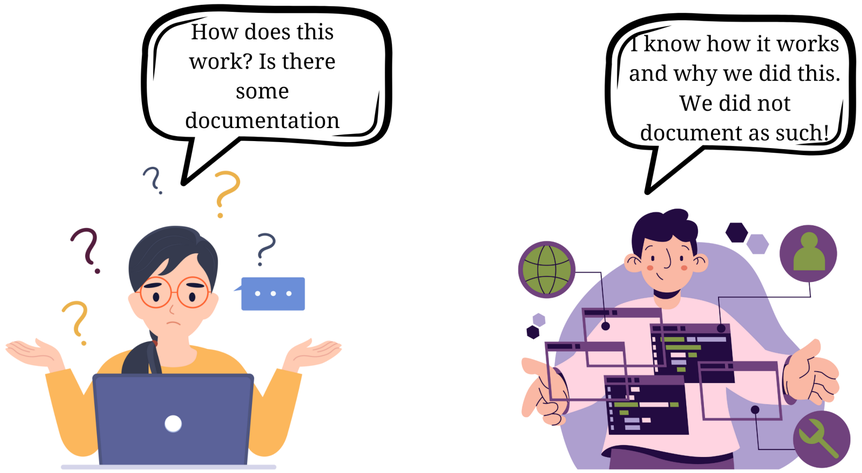


1. Service discovery - Client vs server
2. Services register to Service Registry
3. Eg: Netflix Eureka, Amazon ELB, zookeeper

At the Intersection of SE and AI

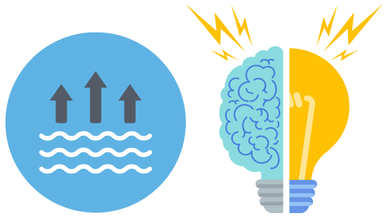


From Lab: AI4SE - LLMs for Architecture Support



Takes away the knowledge!

What happens when the dev/architect leaves?



Knowledge Vaporisation!

Can LLMs Generate Architectural Design Decisions? - An Exploratory Empirical study

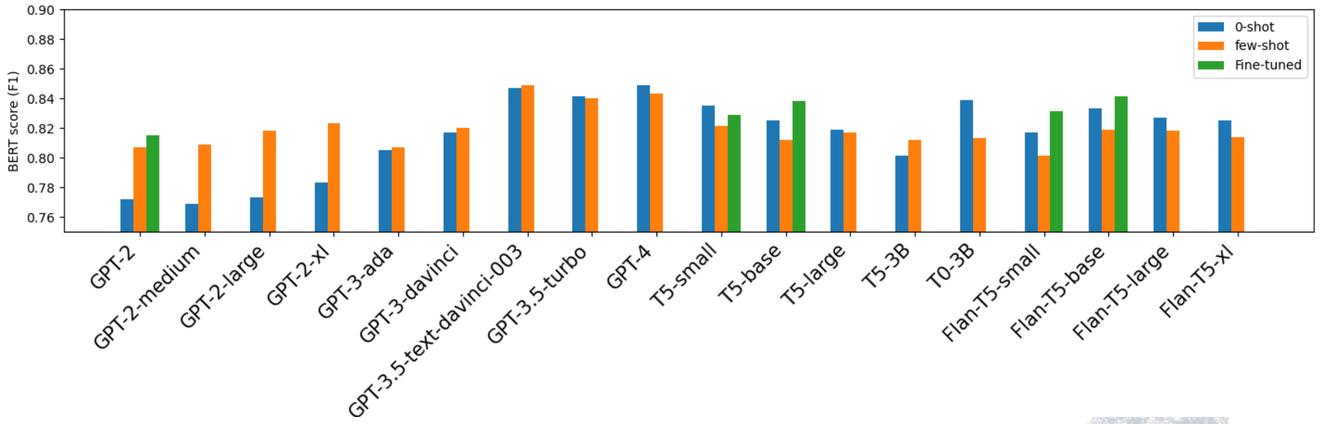
Rudra Dhar
Software Engineering Research Centre
IIIT Hyderabad, India
rudra.dhar@research.iiit.ac.in

Karthik Vaidhyanathan
Software Engineering Research Centre
IIIT Hyderabad, India
karthik.vaidhyanathan@iiit.ac.in

Vasudeva Varma
Language Technologies Research Centre
IIIT Hyderabad, India
vv@iiit.ac.in

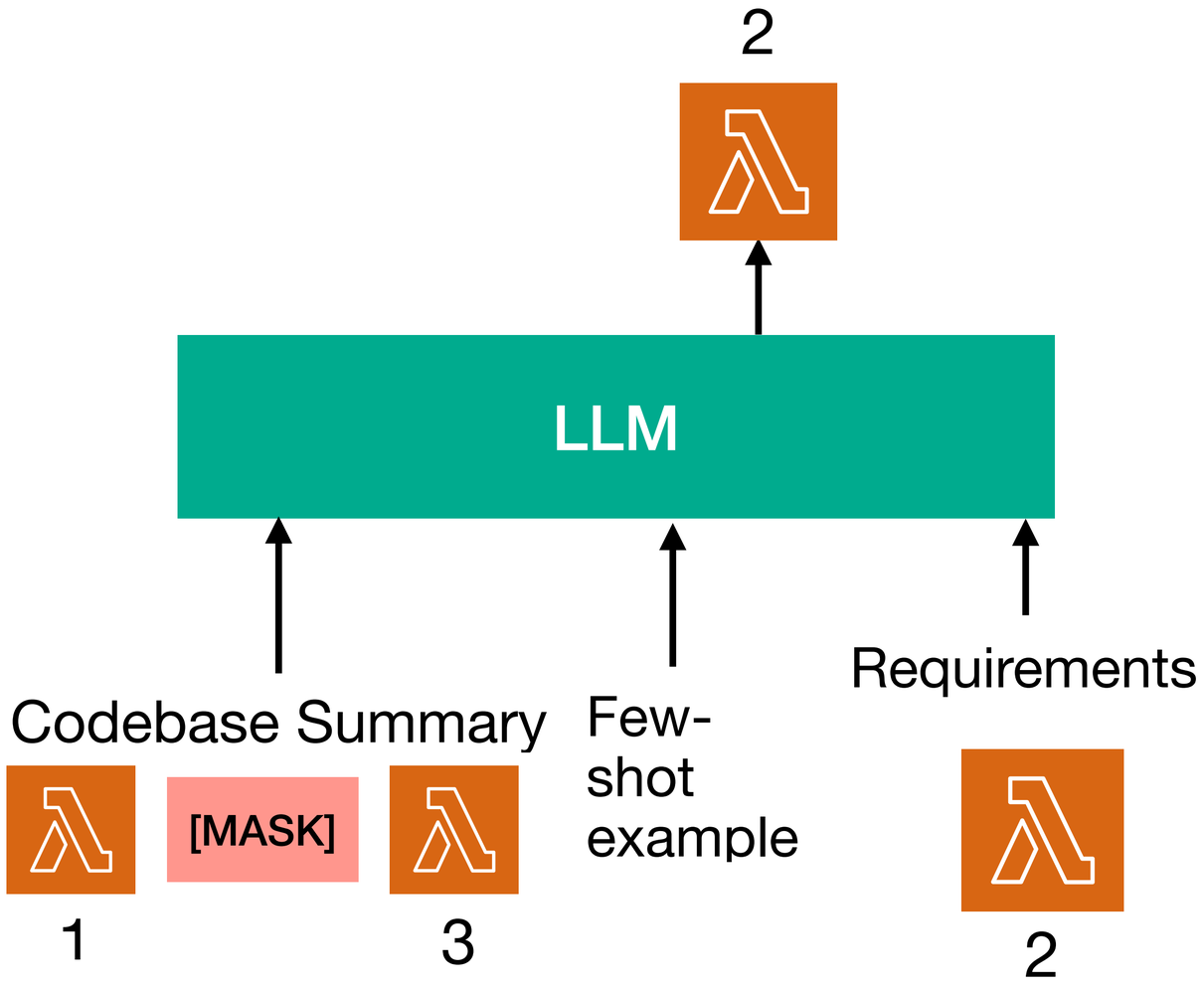
***Abstract*—Architectural Knowledge Management (AKM) involves the organized handling of information related to architectural decisions and design within a project or organization. An essential artefact of AKM is the Architecture Decision Records (ADR), which documents key design decisions. ADRs are documents that capture decision context, decision made and various aspects related to a design decision, thereby promoting** been a crucial reason restricting a wider adoption of AKM approaches, and more research is needed for automatically capturing this knowledge [3].

