Playing with Abstractions: At the Crossroads of Software Architecture and Generative AI

Karthik Vaidhyanathan

Workshop on Generative AI-based Software Engineering, 2025

Feb 20, 2025

SERC
Software Engineering Research Centre

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

ISEC
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

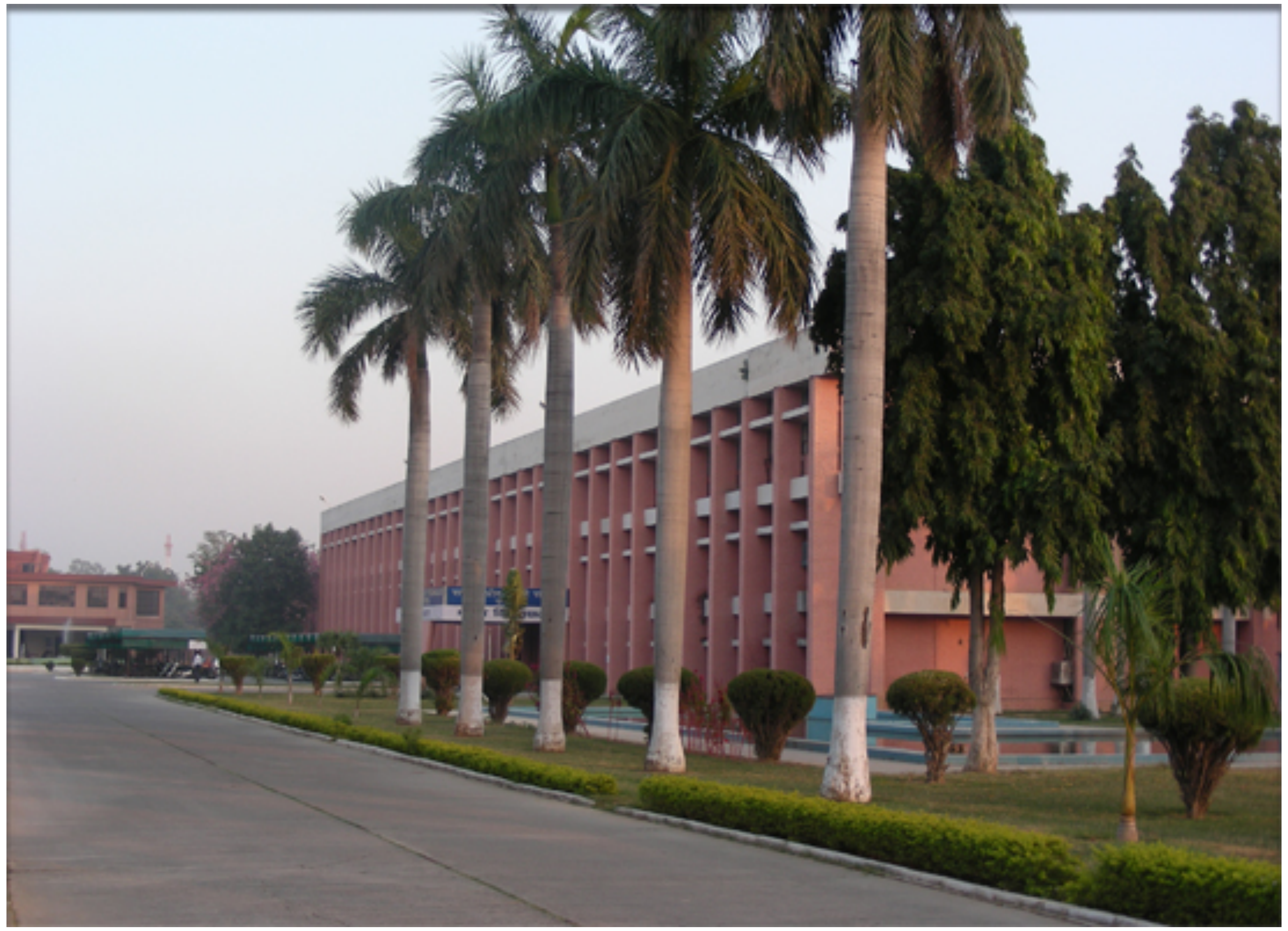
Double Master Degree - Software Architecture and Machine Learning
PhD from GSSI, Italy
Postdoc, University of L'Aquila, Italy

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

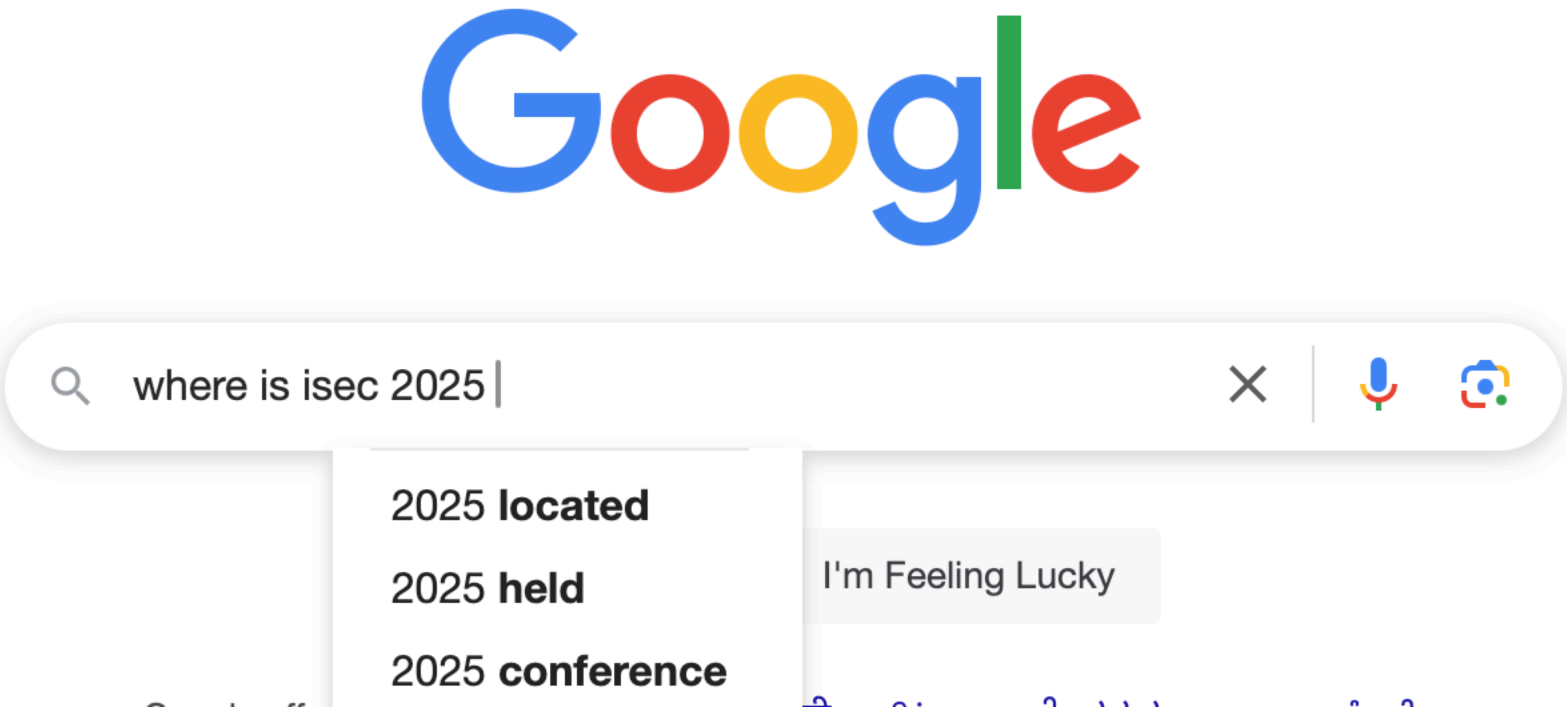
Text: An Abstraction of Reality!



The image depicts a red-brick institutional building lined with tall palm trees on a well-maintained campus.

A symmetrical building with a colonnaded facade stretches horizontally across the frame, flanked by palm trees and lush foliage under a slightly overcast sky.

Next word prediction problem



where is isec 2025

2025 **located**

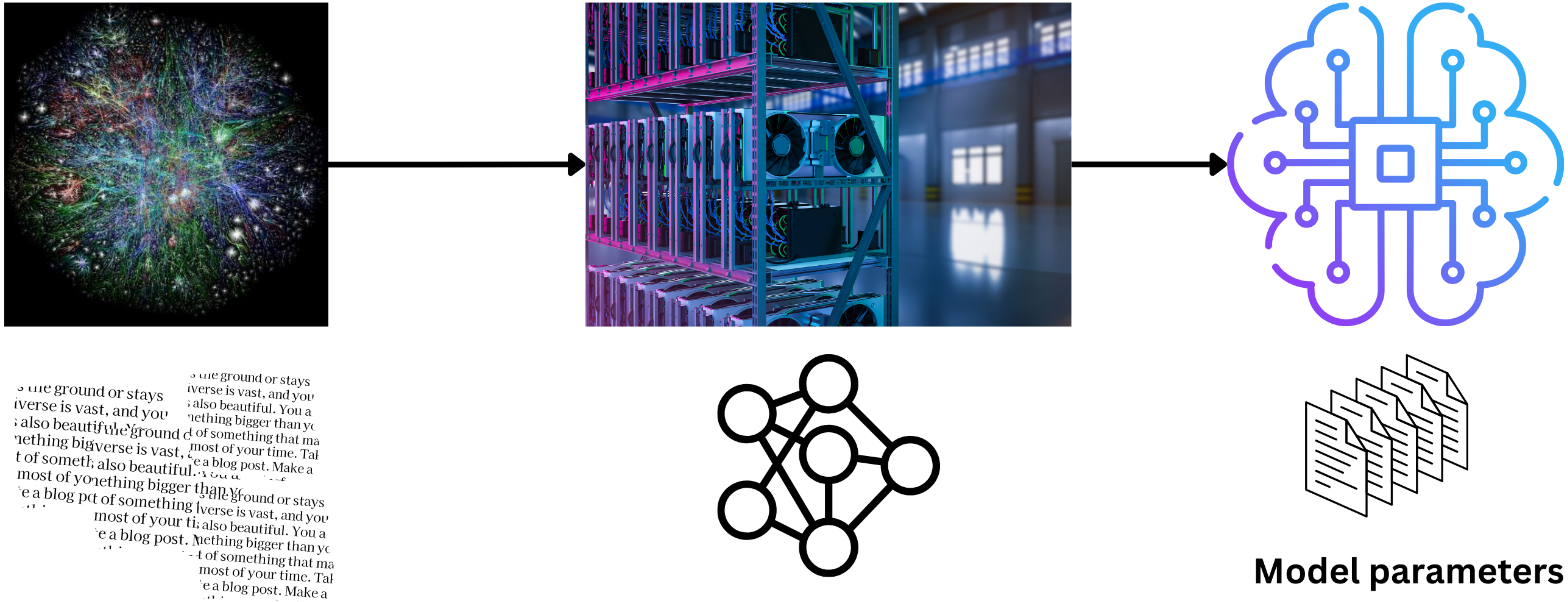
2025 **held**

2025 **conference**

Problem since 1950's!!!

“Large” Language Models (LLM)

Do you have a ton of text and compute power?



Internet

Compute

Foundation model/ Base model

~10B of text from internet

Transformer NN
trained on 1000s of GPUs
for days

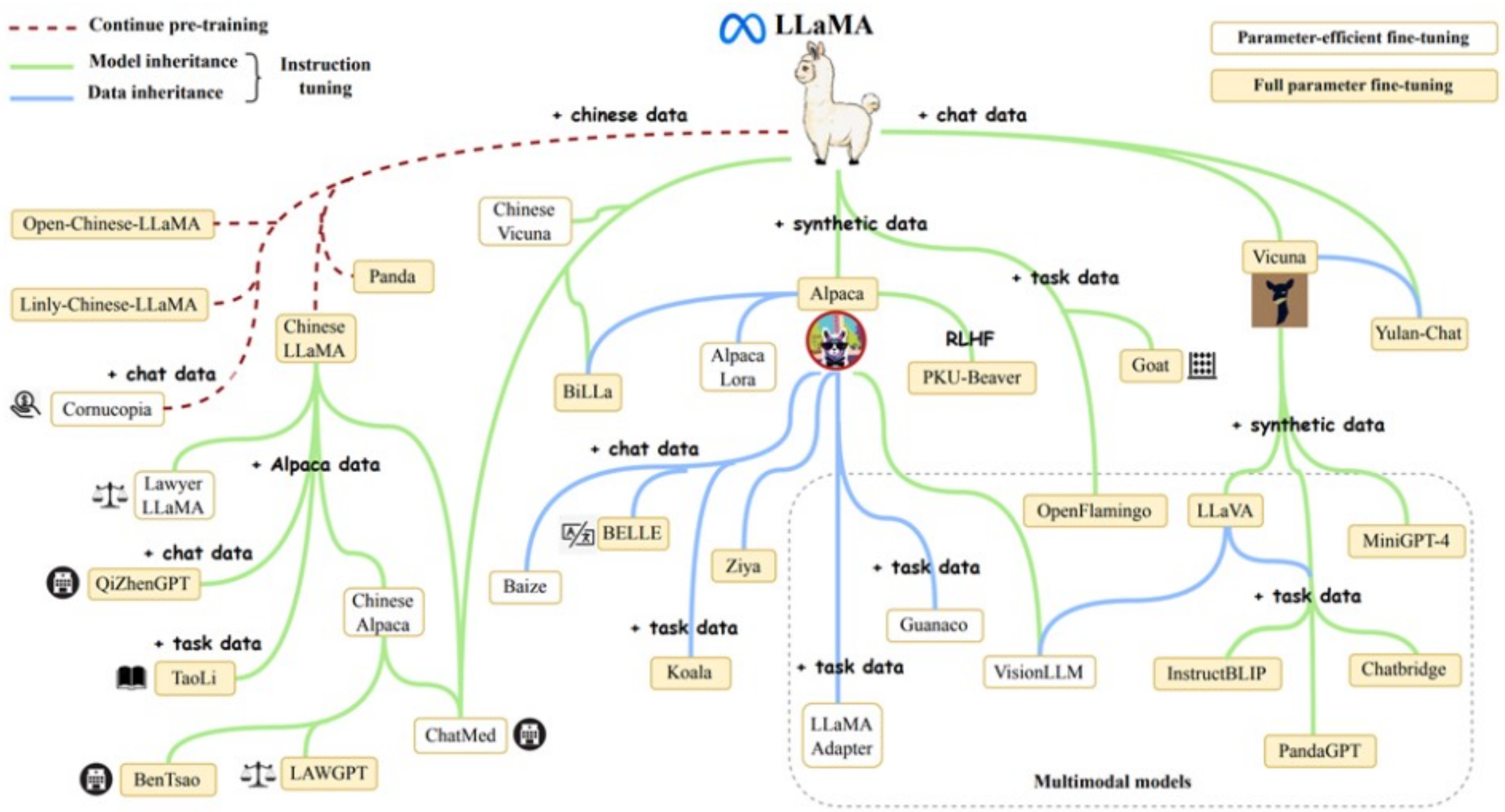
Model parameters
~some billions
100s of GB