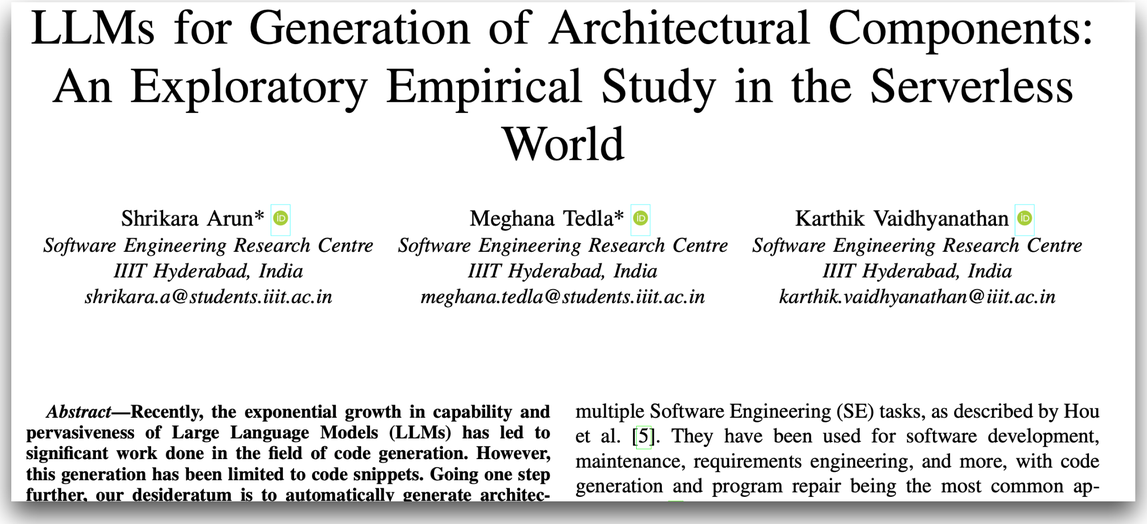
An *Architecture Decision Record (ADR)* is a crucial part of AKM. It entails the idea that software architecture is considered a set of Design Decisions [4]. It is a document used



Study with 18 LLMs - Small models performs well when fine-tuned

From Lab: AI4SE - LLMs for Component Generation





LLMs for Generation of Architectural Components: An Exploratory Empirical Study in the Serverless World

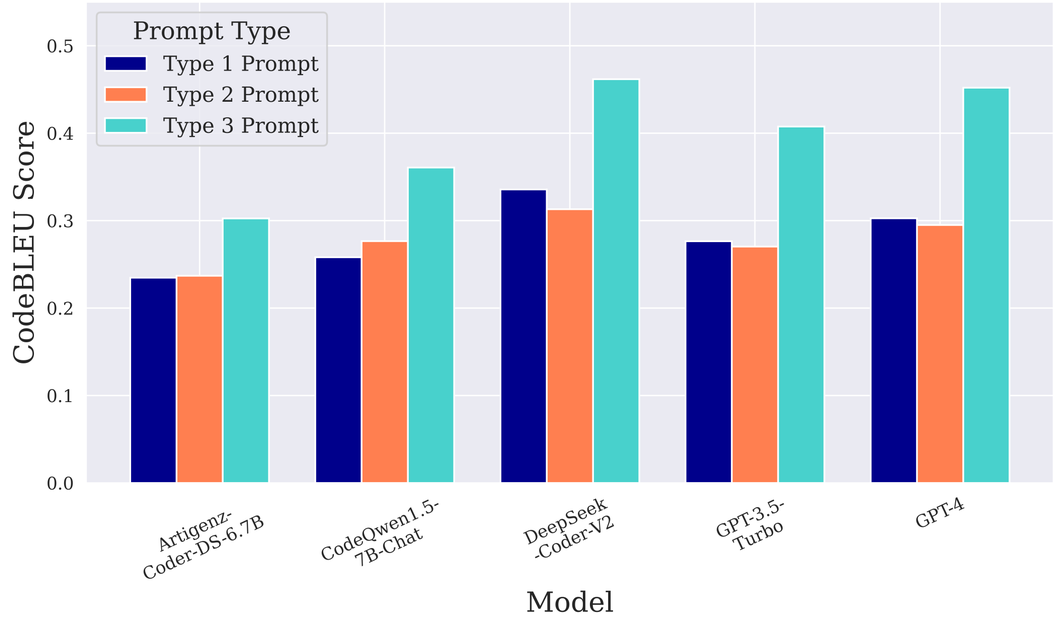
Shrikara Arun*
Software Engineering Research Centre
IIIT Hyderabad, India
shrikara.a@students.iiit.ac.in

Meghana Tedla*
Software Engineering Research Centre
IIIT Hyderabad, India
meghana.tedla@students.iiit.ac.in

Karthik Vaidhyanathan
Software Engineering Research Centre
IIIT Hyderabad, India
karthik.vaidhyanathan@iiit.ac.in

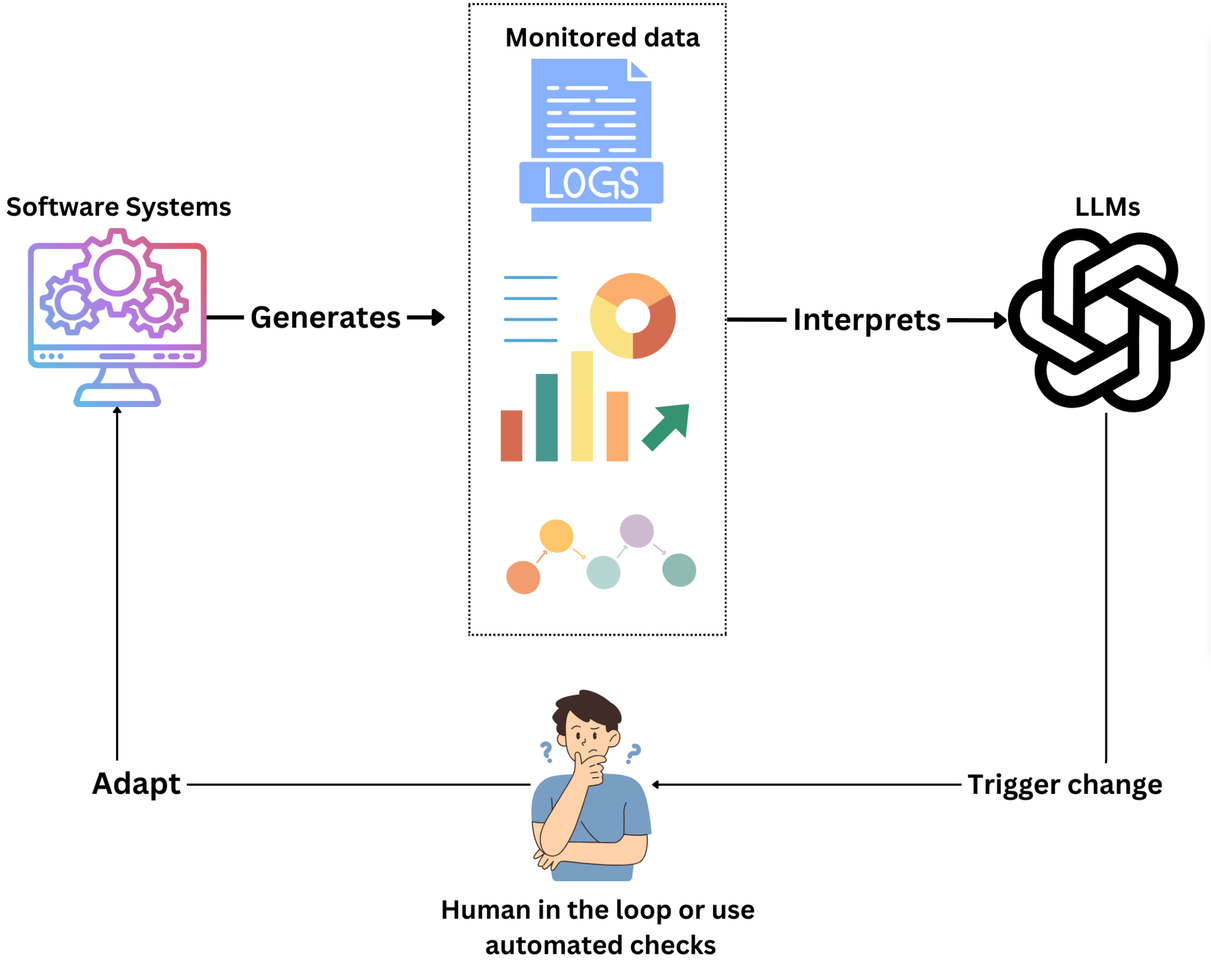
***Abstract*—Recently, the exponential growth in capability and pervasiveness of Large Language Models (LLMs) has led to significant work done in the field of code generation. However, this generation has been limited to code snippets. Going one step further, our desideratum is to automatically generate architec-**

multiple Software Engineering (SE) tasks, as described by Hou et al. [5]. They have been used for software development, maintenance, requirements engineering, and more, with code generation and program repair being the most common ap-



Study with 5 LLMs on 4 repos with 100+ functions - Human Architects + devs => Great Combination

From Lab: AI4SE - Self-adaptation using LLMs



Reimagining Self-Adaptation in the Age of Large Language Models

Raghav Donakanti, Prakhar Jain, Shubham Kulkarni, Karthik Vaidhyanathan
Software Engineering Research Center, IIIT Hyderabad, India
raghav.donakanti@students.iiit.ac.in, prakhar.jain@research.iiit.ac.in, shubham.kulkarni@research.iiit.ac.in,
karthik.vaidhyanathan@iiit.ac.in

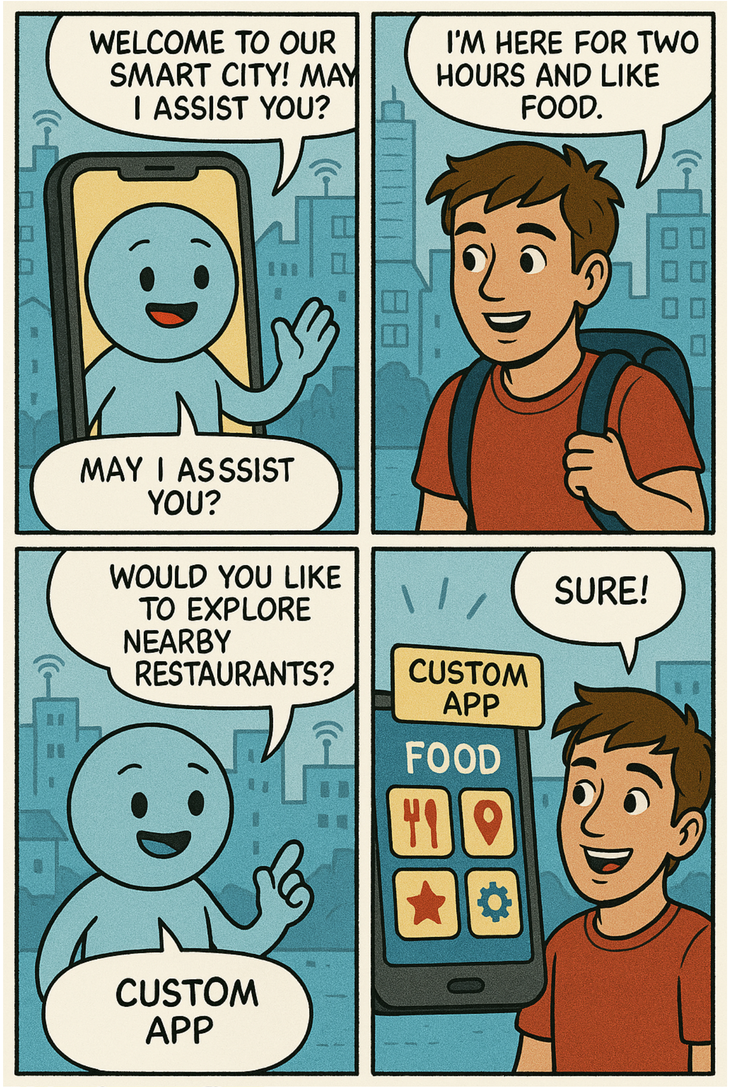
***Abstract*—Modern software systems are subjected to various types of uncertainties arising from context, environment, etc. To this end, self-adaptation techniques have been sought out as potential solutions. Although recent advances in self-adaptation through the use of ML techniques have demonstrated promising results, the capabilities are limited by constraints imposed by the ML techniques, such as the need for training samples, the ability to generalize, etc. Recent advancements in Generative AI (GenAI) open up new possibilities as it is trained on massive amounts of data, potentially enabling the interpretation of uncertainties and synthesis of adaptation strategies. In this context, this paper presents a vision for using GenAI, particularly Large Language Models (LLMs), to enhance the effectiveness and**

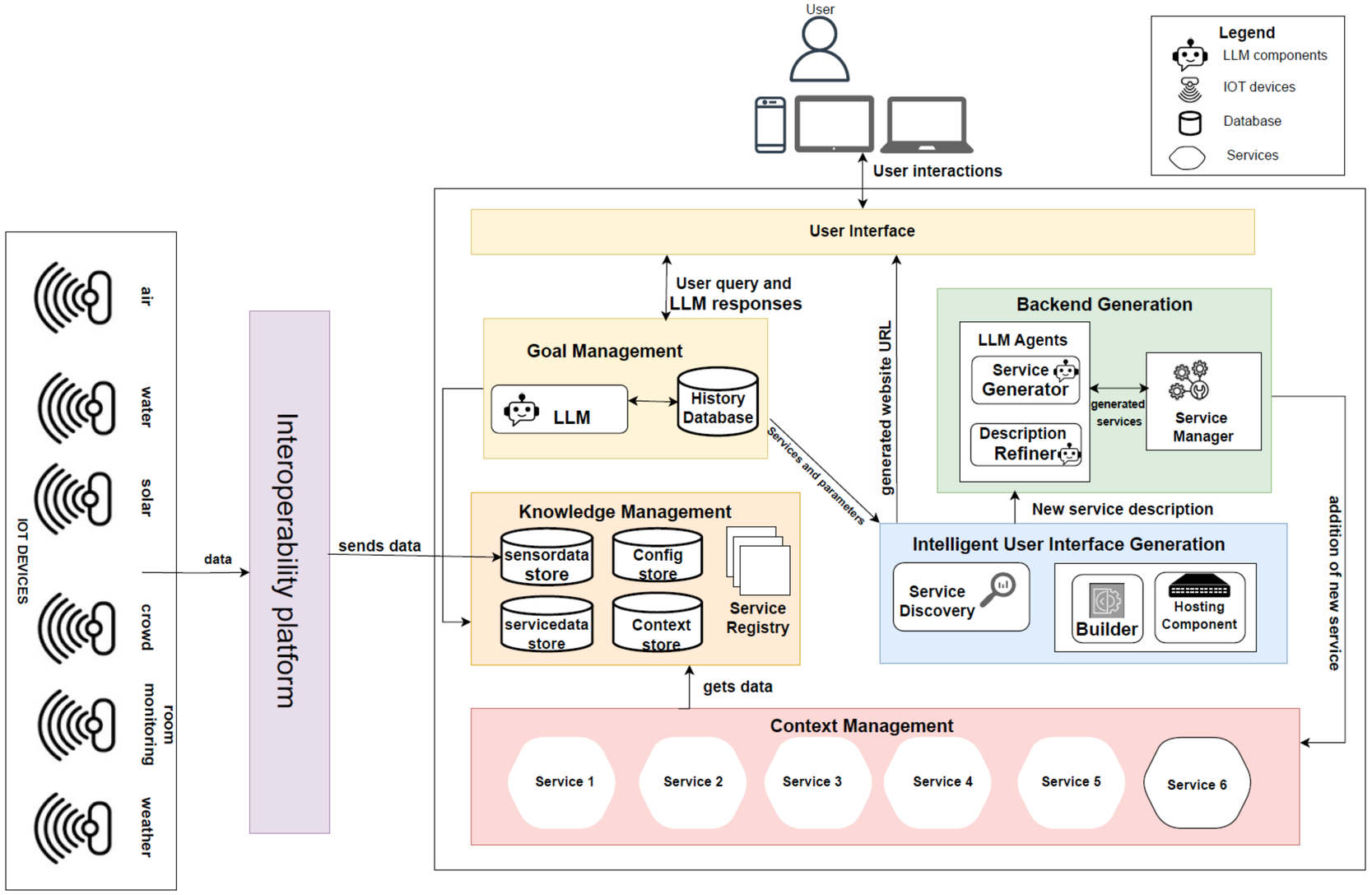
The concept of autonomic computing, as proposed by Kephart and Chess [5], sought to enhance the autonomy of software systems through various strategies. Despite these efforts, a persistent challenge has been the ability of systems to dynamically generate new configurations and components. The advent of GenAI, particularly the capabilities of LLMs, introduces the possibility of developing adaptation strategies directly. This is supplemented by the fact that modern software systems generate vast amounts of data, including logs, metrics, and traces, which system administrators traditionally leverage for tasks such as root cause analysis and resource allocation.

GPT-4 could ensure the system guarantees SLA almost as good as the state-of-the art

Autonomous adaptation with LLMs a possibility!!!

From Lab: AI4SE - Multi-agent for dynamic system generation





Adnan, B., Miryala, S., Sambu, A., Vaidhyanathan, K., De Sanctis, M. and Spalazzese, R.,**Leveraging LLMs for Dynamic IoT Systems Generation through Mixed-Initiative Interaction**. IEEE ICSA 2025 Companion

Our Efforts in GenAI4SE

Generation of user interface mockups [UI]

Checking requirements ambiguity and semantics

Diagrams from Descriptions [Collaboration]

In-house LLM for generating design decisions

Migration from monolith to Microservices

Study of automated code refactoring

Requirements

Design

Development

Study of code smells

Automated generation of serverless functions

Efficiency of generated code

Testing

Generating test cases from requirements

Source identification of a bug (Root Cause Analysis)