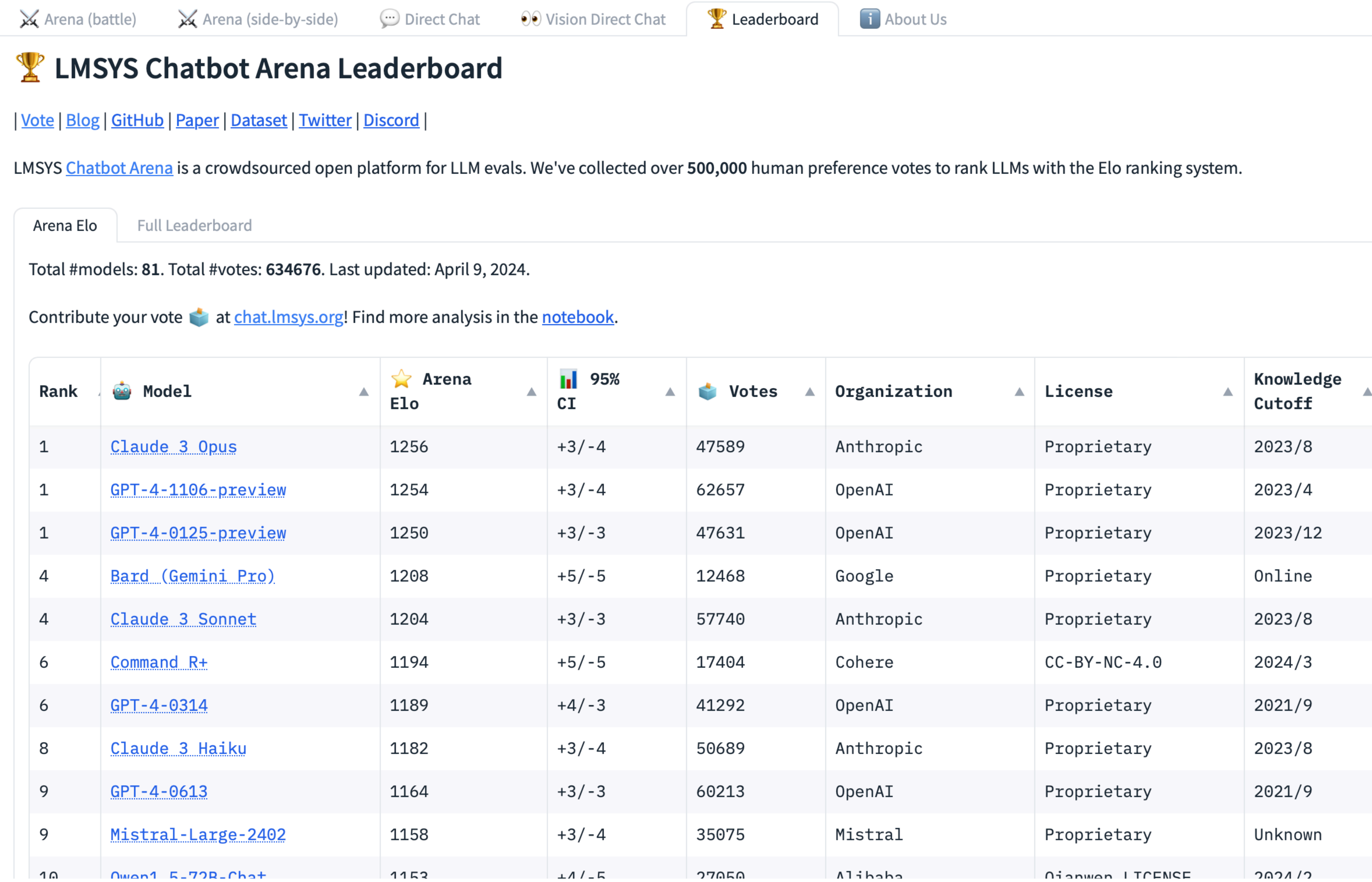
Inspiration from slides of Andrej Karpathy, Introduction to large language models

Today we have different varieties of LLMs

Proprietary

Open Source





🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **500,000** human preference votes to rank LLMs with the Elo ranking system.

Total #models: **81**. Total #votes: **634676**. Last updated: April 9, 2024.

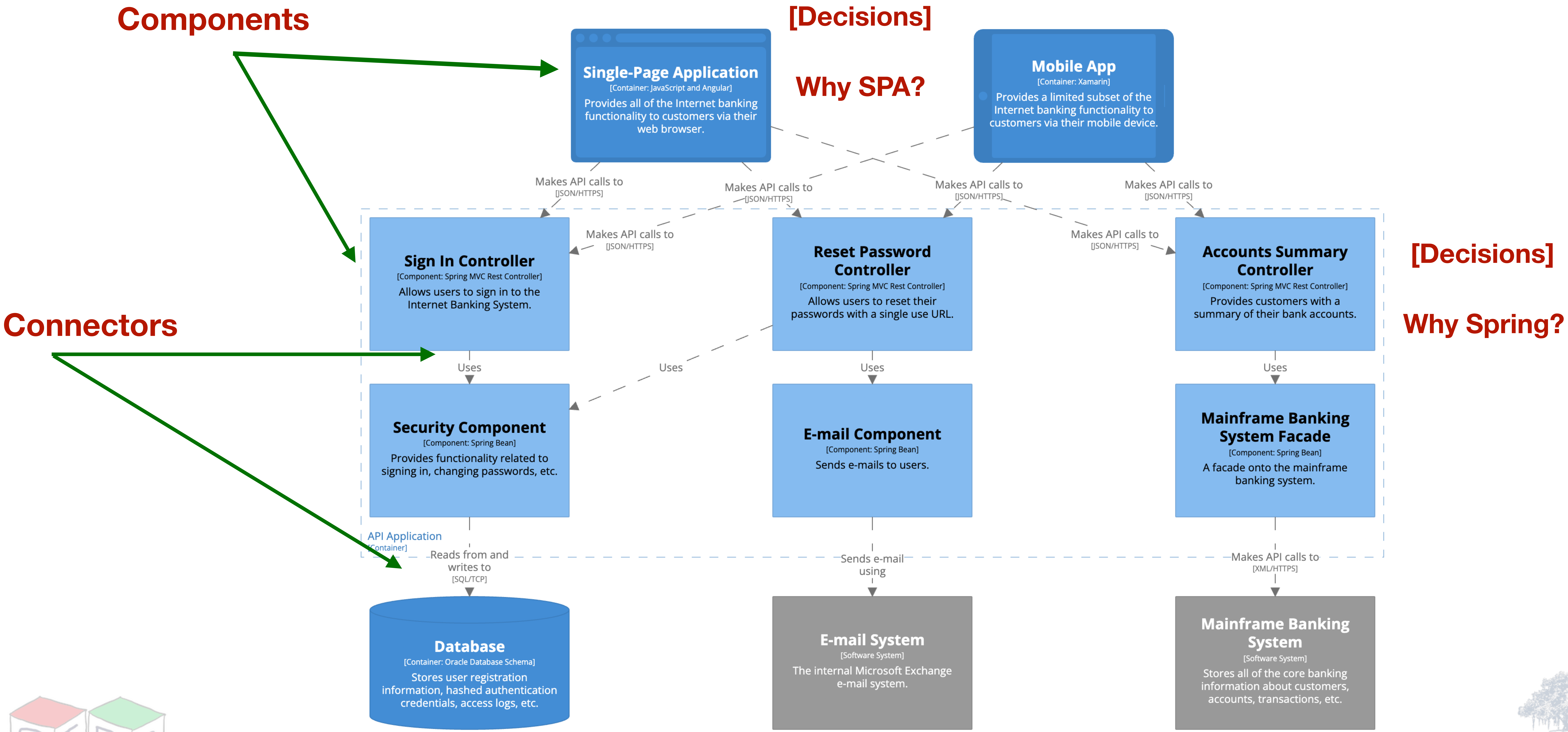
Contribute your vote 🗳️ at chat.lmsys.org! Find more analysis in the notebook.

Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
1	Claude 3 Opus	1256	+3/-4	47589	Anthropic	Proprietary	2023/8
1	GPT-4-1106-preview	1254	+3/-4	62657	OpenAI	Proprietary	2023/4
1	GPT-4-0125-preview	1250	+3/-3	47631	OpenAI	Proprietary	2023/12
4	Bard (Gemini Pro)	1208	+5/-5	12468	Google	Proprietary	Online
4	Claude 3 Sonnet	1204	+3/-3	57740	Anthropic	Proprietary	2023/8
6	Command R+	1194	+5/-5	17404	Cohere	CC-BY-NC-4.0	2024/3
6	GPT-4-0314	1189	+4/-3	41292	OpenAI	Proprietary	2021/9
8	Claude 3 Haiku	1182	+3/-4	50689	Anthropic	Proprietary	2023/8
9	GPT-4-0613	1164	+3/-3	60213	OpenAI	Proprietary	2021/9
9	Mistral-Large-2402	1158	+3/-4	35075	Mistral	Proprietary	Unknown
10	Qwen1.5-72B-Chat	1153	+4/-5	27050	Alibaba	Qianwen LICENSE	2024/2

https://chat.lmsys.org/

Software Engineering Research Centre
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Software Architecture is also about Abstraction



Components

[Decisions]

Why SPA?

Connectors

[Decisions]

Why Spring?

Single-Page Application
[Container: JavaScript and Angular]
Provides all of the Internet banking functionality to customers via their web browser.

Mobile App
[Container: Xamarin]
Provides a limited subset of the Internet banking functionality to customers via their mobile device.

Makes API calls to
[JSON/HTTPS]

Makes API calls to
[JSON/HTTPS]

Makes API calls to
[JSON/HTTPS]

Makes API calls to
[JSON/HTTPS]

Makes API calls to
[JSON/HTTPS]

Makes API calls to
[JSON/HTTPS]

Sign In Controller
[Component: Spring MVC Rest Controller]
Allows users to sign in to the Internet Banking System.

Reset Password Controller
[Component: Spring MVC Rest Controller]
Allows users to reset their passwords with a single use URL.

Accounts Summary Controller
[Component: Spring MVC Rest Controller]
Provides customers with a summary of their bank accounts.

Uses

Uses

Uses

Uses

Security Component
[Component: Spring Bean]
Provides functionality related to signing in, changing passwords, etc.

E-mail Component
[Component: Spring Bean]
Sends e-mails to users.

Mainframe Banking System Facade
[Component: Spring Bean]
A facade onto the mainframe banking system.

API Application
[Container]

Reads from and writes to
[SQL/TCP]

Sends e-mail using

Makes API calls to
[XML/HTTPS]

Database
[Container: Oracle Database Schema]
Stores user registration information, hashed authentication credentials, access logs, etc.

E-mail System
[Software System]
The internal Microsoft Exchange e-mail system.

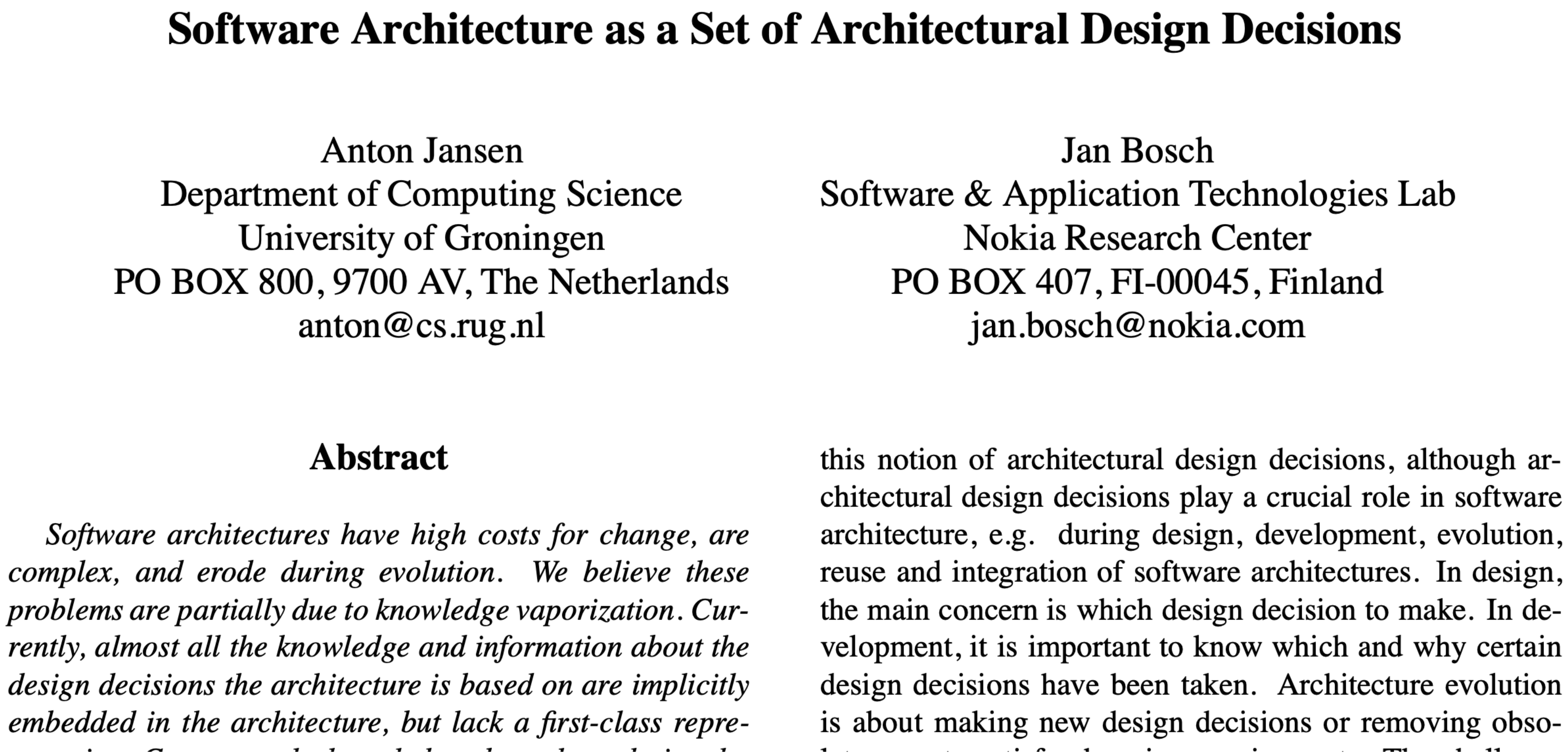
Mainframe Banking System
[Software System]
Stores all of the core banking information about customers, accounts, transactions, etc.

[Component] Internet Banking System - API Application

The component diagram for the API Application - diagram created with Structurizr.
Saturday, 11 November 2023 at 09:04 Greenwich Mean Time

How can all these help Software Architect?

Design Decisions is all you need!