Deployment

On-device SLMs for function calling [Study and development] - Collaboration

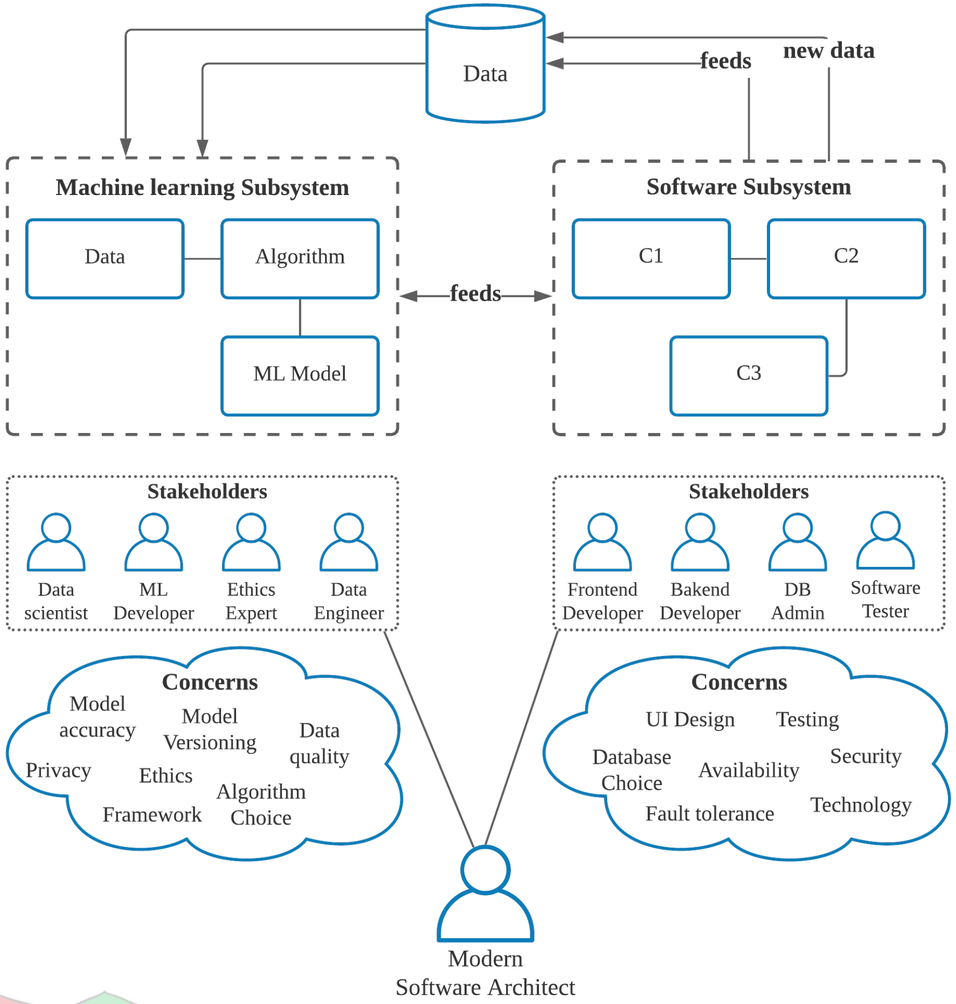
Generating IaaC [Collaboration]

Autonomous CloudOps

Maintenance

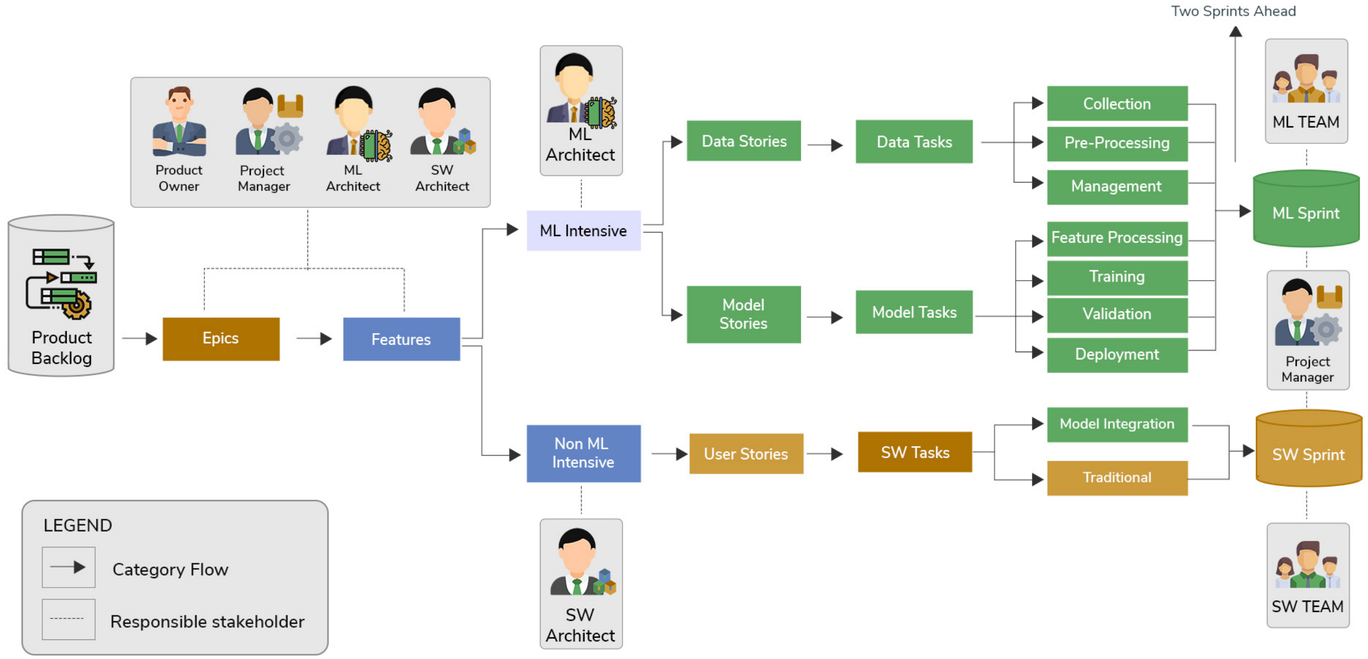
LLLMs for autonomous self-adaptation (multi-agent)

SE4AI: Calls for a Paradigm Shift (Agentic AI just adds to it)



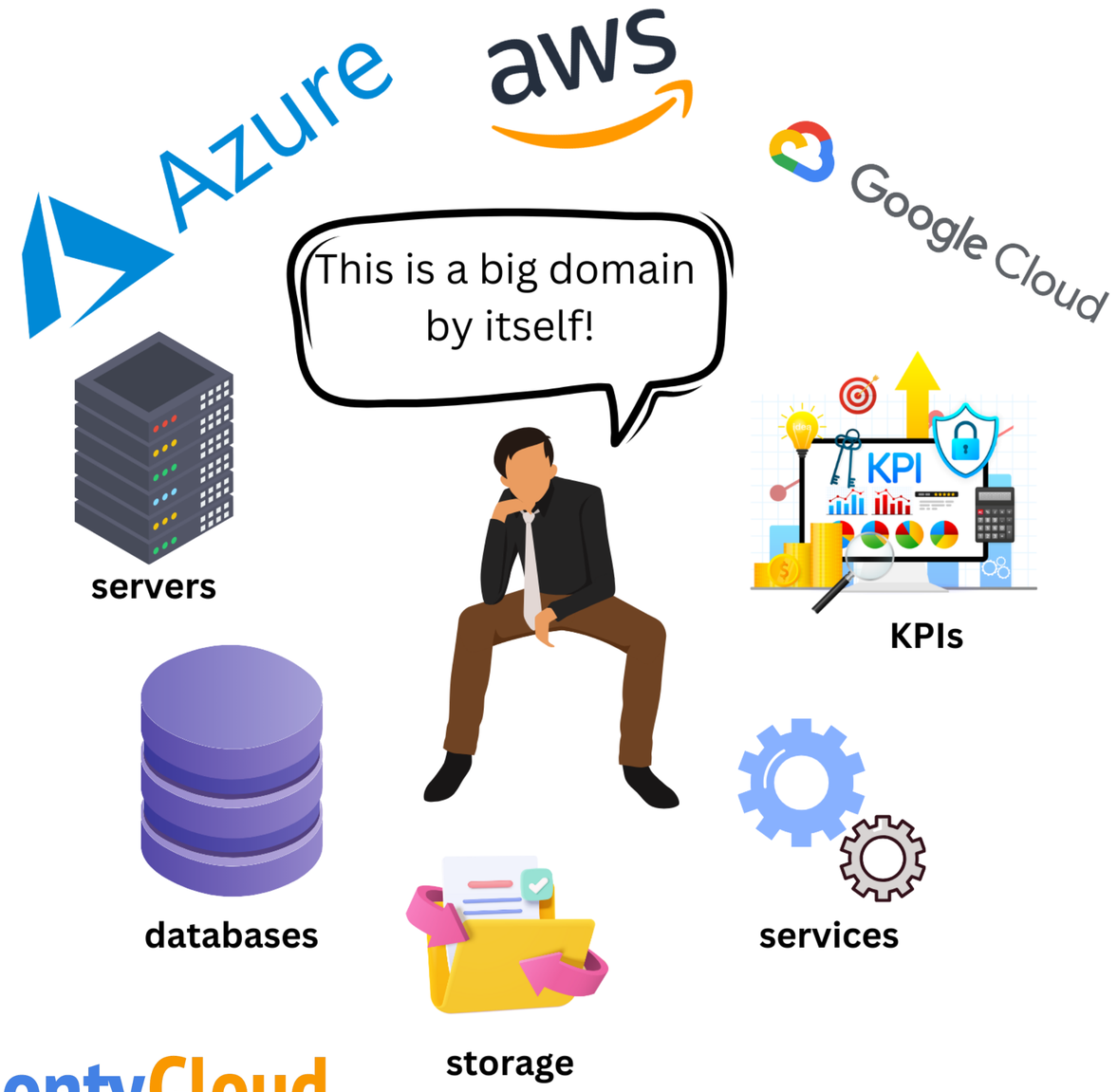
H. Muccini, K. Vaidhyanathan: **Software Architecture for ML-based Systems: What Exists and What Lies Ahead.** WAIN@ICSE 2021

> 50% of ML systems do not make it into production - - Gartner



K. Vaidhyanathan, A. Chandran, H. Muccini and R. Roy, **Agile4MLS—Leveraging Agile Practices for Developing Machine Learning-Enabled Systems: An Industrial Experience** in IEEE Software, 2022

To Land: SE4AI - Autonomous CloudOps



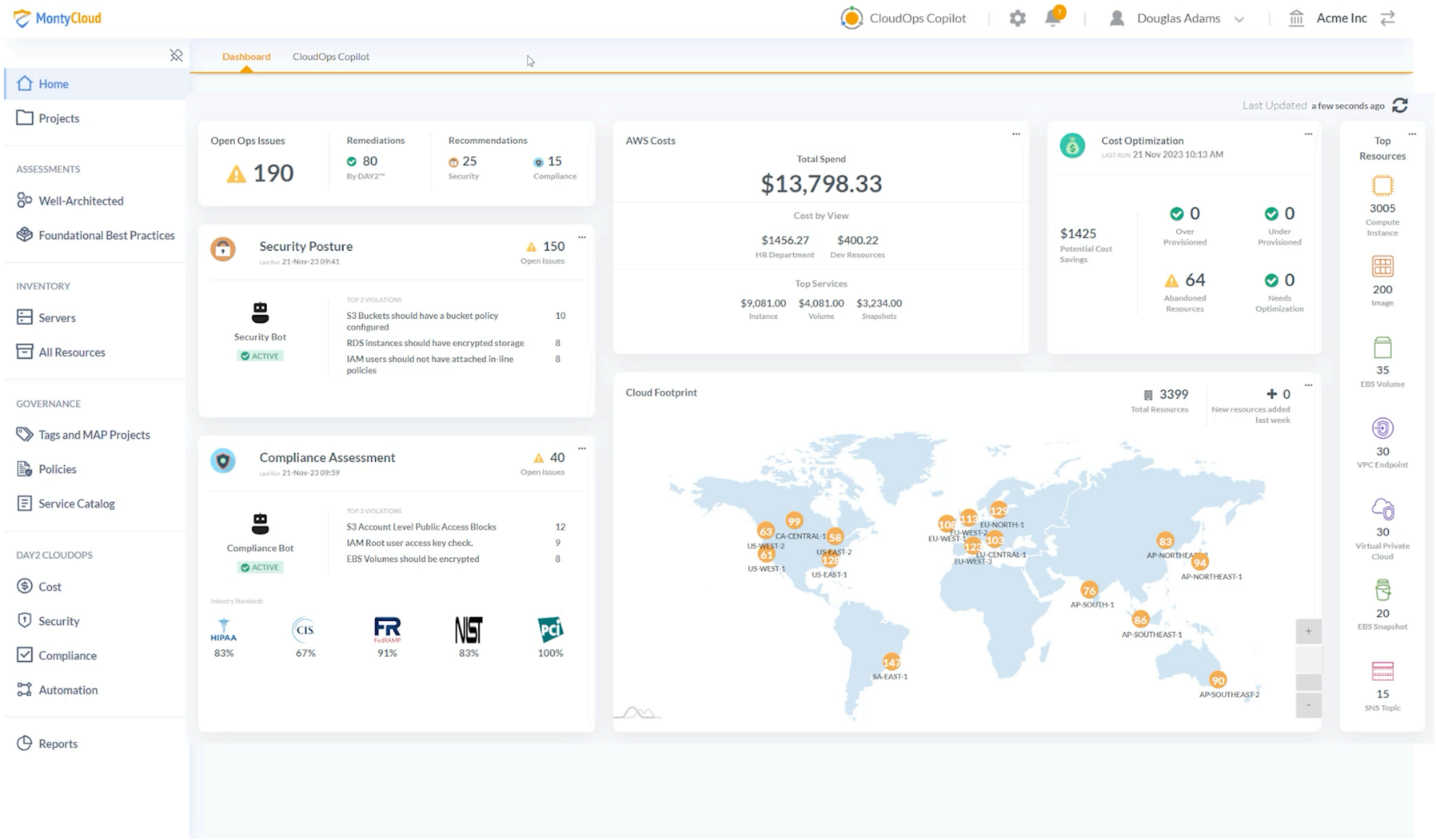
CloudOps - *Run, Manage, Evolve*

AWS Well Architected Framework

Helps cloud architects build resilient, secure and high performing infrastructure

- **Build around six pillars**
  - Operational Efficiency
  - Security
  - Reliability
  - Performance Efficiency
  - Sustainability
  - Cost

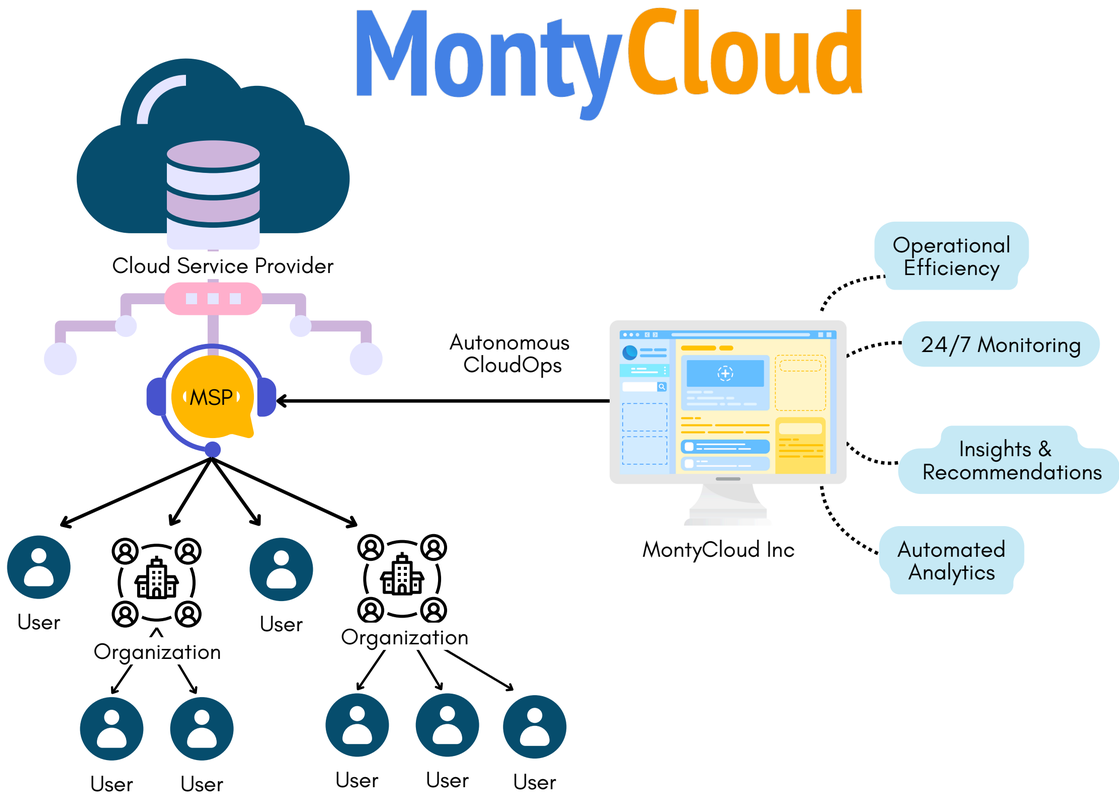
Ideas into Production: CloudOps CoPilot



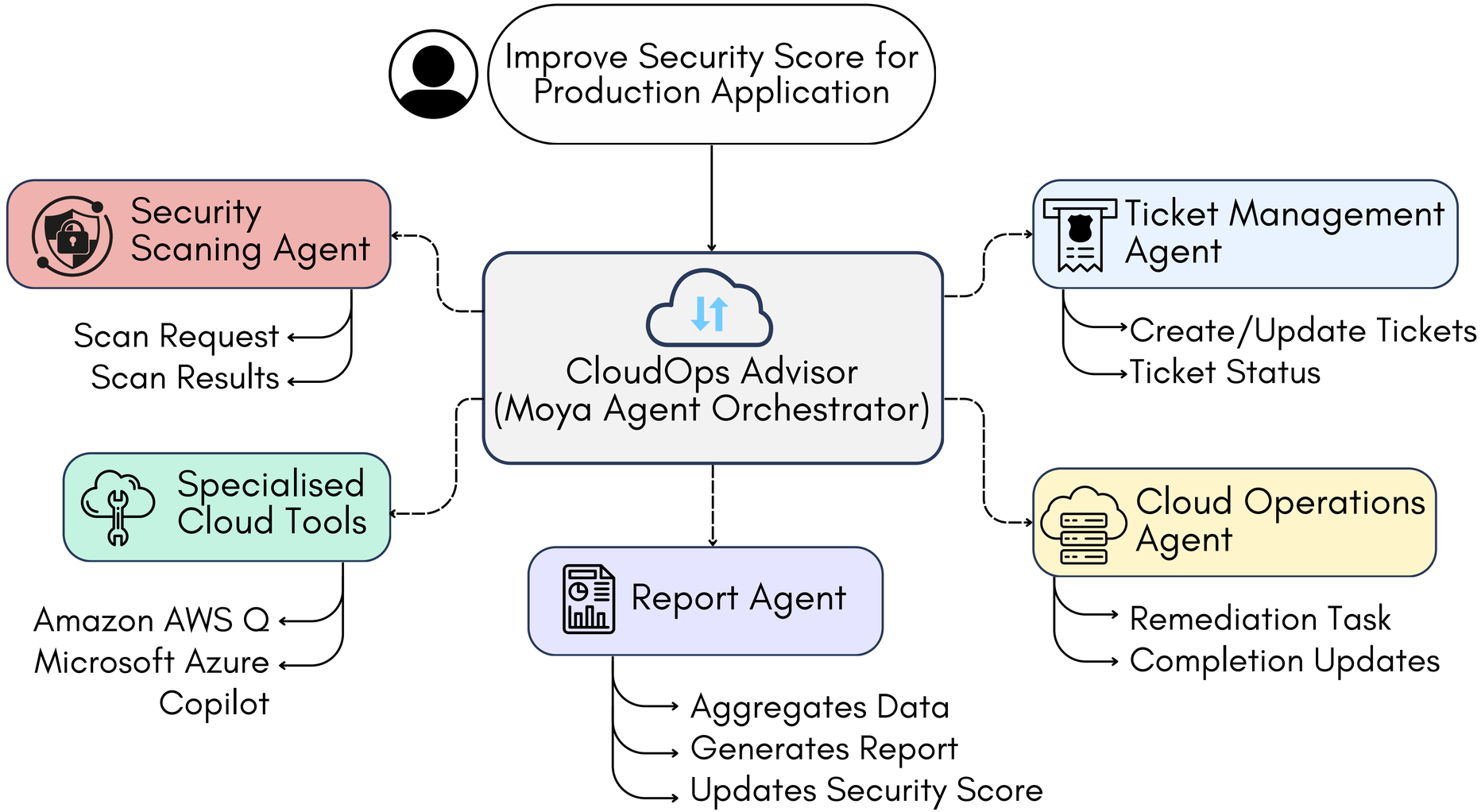
Work done in collaboration with MontyCloud Inc.