Software Architecture as a Set of Architectural Design Decisions

Anton Jansen
Department of Computing Science
University of Groningen
PO BOX 800, 9700 AV, The Netherlands
anton@cs.rug.nl

Jan Bosch
Software & Application Technologies Lab
Nokia Research Center
PO BOX 407, FI-00045, Finland
jan.bosch@nokia.com

Abstract

Software architectures have high costs for change, are complex, and erode during evolution. We believe these problems are partially due to knowledge vaporization. Currently, almost all the knowledge and information about the design decisions the architecture is based on are implicitly embedded in the architecture, but lack a first-class repre-

this notion of architectural design decisions, although architectural design decisions play a crucial role in software architecture, e.g. during design, development, evolution, reuse and integration of software architectures. In design, the main concern is which design decision to make. In development, it is important to know which and why certain design decisions have been taken. Architecture evolution is about making new design decisions or removing obso-

Software Architecture is a set of key design decisions!!!

Jansen, A. and Bosch, J., 2005, November. Software architecture as a set of architectural design decisions. In *5th Working IEEE/IFIP Conference on Software Architecture (WICSA'05)*

The Key Issue

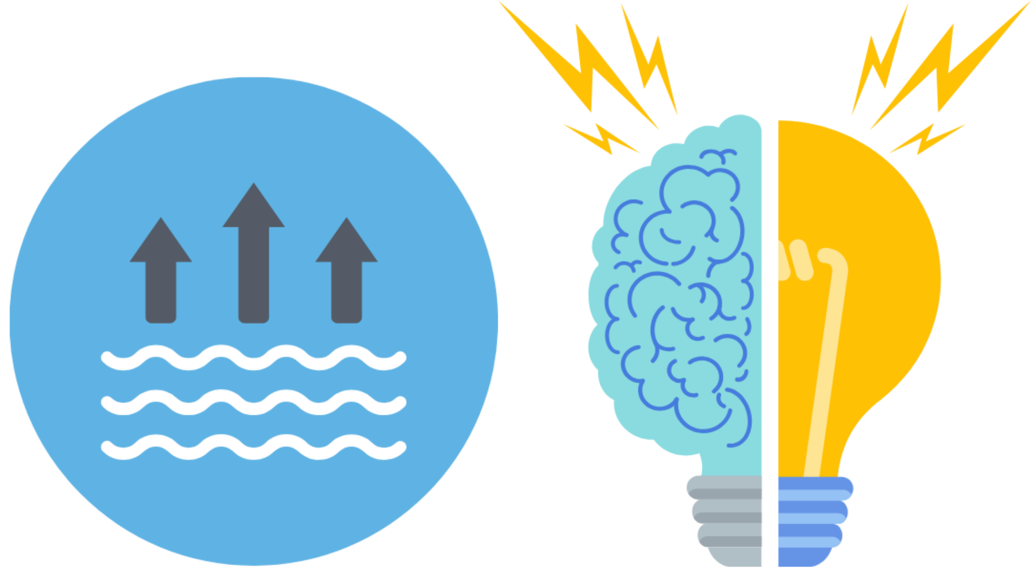
How does this work? Is there some documentation

I know how it works and why we did this. We did not document as such!

Takes away the knowledge!

?

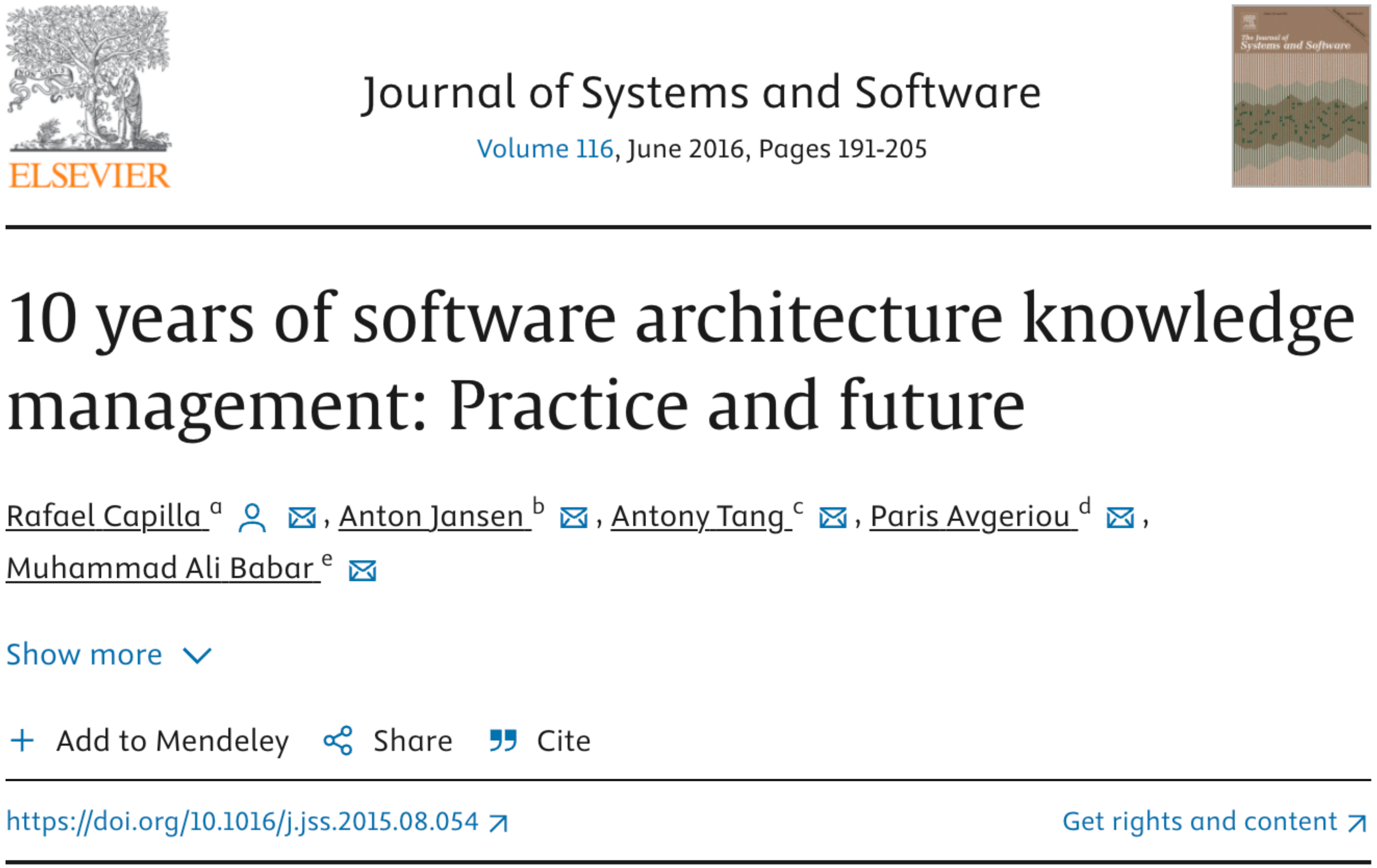
What happens when the dev/architect leaves?



Knowledge Vaporisation!

Architecture Knowledge Management

*Architecture knowledge management (AKM) aims to **codify and maintain** the Architectural knowledge of a software system in a form that can be **easily accessed** by different stakeholders*



ELSEVIER

Journal of Systems and Software
Volume 116, June 2016, Pages 191-205

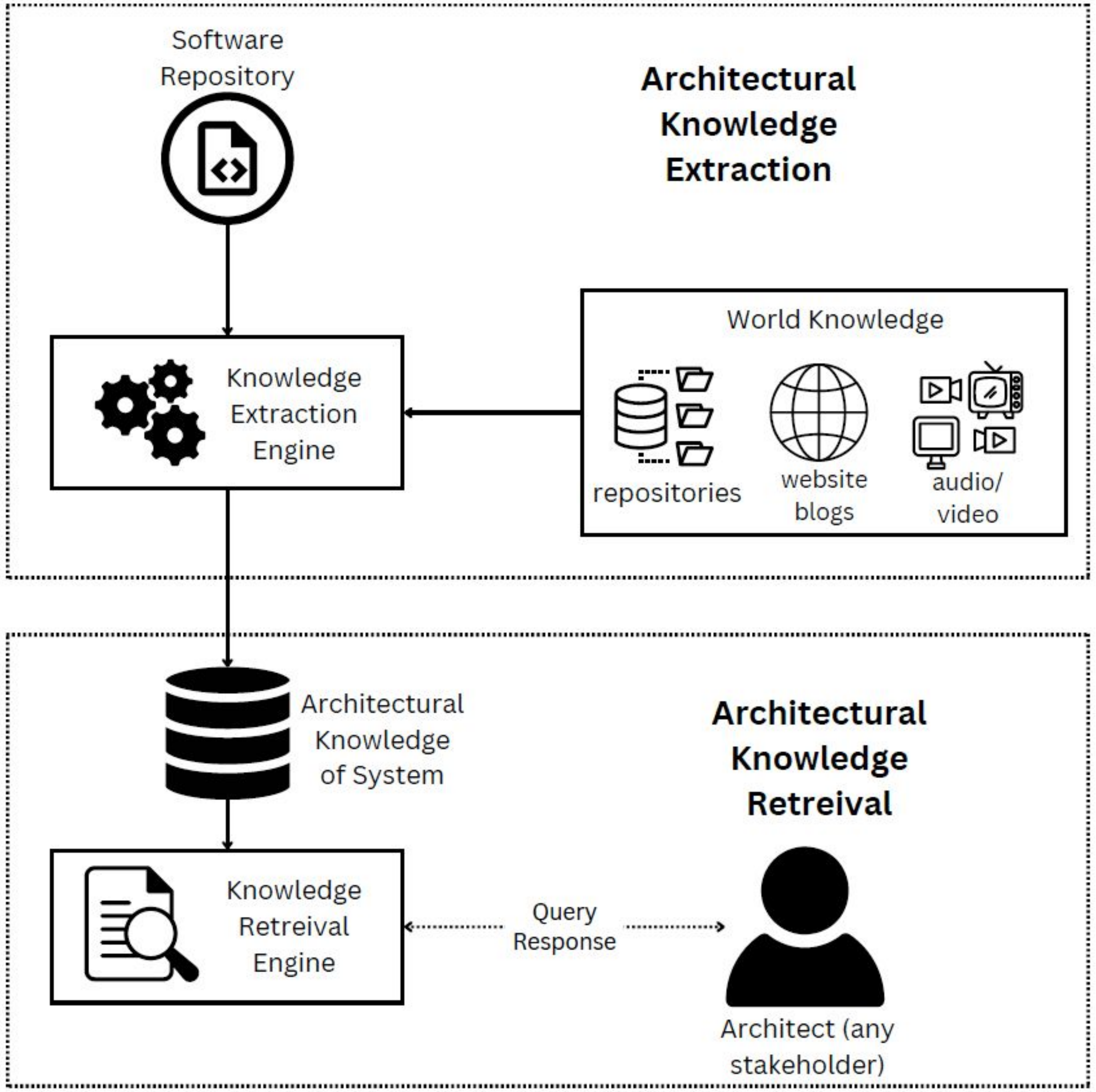
10 years of software architecture knowledge management: Practice and future

Rafael Capilla [a], Anton Jansen [b], Antony Tang [c], Paris Avgeriou [d], Muhammad Ali Babar [e]

https://doi.org/10.1016/j.jss.2015.08.054

Need for better tools => Automate using **ArchBots or a co-pilot**

Generative AI for Architectural Knowledge Management



Architectural Knowledge Extraction

Software Repository

Knowledge Extraction Engine

World Knowledge

repositories

website blogs

audio/ video

Architectural Knowledge of System

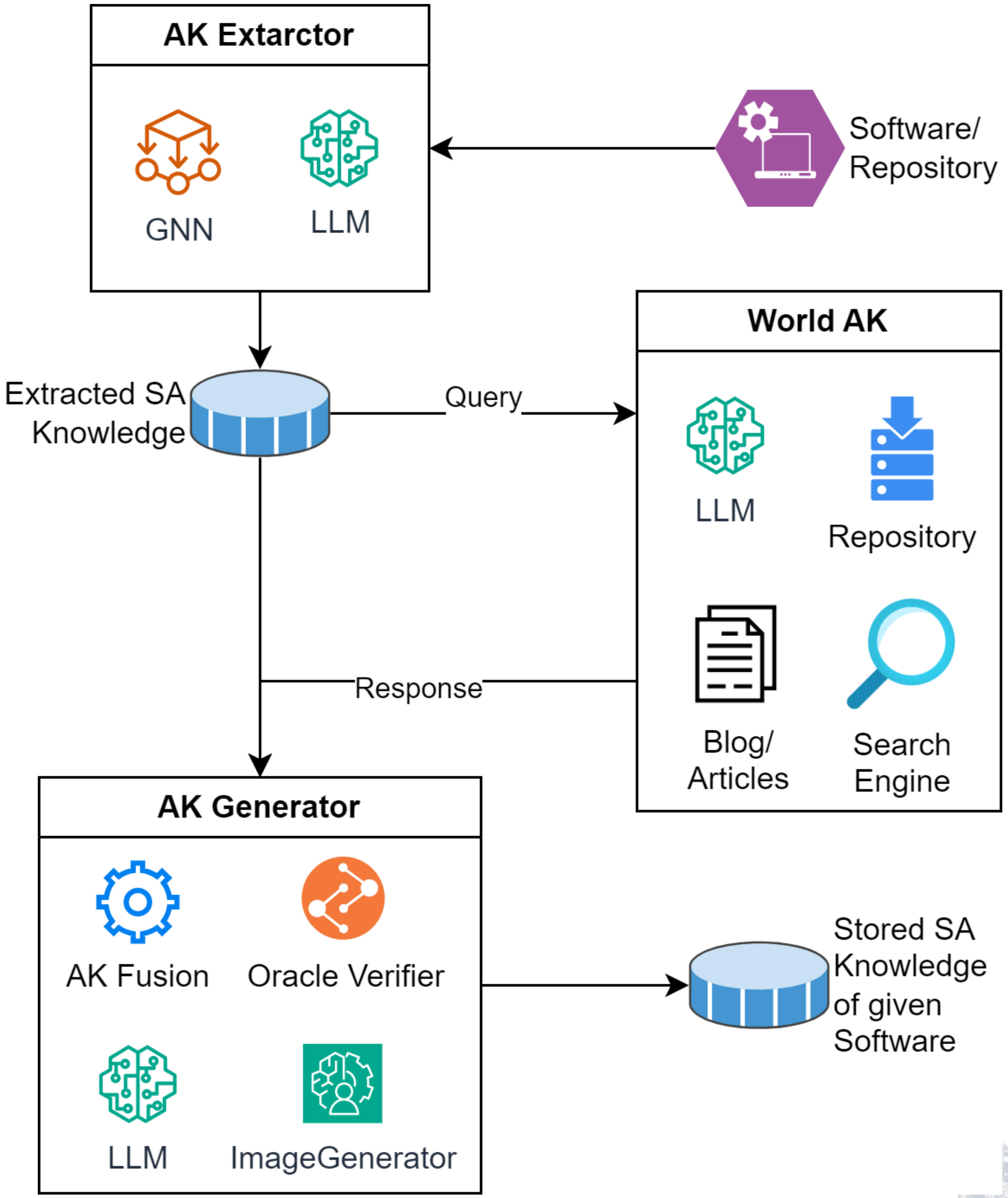
Architectural Knowledge Retreival

Knowledge Retreival Engine

Query

Response

Architect (any stakeholder)