Complex Engineering Challenges

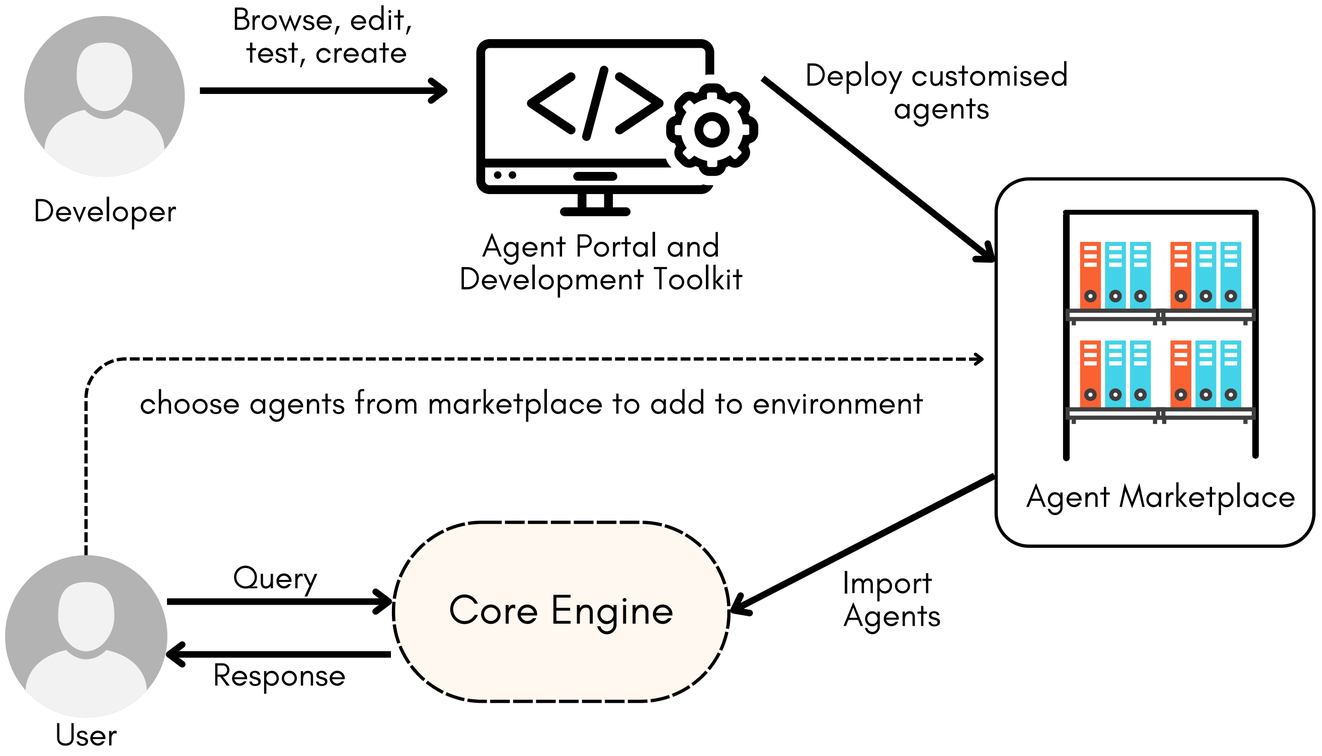
- **Managing Distributed Data**
  - Diverse data sources
- **Maintainability**
  - Large code base, time for updates
- **Extensibility and Modularity**
  - Single vendor, ease of extensions!
- **Monolithic nature of existing frameworks**
  - Limited support, vendor lock-in, learning curve



Can we go Multi-agent?



Enters MOYA: Meta Orchestration Framework of Your Agents



Meta orchestration Framework

Engineering LLM Powered Multi-agent Framework for Autonomous CloudOps

Kannan Parthasarathy*, Karthik Vaidhyanathan†, Rudra Dhar†, Venkat Krishnamachari*, Basil Muhammed*, Adyansh Kakran†, Sreemaee Akshathala†, Shrikara Arun†, Sumant Dubey*, Mohan Veerubhotla*, Amey Karan†
*MontyCloud Inc
†Software Engineering Research Center, IIIT Hyderabad, India
Email: karthik.vaidhyanathan@iiit.ac.in, {kannan, venkat, basil, sumant, mohan}@montycloud.com,
{rudra.dhar, adyansh.kakran, sreemaee.akshathala, amey.karan}@research.iiit.ac.in, shrikara.a@students.iiit.ac.in

***Abstract*—Cloud Operations (CloudOps) is a rapidly growing field focused on the automated management and optimization of cloud infrastructure which is essential for organizations navigating increasingly complex cloud environments. MontyCloud Inc is one of the major companies in the CloudOps domain that leverages autonomous bots to manage cloud compliance, security, and continuous operations. To make their platform more accessible and effective to the customers, MontyCloud worked with us to leverage the use of GenAI.**

Developing a GenAI-based solution for autonomous CloudOps for the existing MontyCloud system presented us with various challenges such as i) diverse data sources; ii) orchestration of multiple processes and iii) handling complex workflows to automate routine tasks. To this end, we developed MOYA, a multi-agent framework that leverages GenAI and balances autonomy with the necessary human control. This framework integrates various internal and external systems and is optimised for factors like task orchestration, security, and error mitigation while producing accurate, reliable and relevant insights by utilising Retrieval Augmented Generation (RAG). Evaluations of our multi-agent system with the help of practitioners as well as using automated checks demonstrate enhanced accuracy, responsiveness, and effectiveness over non-agentic approaches across complex workflows.

a well-defined CloudOps practice to effectively manage their share of duties.

CloudOps, or Cloud Operations, refers to the practices, tools, and processes to manage, optimise, and secure applications and infrastructure in the cloud. Alonso et al. [1] defines it as a framework that extends *DevOps* practices to cloud management by adding components like resource discovery, self-healing, and real-time monitoring. By focusing on automation, monitoring, cost management, and compliance, CloudOps enables organisations to maintain efficient, resilient, and scalable cloud environments. However, the complex and dynamic nature of cloud services makes manual management time-intensive, challenging, and prone to errors.

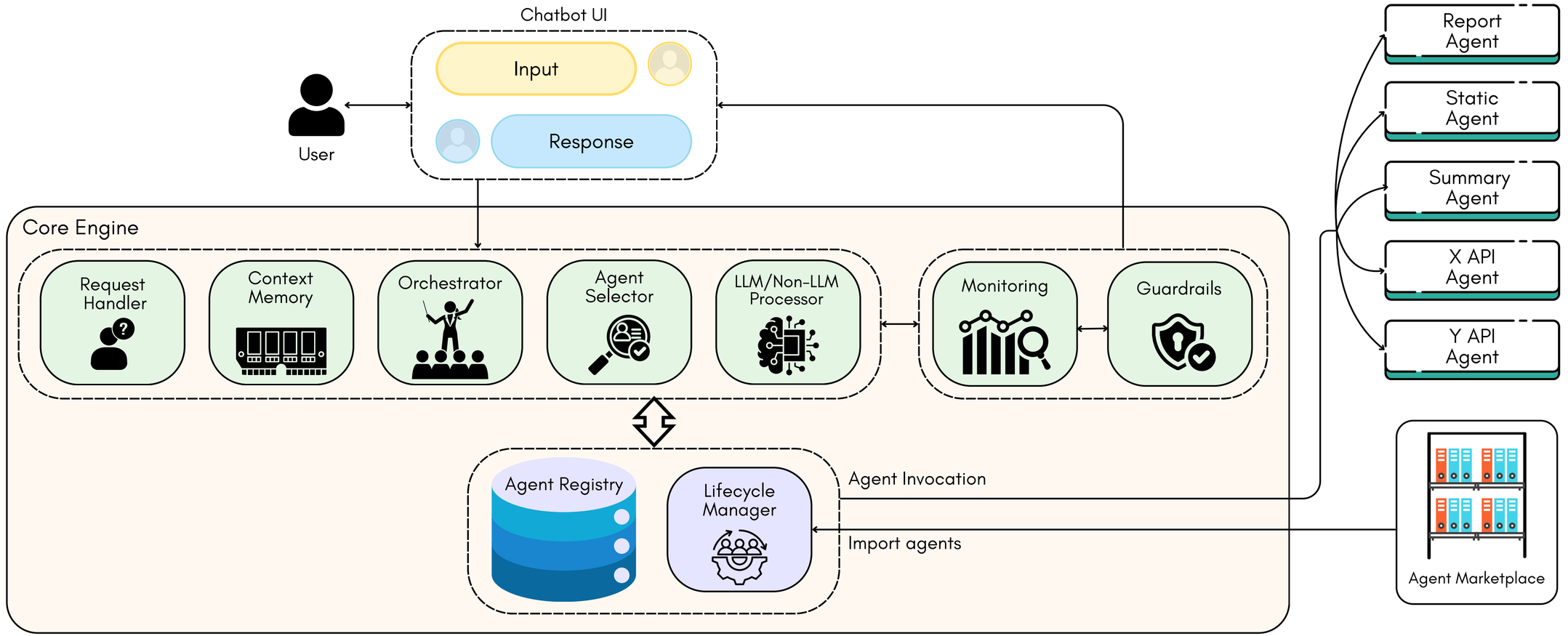
MontyCloud's Autonomous CloudOps platform addresses these challenges by automating workflows to streamline operations and provide real-time visibility into inventory, security, and costs [4]. The platform tackles challenges such as navigating the complexity of hundreds of services, establishing secure and cost-effective cloud governance, ensuring a strong security posture, and adhering to evolving compliance standards.

Best paper candidate@CAIN, ICSE 2025

Basic idea: support custom agents, in-house, etc. - Orchestrate with guardrails

The MOYA Multi-agent Framework

MOYA repo

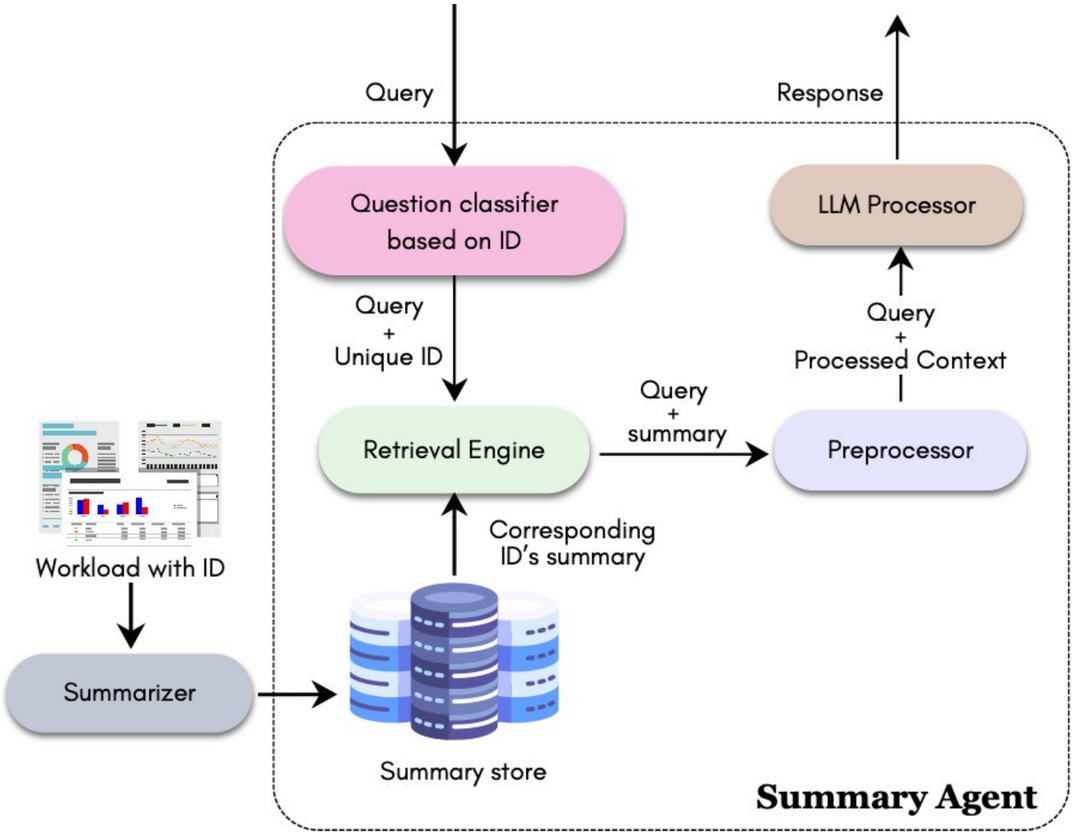


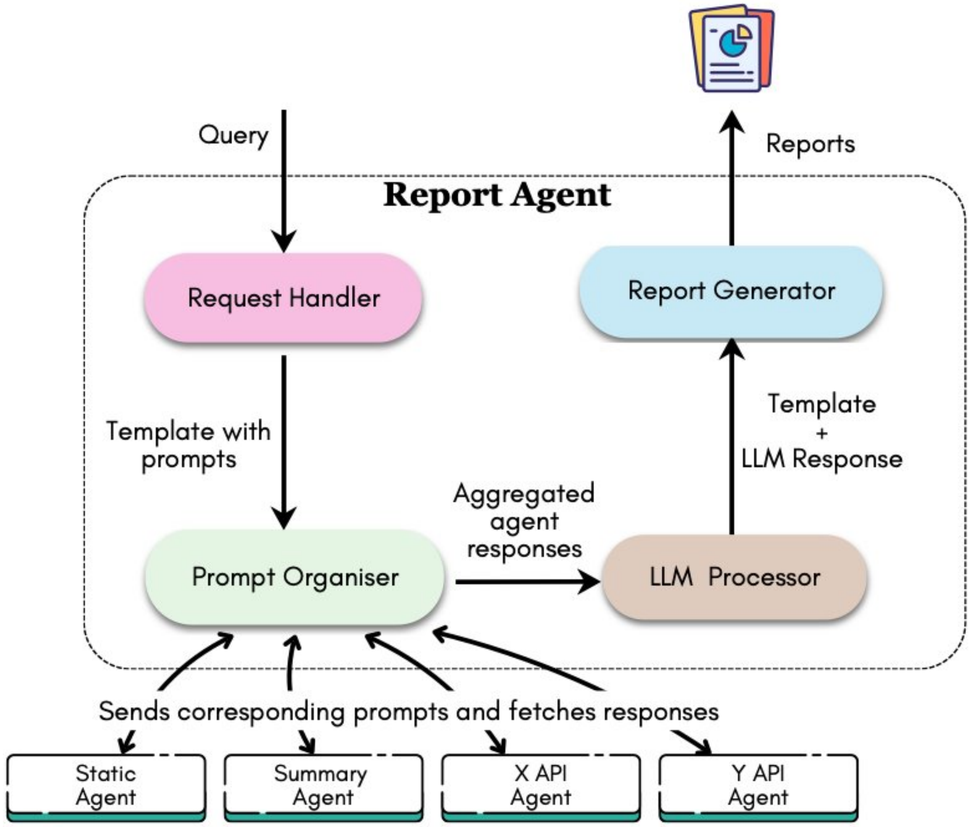
Parthasarathy, K., Vaidhyanathan, K., Dhar, R., Krishnamachari, V., Muhammed, B., Kakran, A., Akshathala, S., Arun, S., Dubey, S., Veerubhotla, M. and Karan, A., 2025. **Engineering LLM Powered Multi-agent Framework for Autonomous CloudOps**. *arXiv preprint arXiv:2501.08243,* CAIN 2025.