AK Extarctor

GNN

LLM

Software/ Repository

Extracted SA Knowledge

Query

World AK

LLM

Repository

Blog/ Articles

Search Engine

Response

AK Generator

AK Fusion

Oracle Verifier

LLM

ImageGenerator

Stored SA Knowledge of given Software

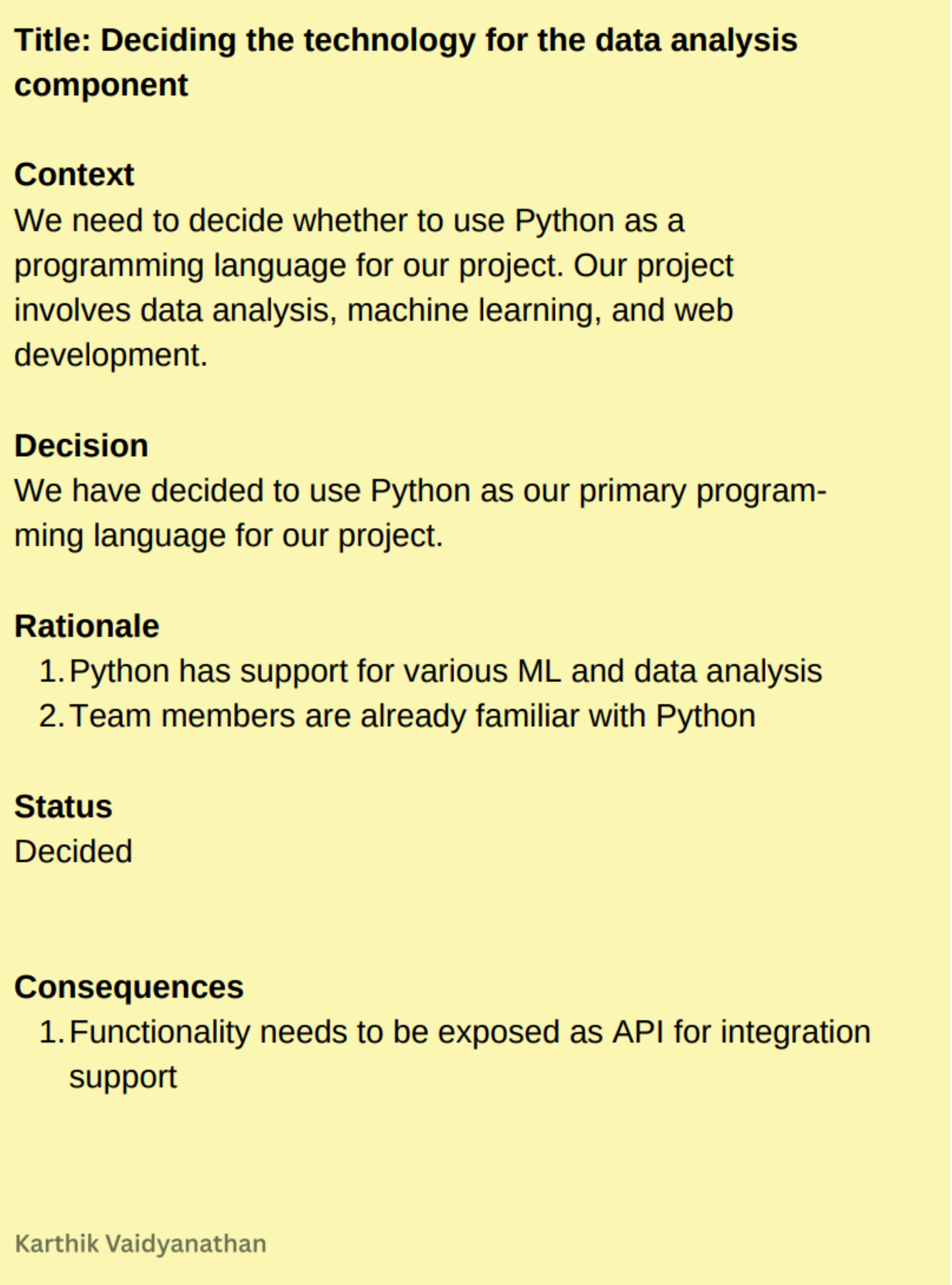
Rudra Dhar, Karthik Vaidhyanathan, Vasudeva Varma, **GenAI for Architecture Knowledge Management**, ICSA 2024, Poster

Software Engineering Research Centre
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Starting with Design Decisions

- **Architecture Decision Records: ADR**
- Lightweight mechanism for documenting decisions
- Design decisions require careful considerations of various parameters
  - This requires broader understanding of domain as well as expertise
  - **Can we use LLMs to generate architecture design decisions?**
  - **Can LLMs be used to extract architectural information from design decisions?**

https://github.com/joelparkerhenderson/architecture-decision-record



Title: Deciding the technology for the data analysis component

Context

We need to decide whether to use Python as a programming language for our project. Our project involves data analysis, machine learning, and web development.

Decision

We have decided to use Python as our primary programming language for our project.

Rationale

1. Python has support for various ML and data analysis
2. Team members are already familiar with Python

Status

Decided

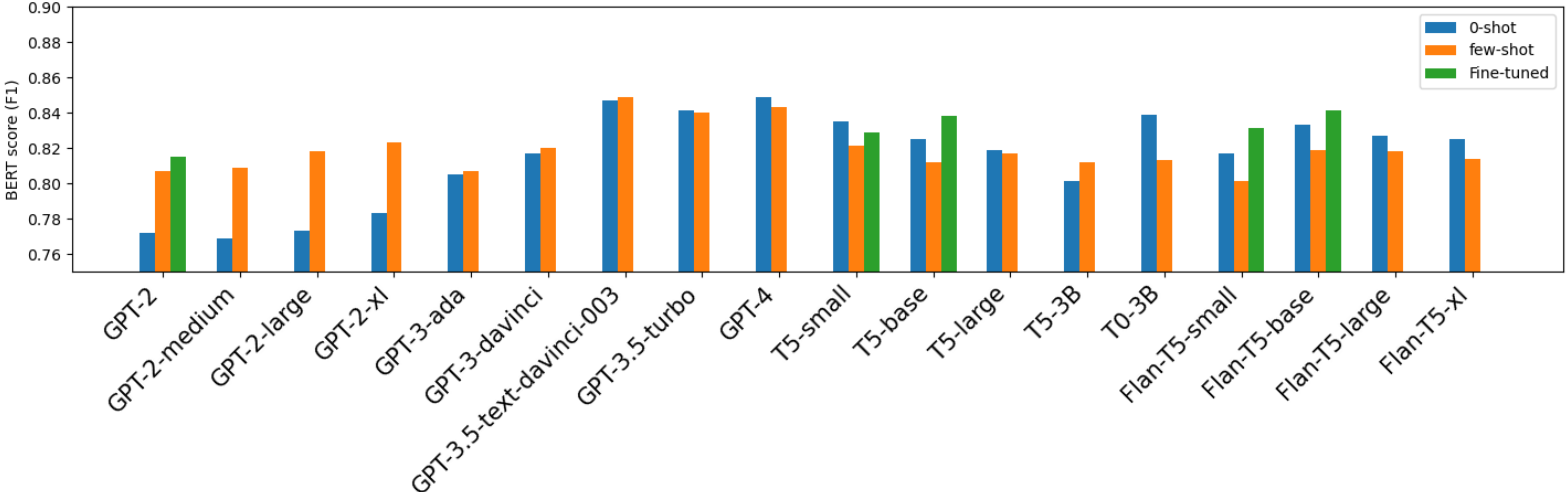
Consequences

1. Functionality needs to be exposed as API for integration support

Karthik Vaidyanathan

The Overall Results

Study repo



Smaller models perform well!…there is a need for more data and context!

Rudra Dhar, Karthik Vaidhyanathan, Vasudeva Varma, **Can LLMs generate Architecture Design Decisions: An Exploratory Empirical Study**, ICSA 2024 https://arxiv.org/abs/2403.01709

Generated Decisions

GPT4 zero shot

We will use Python as our programming language for this project. Python has emerged as a reliable, versatile language with a rich ecosystem of libraries and frameworks that cater to our specific needs including data analysis (Pandas, NumPy), machine learning (scikit-learn, TensorFlow), and web

GPT3.5 few shot

After considering the various criteria discussed, we have decided to go ahead and use Python for our project.

Python is a versatile, powerful language with a wide range of features, .. community of developers who can provide support.. Python is also open-source, free to use

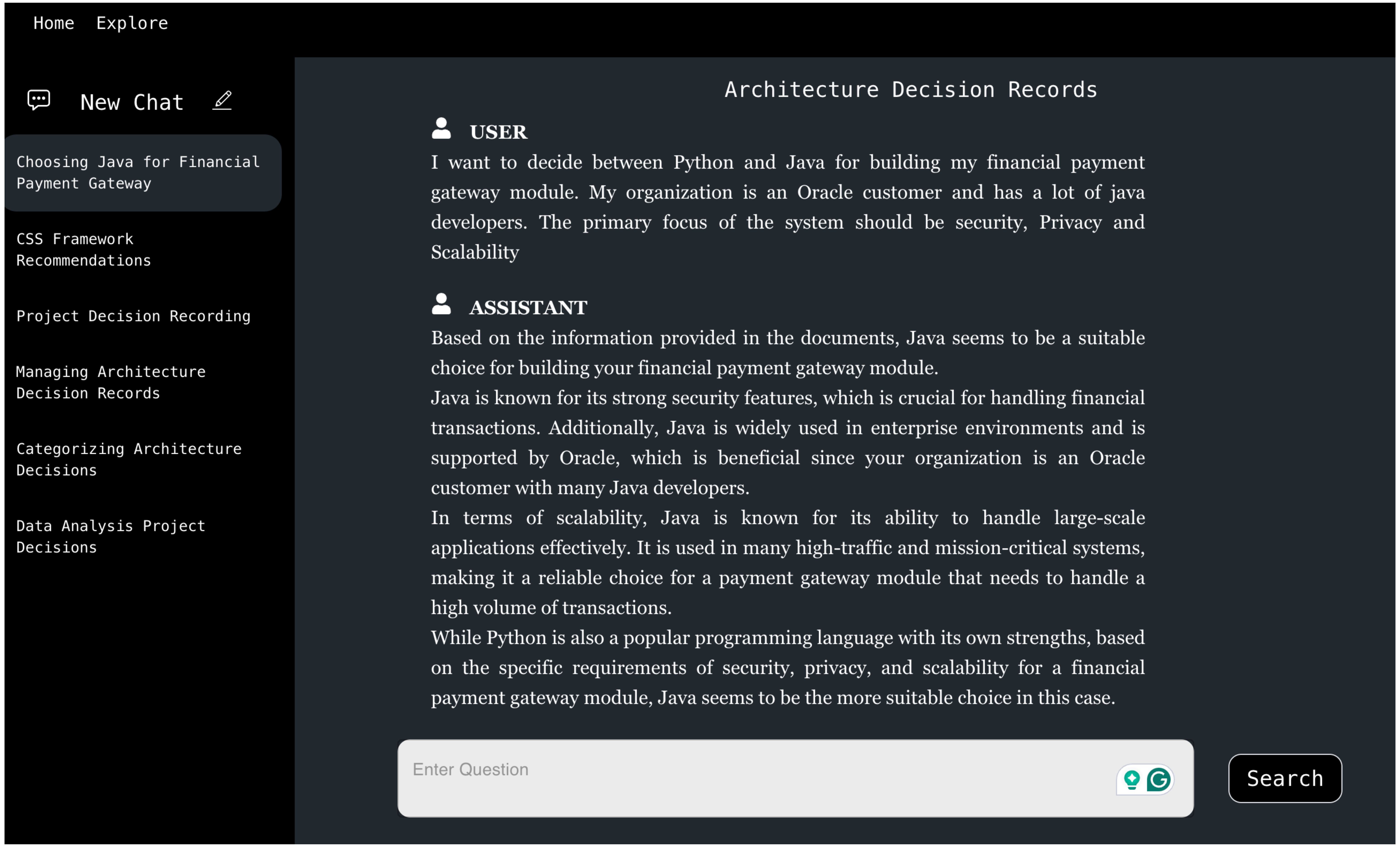
Python is also a great choice in terms of **performance and scalability….**, Python is an interpreted language..easier to debug and **maintain** than compiled languages.

Overall, Python is a great choice for our

Flan-T5 fine-tuned

It will have a large dataset, which will display a lot of data. This can help us in preparing the application. We need to choose Python as a programming language, because all of our projects will require Python to use it.

Introducing the ArchCopilot (Support RAG)


Home
Explore
New Chat
Choosing Java for Financial Payment Gateway
CSS Framework Recommendations
Project Decision Recording
Managing Architecture Decision Records
Categorizing Architecture Decisions
Data Analysis Project Decisions
Enter Question
Search

Architecture Decision Records

USER

I want to decide between Python and Java for building my financial payment gateway module. My organization is an Oracle customer and has a lot of java developers. The primary focus of the system should be security, Privacy and Scalability

ASSISTANT

Based on the information provided in the documents, Java seems to be a suitable choice for building your financial payment gateway module.

Java is known for its strong security features, which is crucial for handling financial transactions. Additionally, Java is widely used in enterprise environments and is supported by Oracle, which is beneficial since your organization is an Oracle customer with many Java developers.

In terms of scalability, Java is known for its ability to handle large-scale applications effectively. It is used in many high-traffic and mission-critical systems, making it a reliable choice for a payment gateway module that needs to handle a high volume of transactions.

While Python is also a popular programming language with its own strengths, based on the specific requirements of security, privacy, and scalability for a financial payment gateway module, Java seems to be the more suitable choice in this case.

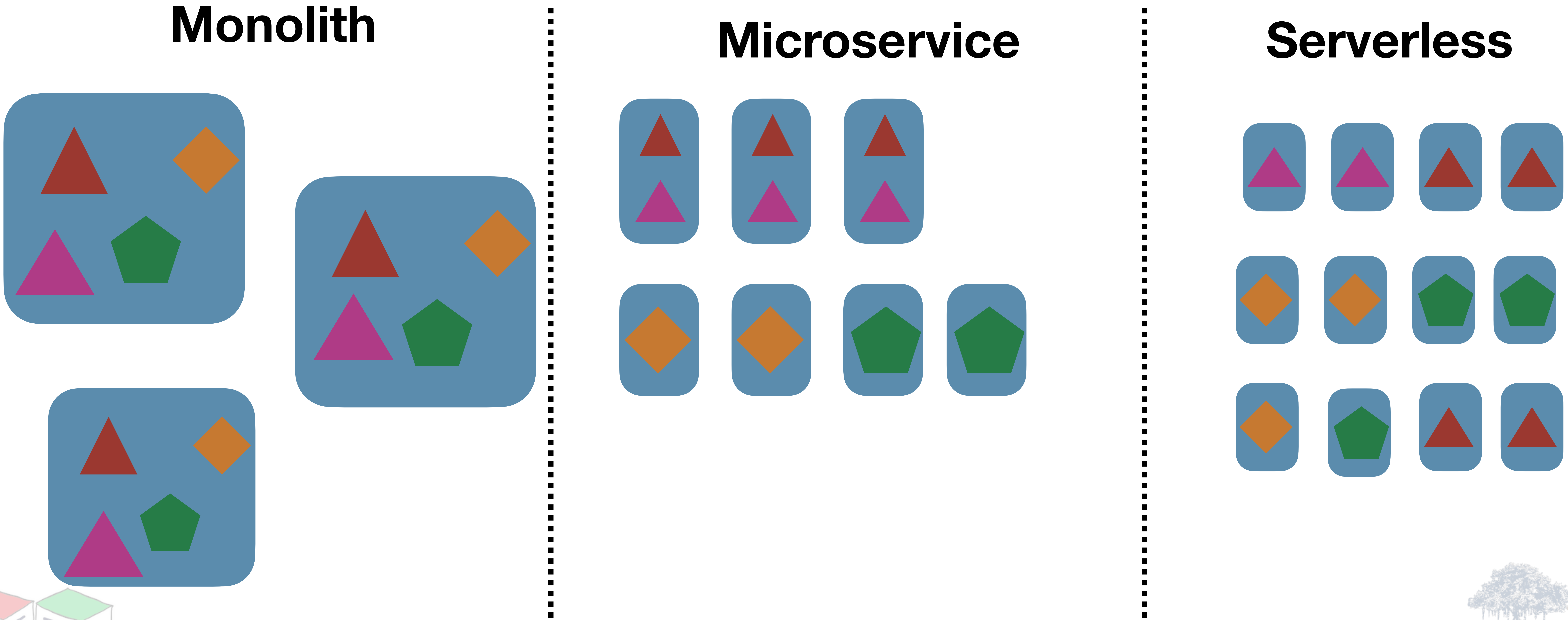
1. Add knowledge - decisions..
2. Search for similar decisions
3. Retrieve knowledge
4. Generate decision records for a given context

A companion for architect!

Thanks to Nikunj Garg, Pabba Ananya, Vansh Pravin, Yatharth Gupta, Ananya Sudi, Rupasree Yeduru and Rudra Dhar

Going beyond Design Decisions!

Can we automate generation of architectural components?



Monolith

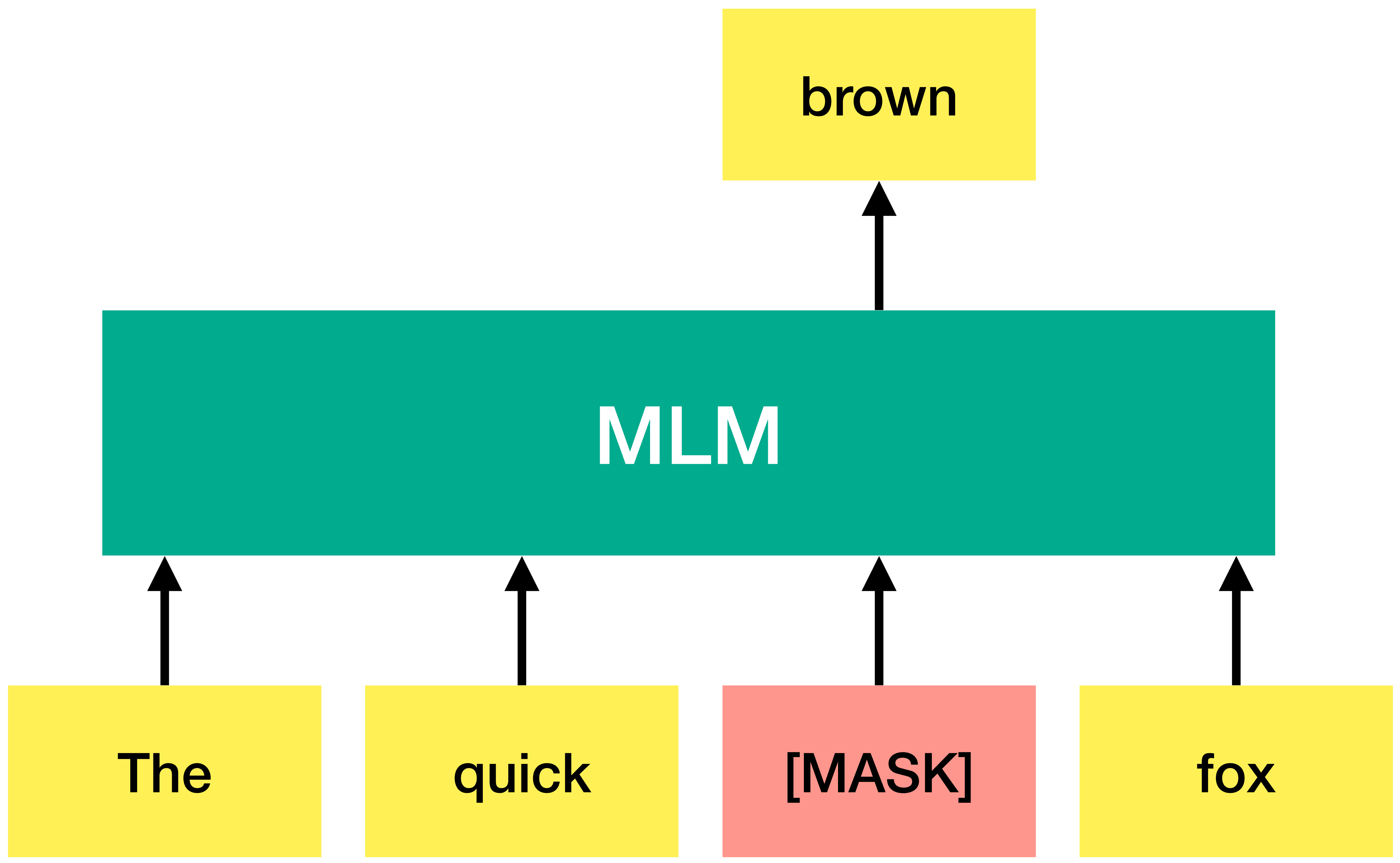
Microservice

Serverless

Image credits: Shrikara A

Study Design

Inspired from the concept of masked language modeling



brown

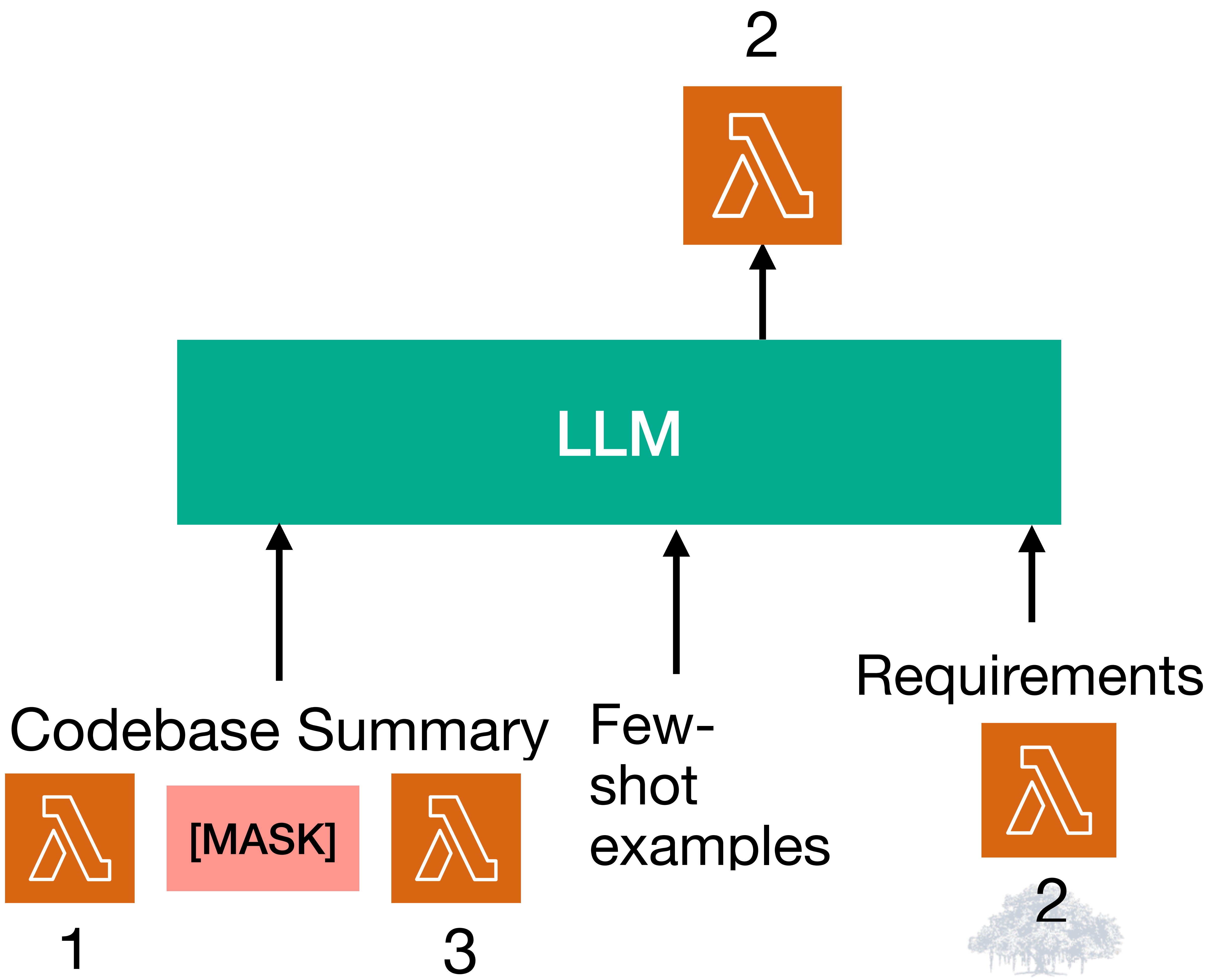
MLM

The

quick

[MASK]

fox



2

LLM

Codebase Summary

1

[MASK]

3

Few-shot examples

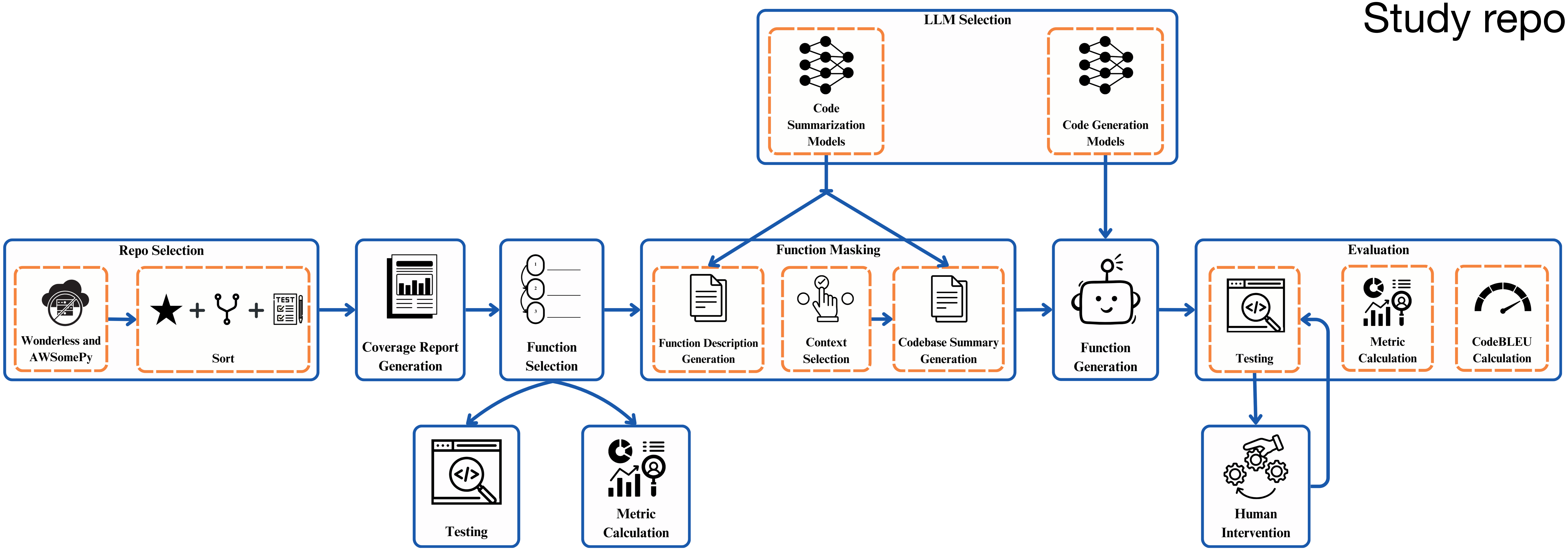
Requirements

2

Image credits: Shrikara A

Study Design

Study repo

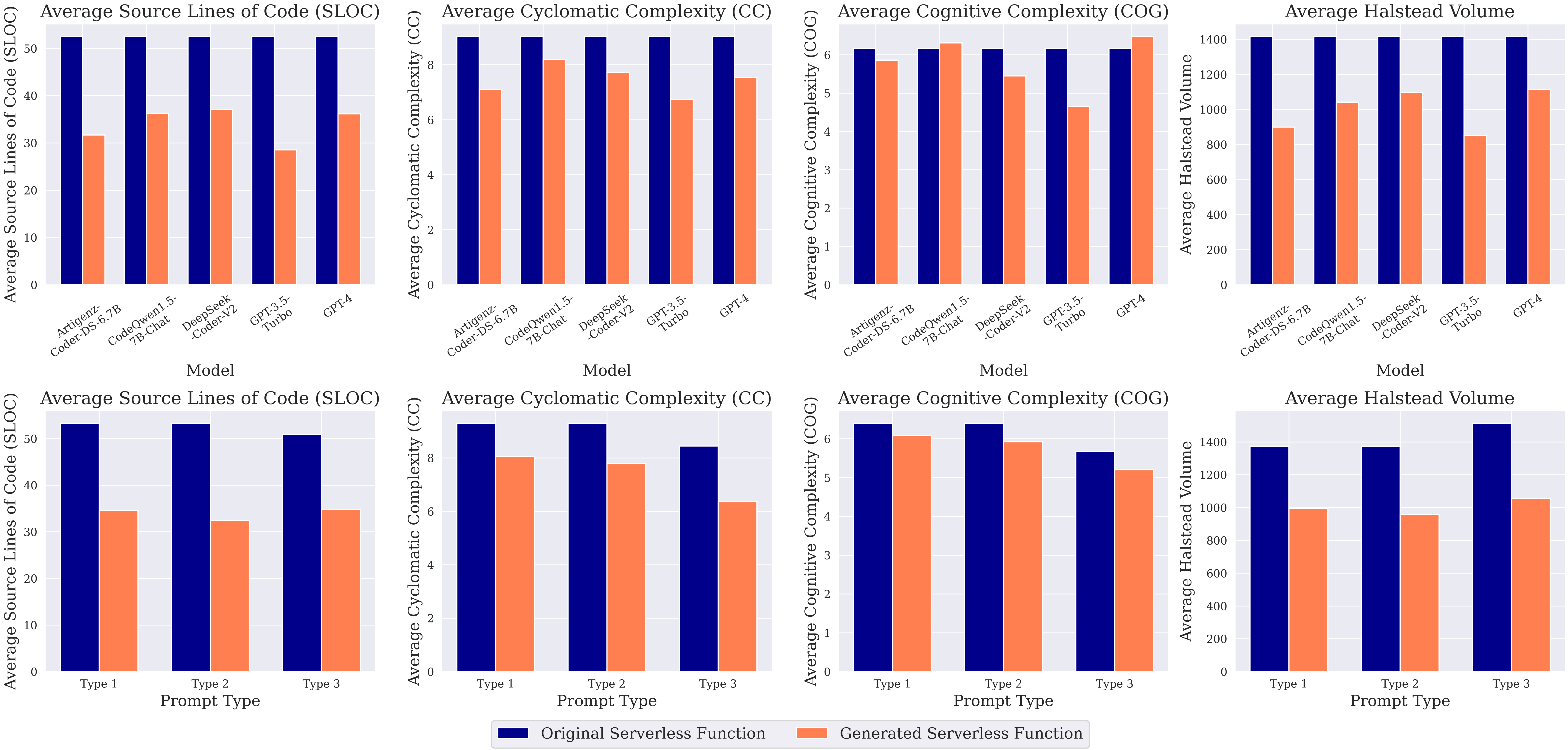


3 types of abstraction in prompt, 4 serverless repo, 5 models, 145 functions tested

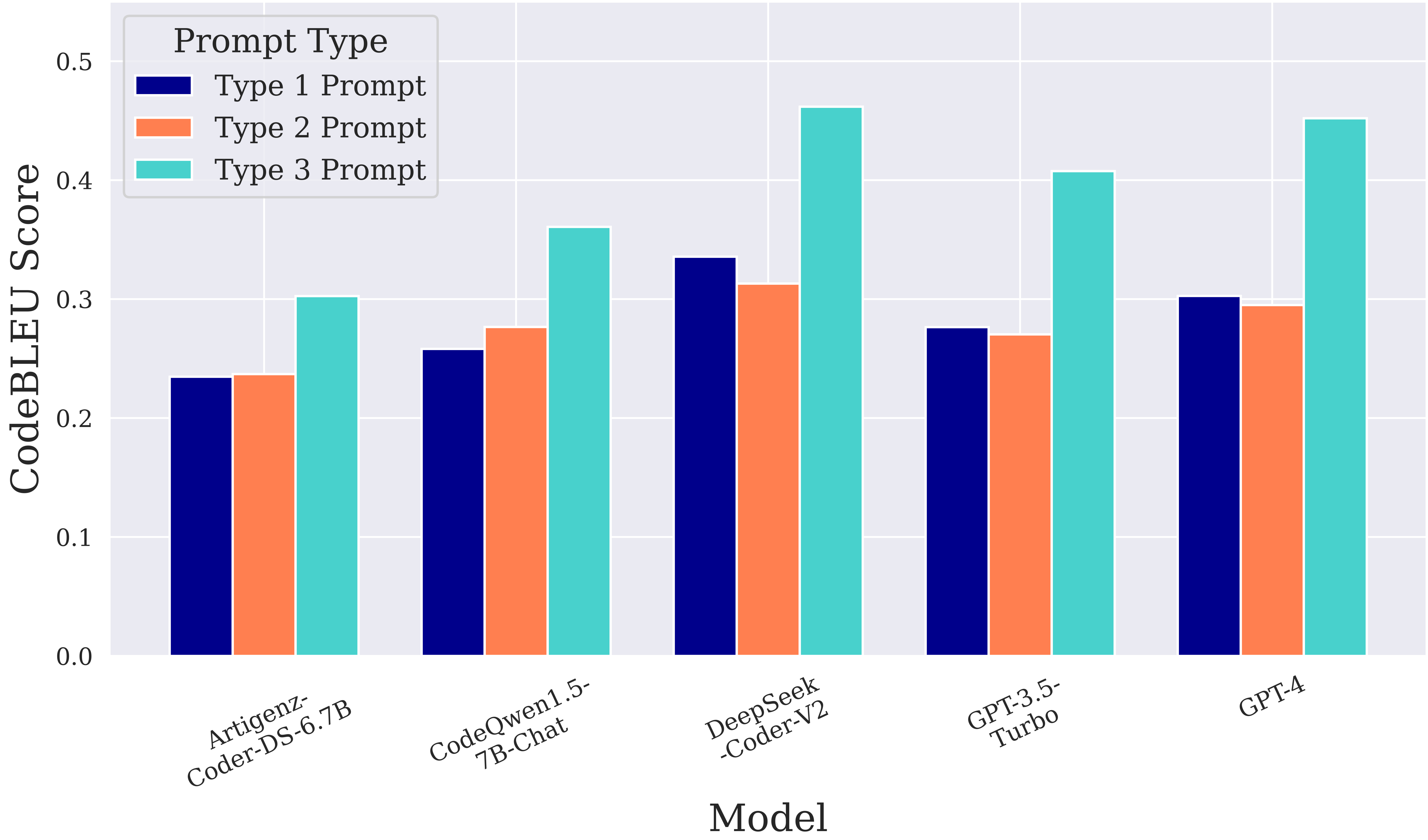
Some Key Results

Model	Initial Test Pass Rate	Type 1 (No interventio)	Type 1 (intervention)	Type 2 (No intervention)	Type 2 (intervention)	Type 3 (No intervention)	Type 3 (intervention)
Artigenz-Coder-DS-6.7B	100	0	0	0	0	0	0
CodeQwen1.5-7B-Chat	100	**3**	3	0	22	7	33
DeepSeek-Coder-V2	100	0	18	0	39	**13**	**71**
GPT-3.5-Turbo	100	0	3	0	50	4	64
GPT-4	100	0	**24**	0	**61**	10	50
Average	100	1	10	0	34	7	44

Some Key Results - Code Quality



Some Key Results - Code Quality

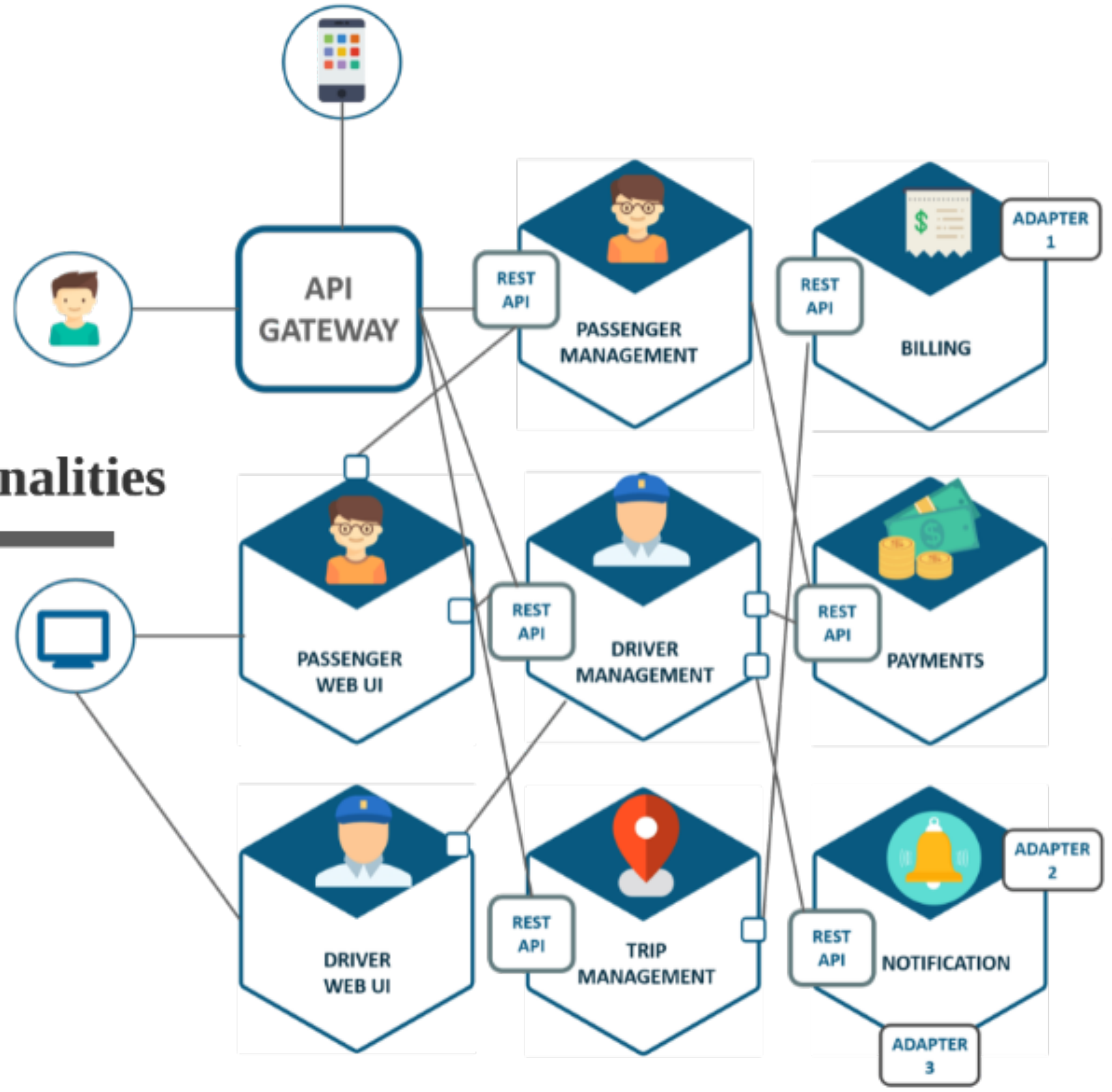


Human Architects + Models + devs => Great combination!!

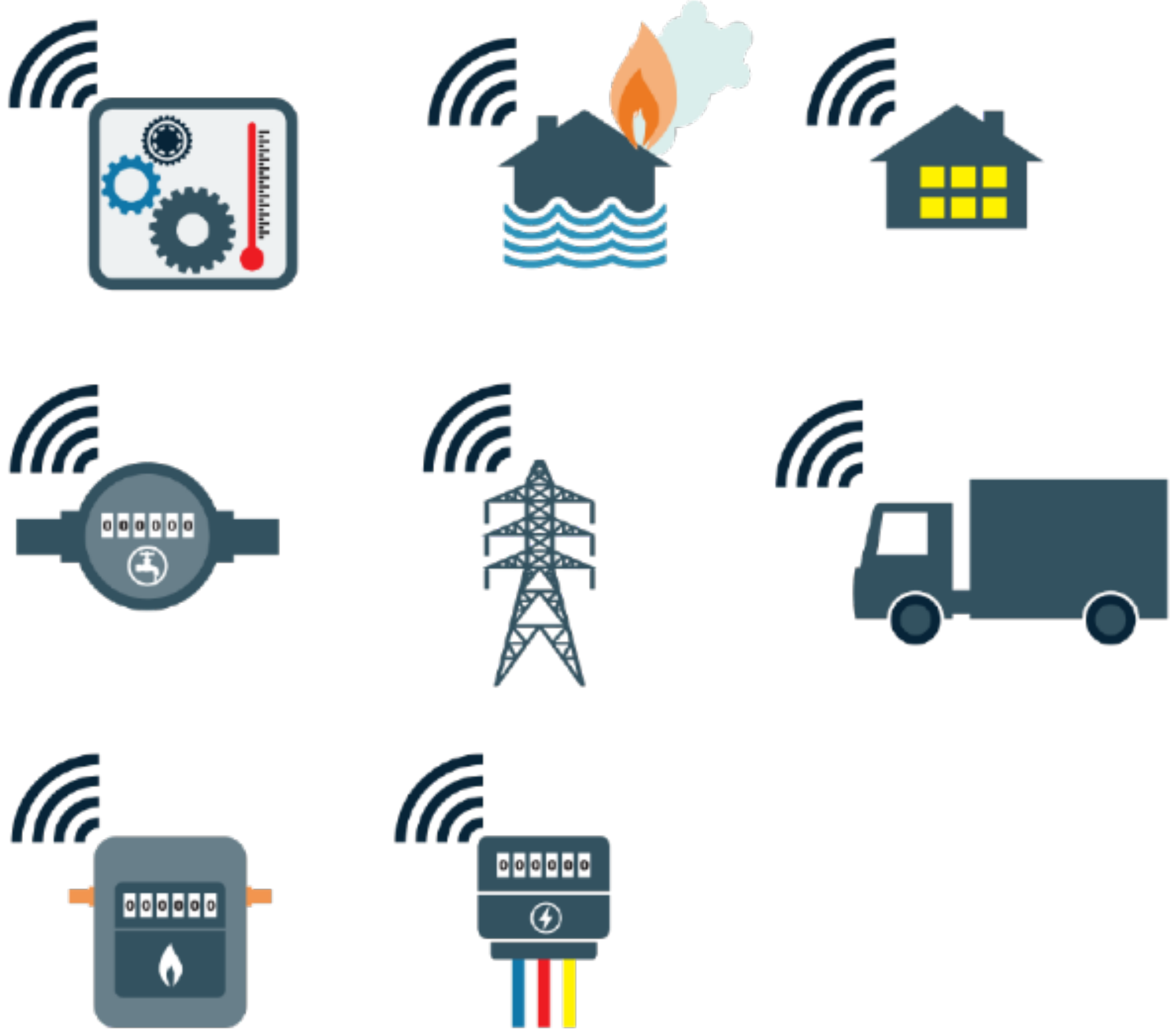
Dynamically Generate Service@runtime

Applying to IoT Systems

Functionalities

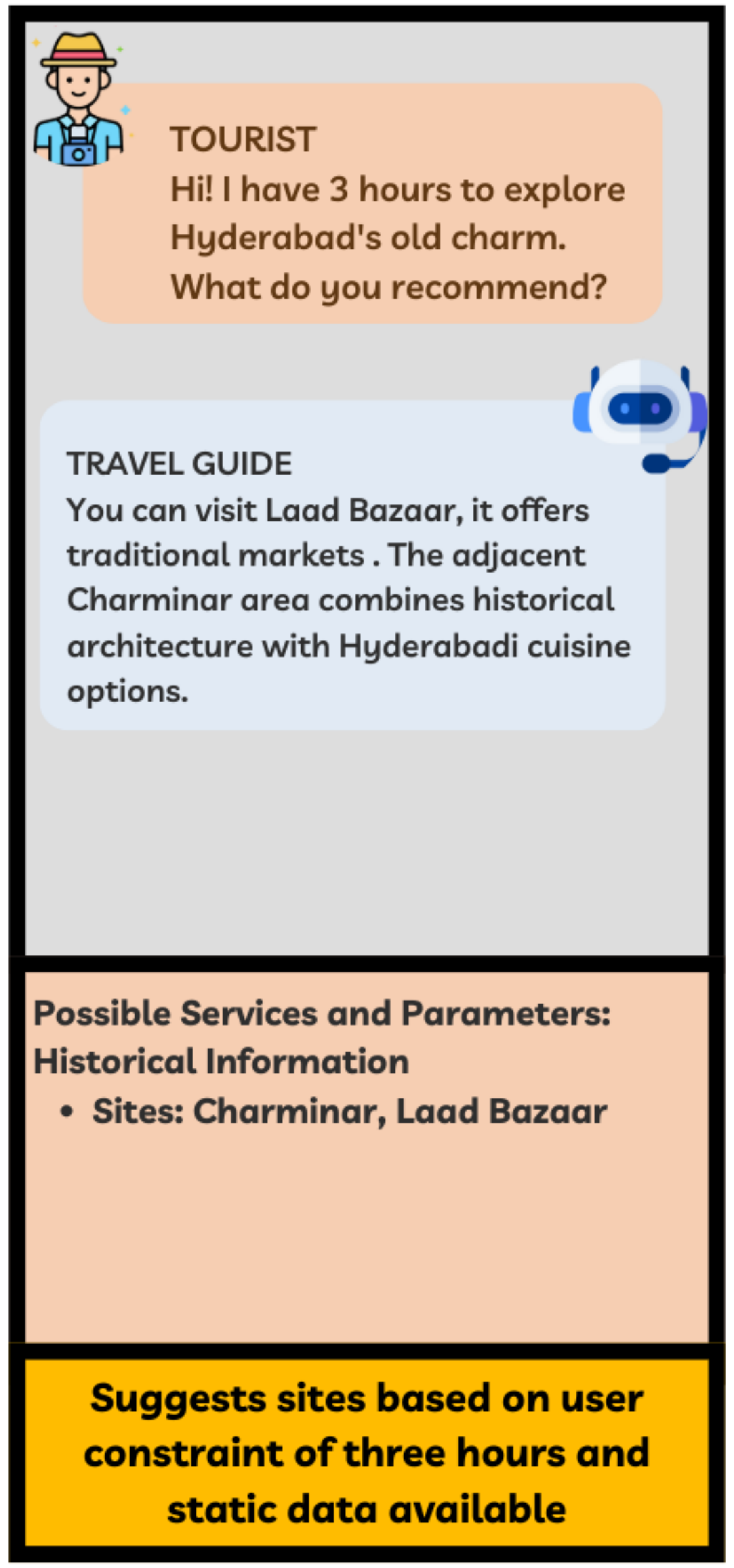


Data



Applying to IoT Systems

Dynamically generate services

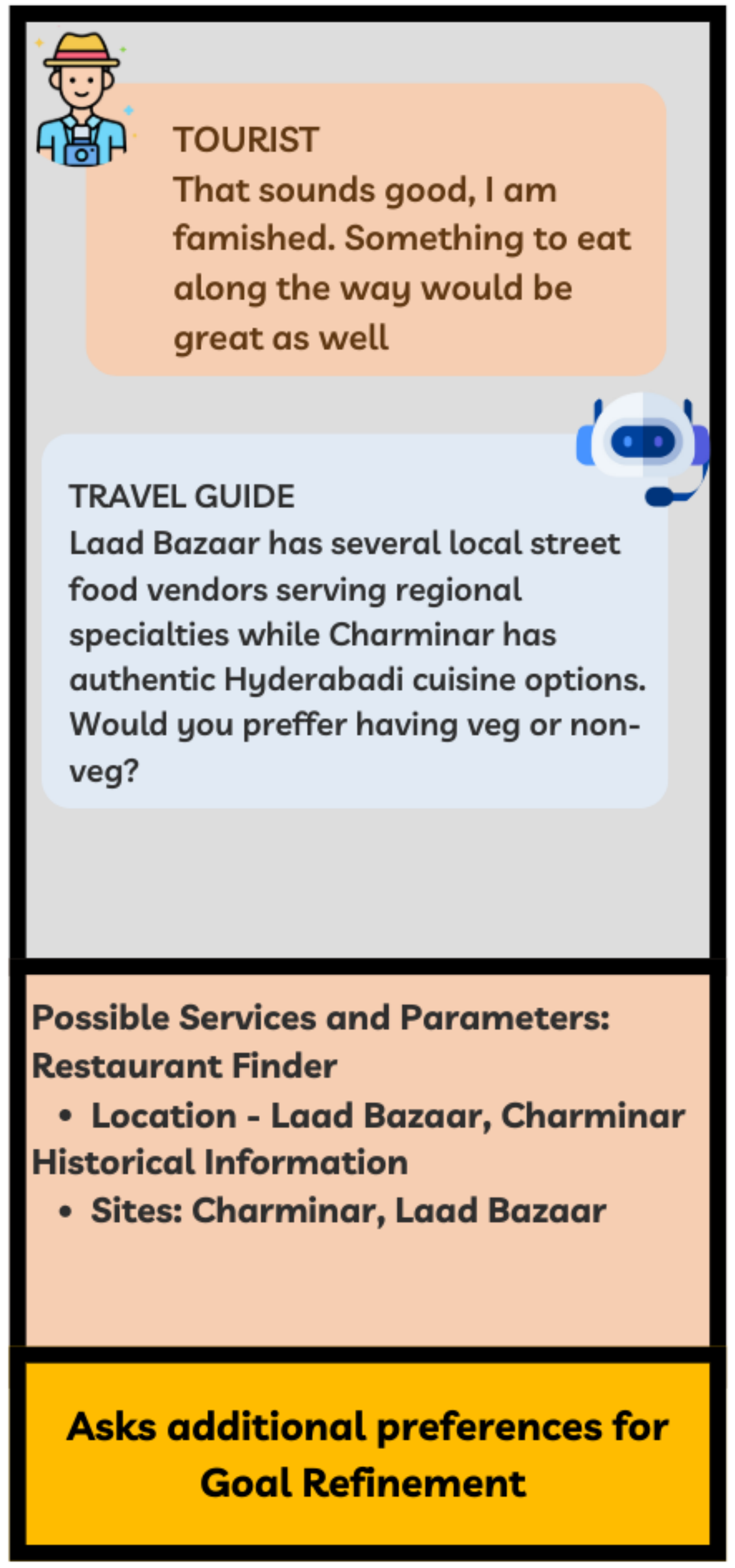


TOURIST
Hi! I have 3 hours to explore Hyderabad's old charm. What do you recommend?

TRAVEL GUIDE
You can visit Laad Bazaar, it offers traditional markets . The adjacent Charminar area combines historical architecture with Hyderabadi cuisine options.

Possible Services and Parameters:
Historical Information
- **Sites: Charminar, Laad Bazaar**

Suggests sites based on user constraint of three hours and static data available



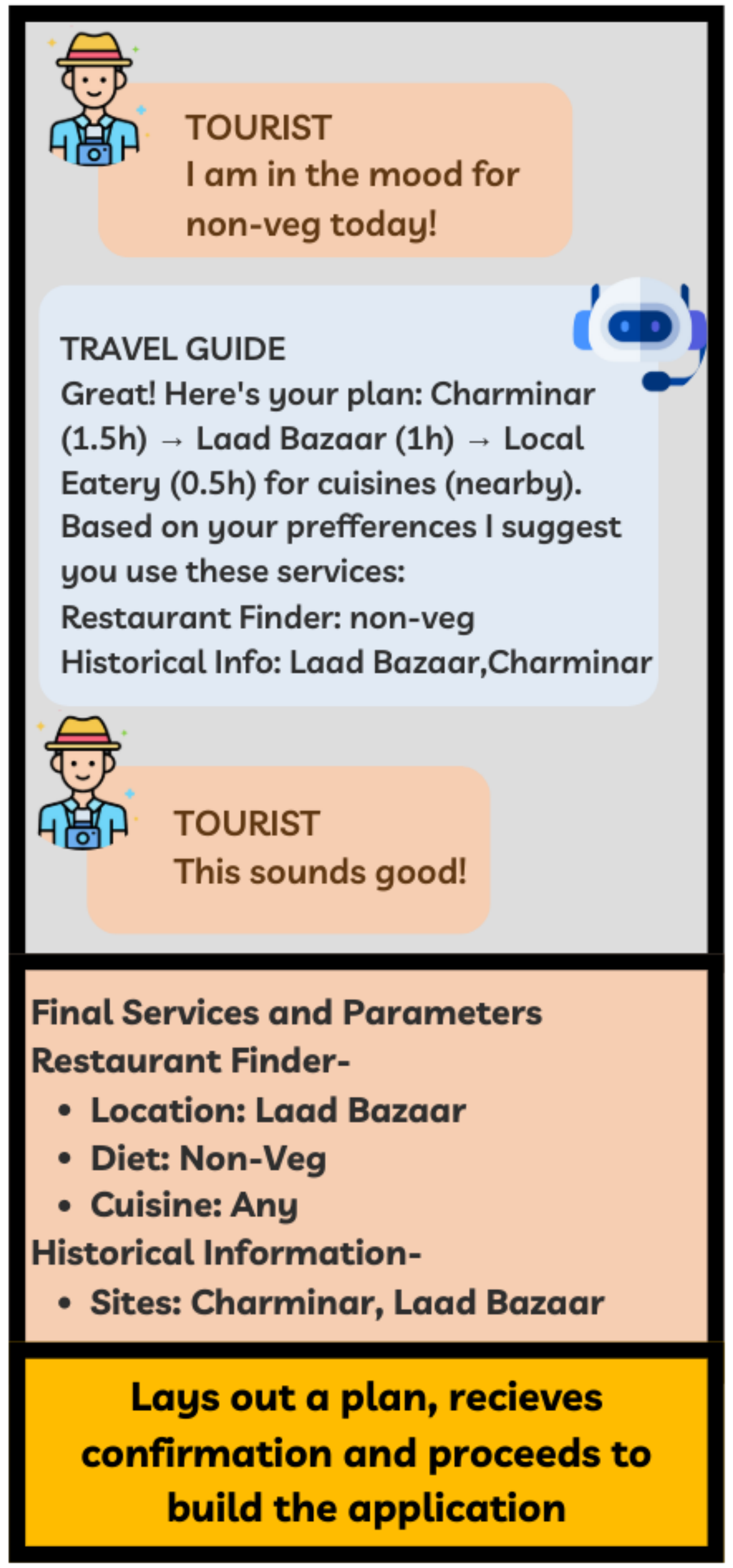
TOURIST
That sounds good, I am famished. Something to eat along the way would be great as well

TRAVEL GUIDE
Laad Bazaar has several local street food vendors serving regional specialties while Charminar has authentic Hyderabadi cuisine options. Would you preffer having veg or non-veg?

Possible Services and Parameters:
Restaurant Finder
- **Location - Laad Bazaar, Charminar**

Historical Information
- **Sites: Charminar, Laad Bazaar**

Asks additional preferences for Goal Refinement



TOURIST
I am in the mood for non-veg today!

TRAVEL GUIDE
Great! Here's your plan: Charminar (1.5h) → Laad Bazaar (1h) → Local Eatery (0.5h) for cuisines (nearby).
Based on your prefferences I suggest you use these services:
Restaurant Finder: non-veg
Historical Info: Laad Bazaar,Charminar

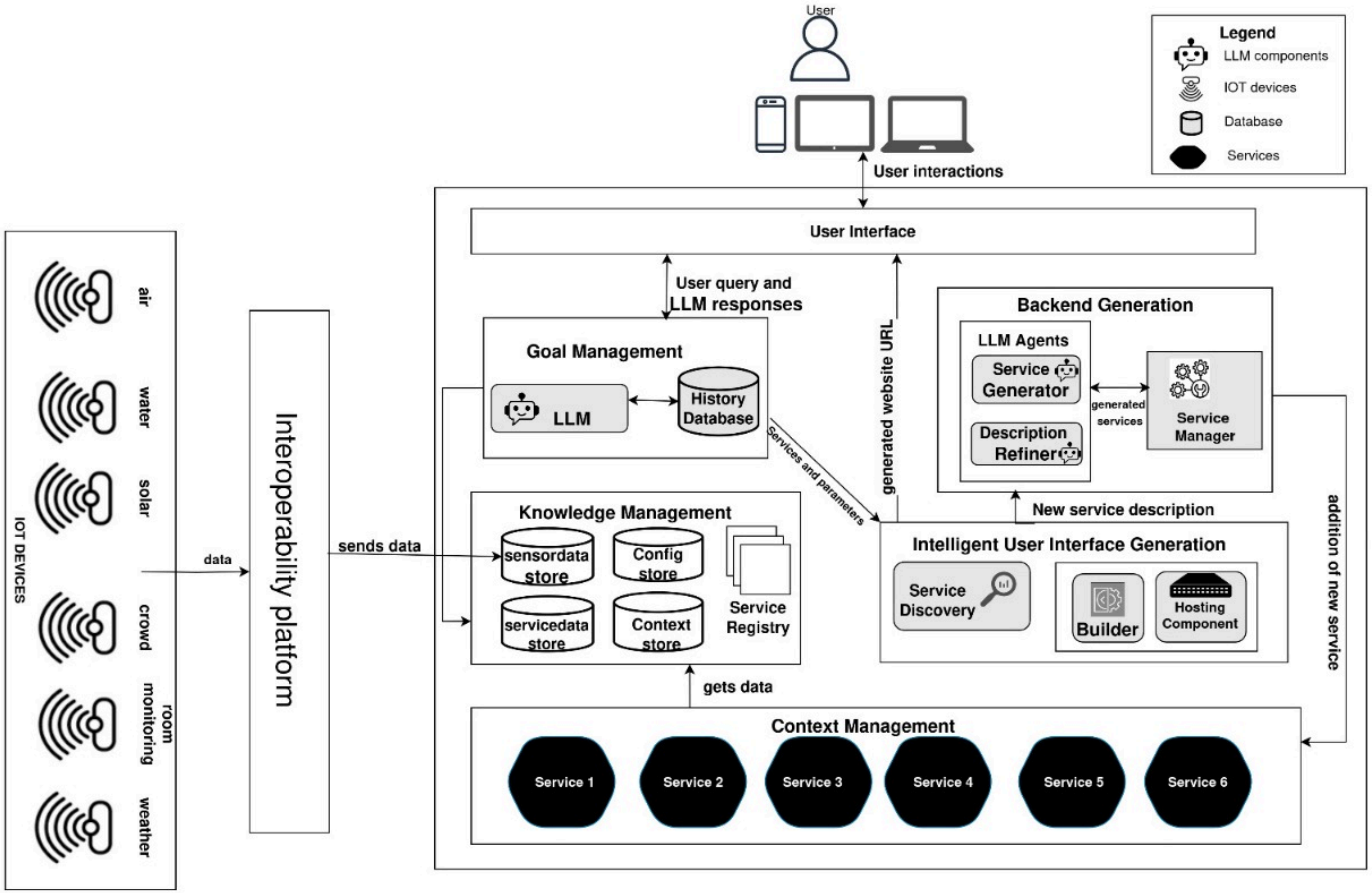
TOURIST
This sounds good!

Final Services and Parameters
Restaurant Finder-
- **Location: Laad Bazaar**
- **Diet: Non-Veg**
- **Cuisine: Any**

Historical Information-
- **Sites: Charminar, Laad Bazaar**

Lays out a plan, recieves confirmation and proceeds to build the application

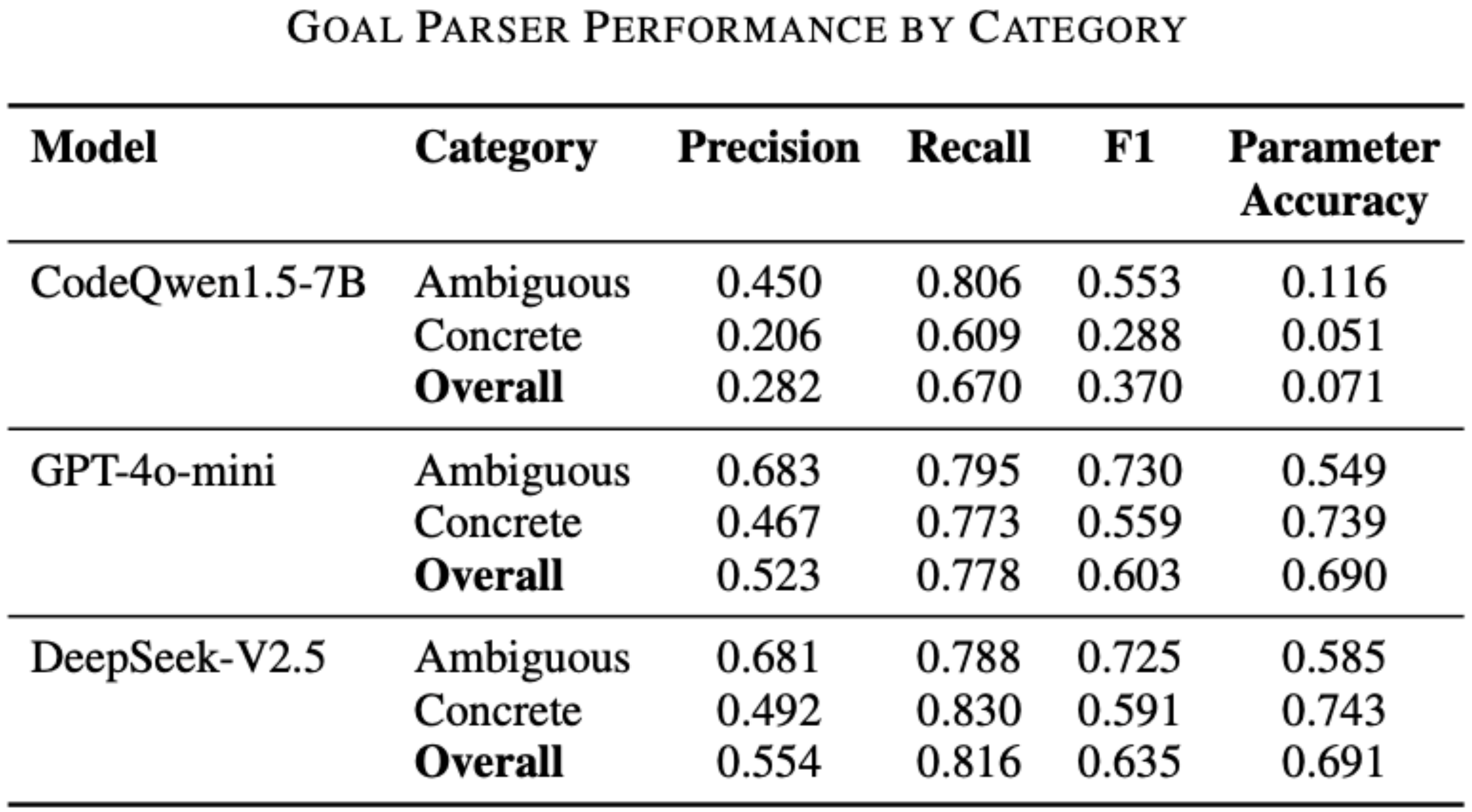
IoT-Together: Mixed Initiative Interactions



Adnan, B., Miryala, S., Sambu, A., Vaidhyanathan, K., De Sanctis, M. and Spalazzese, R., 2025. **Leveraging LLMs for Dynamic IoT Systems Generation through Mixed-Initiative Interaction**.ICSA 2025 Companion, To appear. *arXiv preprint arXiv:2502.00689*.

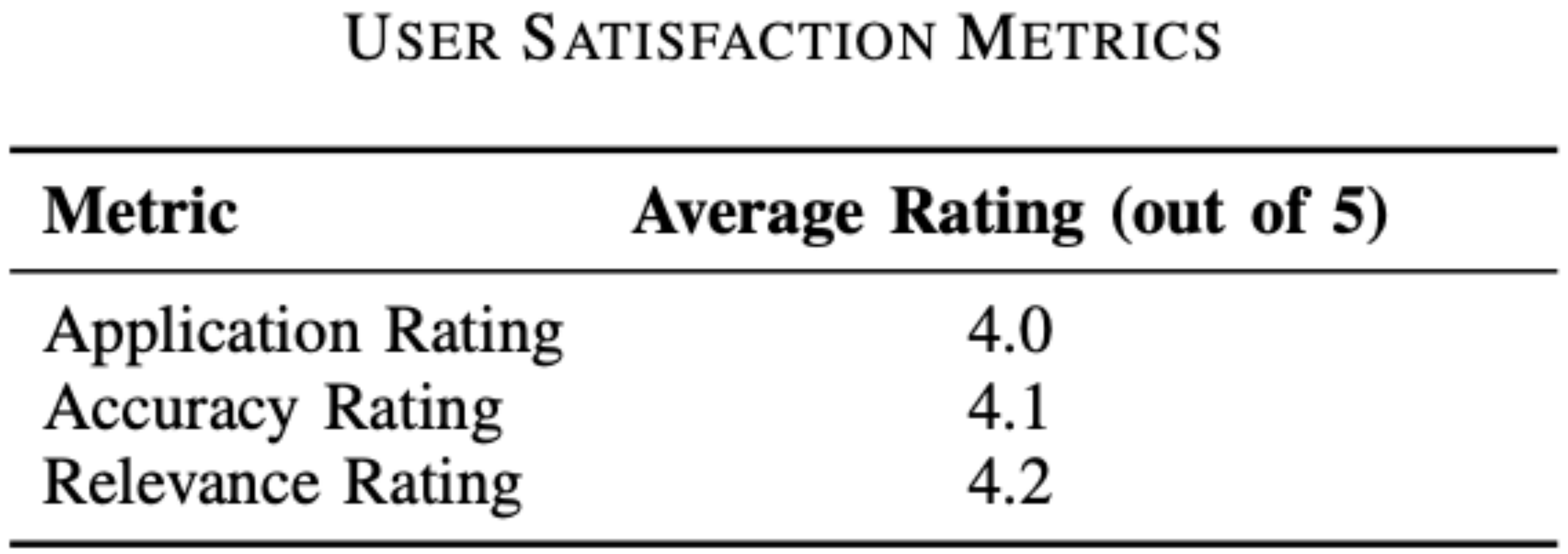
Some Results

Website



GOAL PARSER PERFORMANCE BY CATEGORY

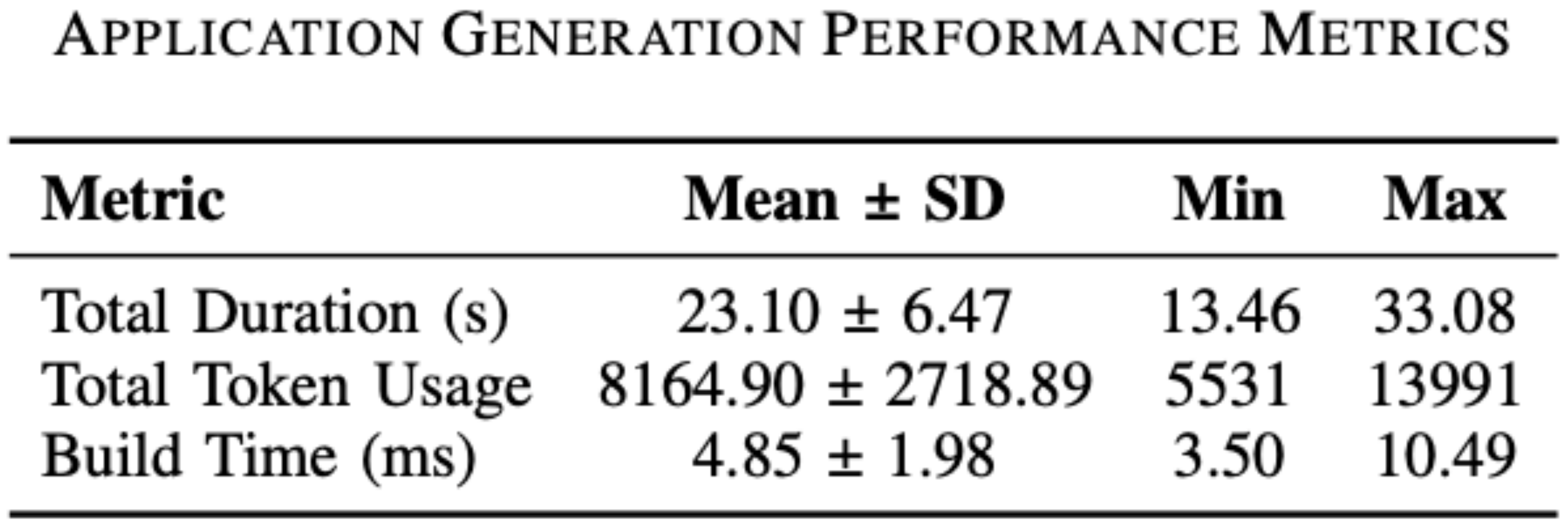
Model	Category	Precision	Recall	F1	Parameter Accuracy
CodeQwen1.5-7B	Ambiguous	0.450	0.806	0.553	0.116
	Concrete	0.206	0.609	0.288	0.051
	Overall	0.282	0.670	0.370	0.071
GPT-4o-mini	Ambiguous	0.683	0.795	0.730	0.549
	Concrete	0.467	0.773	0.559	0.739
	Overall	0.523	0.778	0.603	0.690
DeepSeek-V2.5	Ambiguous	0.681	0.788	0.725	0.585
	Concrete	0.492	0.830	0.591	0.743
	Overall	0.554	0.816	0.635	0.691



USER SATISFACTION METRICS

Metric	Average Rating (out of 5)
Application Rating	4.0
Accuracy Rating	4.1
Relevance Rating	4.2

Preliminary study with 15 users



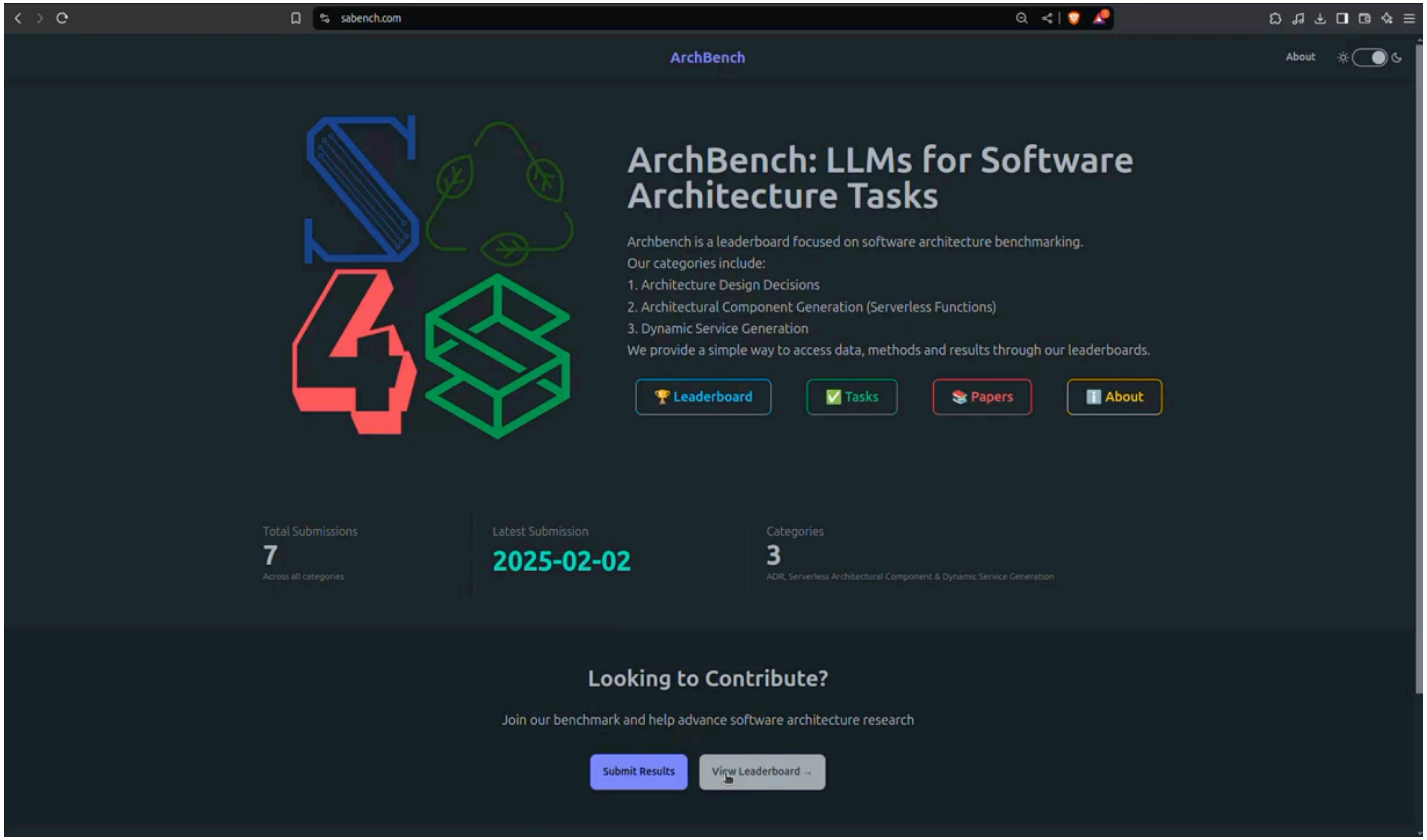
APPLICATION GENERATION PERFORMANCE METRICS

Metric	Mean ± SD	Min