Some Agents in MOYA

Following Principles of Domain Driven Design

MOYA repo





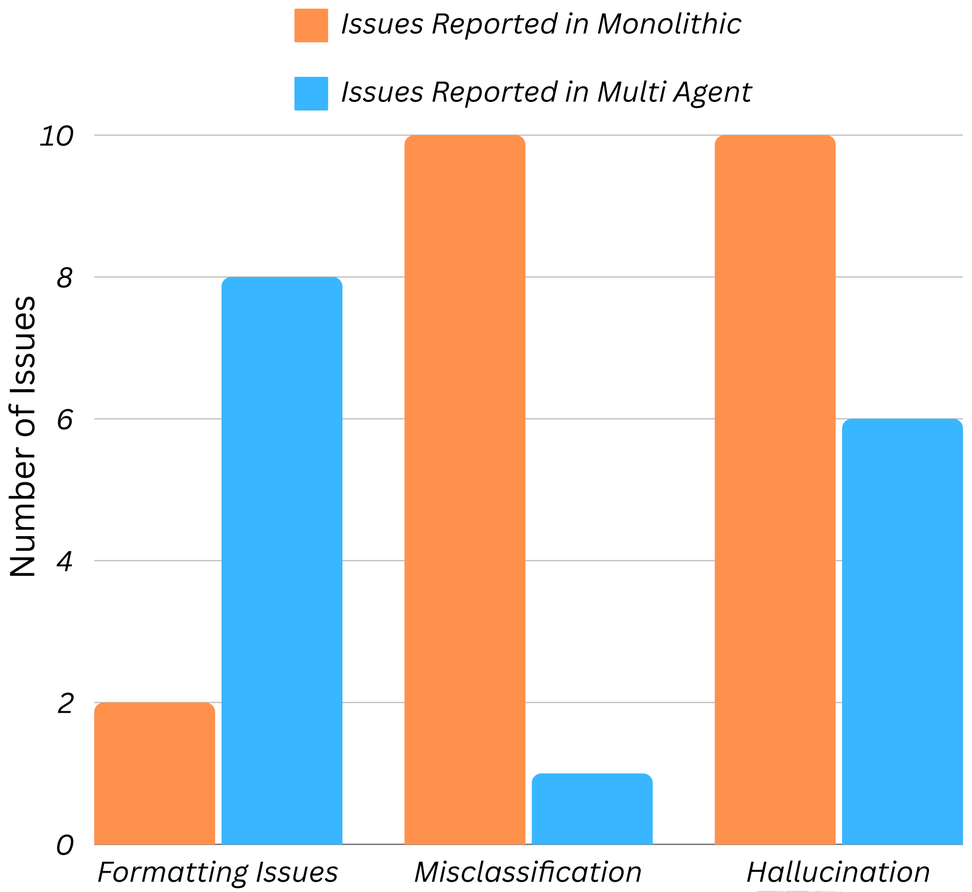
MOYA has support for different memory; MCP and A2A compliance on the way

Evaluating MOYA

MOYA repo

- Combination of automated and manual evaluations
- Ground truth of 260 prompts and responses
  - Curated with support of domain experts and LLMs

Approach	Rouge-1	bleu	Meteor	BERT score		
				Precision	Recall	f1
Monolith	0.321	0.102	0.265	0.854	0.834	0.843
MOYA	**0.448**	**0.221**	**0.423**	**0.867**	**0.869**	**0.868**



MOYA preformed much better -> Integration to the product

MOYA in action

MOYA Hackathon@IIITH

- 20+ teams with about 100 students
- 16 use cases across different domains
  - Framework extensions
  - Open track
- Some outputs/feedbacks
  - Generalisability of MOYA
  - Ease of use
  - Suggestions for improvement

Source: https://blogs.iiit.ac.in/moya/

MontyCloud and IIIT Hyderabad Present Groundbreaking Framework for Autonomous Agent Orchestration at CAIN

NEWS PROVIDED BY
MontyCloud, Inc.
January 16, 2025, 13:00 GMT

SHARE THIS ARTICLE

Joint Industry-Academia Research Unveils Novel Framework on Multi-Agents

March 11, 2025 | Sarita Chebbi

Meme Generator

Meta Solver

More..

Mental Wellness Assistant

Team Orchestrator

Start Thinking in Agents

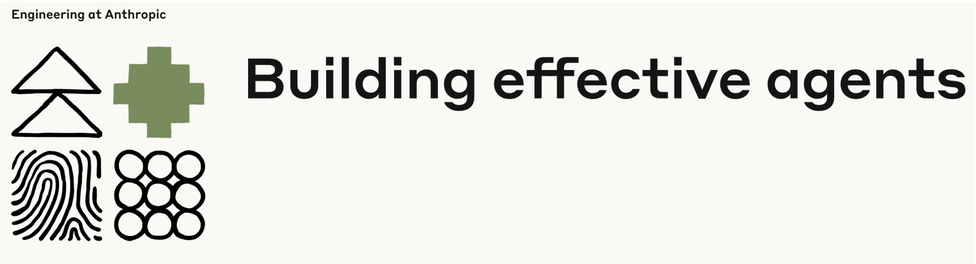
Build them in the right way

- There are emerging patterns
- Not every time we need to build agents
  - **Simple chatbots:** LLMs with RAG
  - **Workflows:** Orchestrated flows where LLM calls a tool
  - **Agents:** Back and forth communication to accomplish a task - Dynamic nature
- **Engineering plays the key:** DDD, Separation of Concerns, Trade-offs..

AGENT DESIGN PATTERN CATALOGUE:
A COLLECTION OF ARCHITECTURAL PATTERNS
FOR FOUNDATION MODEL BASED AGENTS

Yue Liu, Sin Kit Lo, Qinghua Lu, Liming Zhu, Dehai Zhao, Xiwei Xu, Stefan Harrer, Jon Whittle
Data61, CSIRO, Australia
Email: *firstname.lastname*@data61.csiro.au

November 7, 2024



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

Key Takeaways

Agentic AI is shifting the way we think about building software/software services

- We **need a change in mindset** when it comes to development
- Lot of **support for automation** (eg: modernization)
- **Reliability, Robustness, Responsibility** - Engineering is the key!
- **Domain specific LLMs** which are smaller shall be the way forward - collection of SLMs (helps agents)
- **Need for better processes** to architect/engineer systems around AI agents
- **Agentic thinking -** SaaS as such is not dead but the **way we build/develop!**
- **AgenticAI - Reimagine Autonomy, Sustainability and intelligence at scale!**

Credits: Kannan Parthasarathy, MontyCloud

SA4S@SERC


Rudra Dhar


Akhila Matathammal


Hiya Bhatt

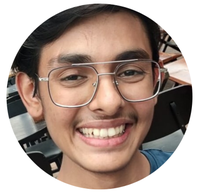

Chandrasekar S

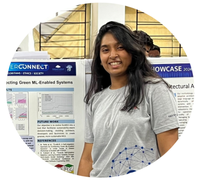

Shubham Kulkarni


Adyansh Kakran


Prakhar Jain


Shrikara A


Arya Pravin Marda


Meghana Tedla


Miryala Sathvika


Prakhar Singhal


Amey Karan


Bassam Adnan


Aneesh Sambu


Shaunak Biswas


Shailender Goyal

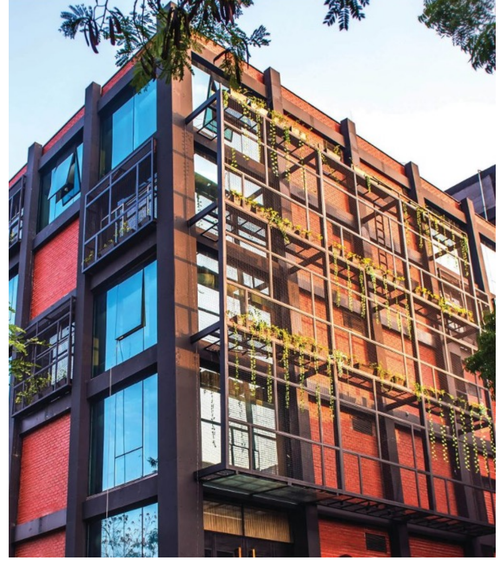

Sreemaee Akshathala


Divyansh Pandey


Maddireddy Kritin


Santosh Kotekal

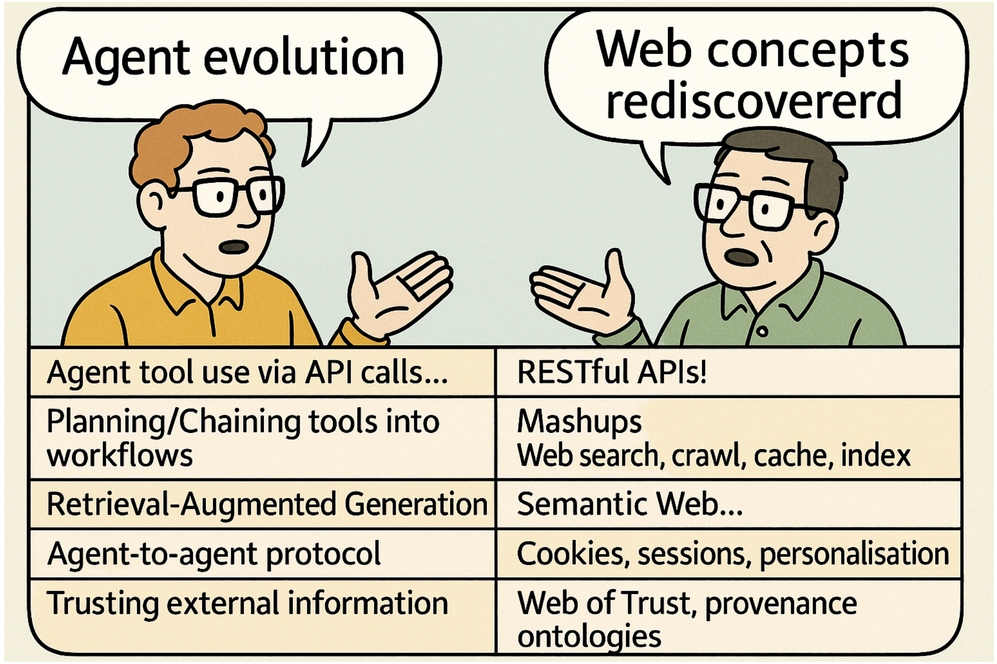

Vyakhya Gupta



https://serc.iiit.ac.in

Team SA4S

https://sa4s-serc.github.io



Concept credit: Liming Zhu

SCAN ME

Thank you

IEEE Software Magazine

Web: karthikvaidhyanathan.com
Email: karthik.vaidhyanathan@iiit.ac.in
Twitter: @karthi_ishere

Software Engineering Research Centre

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD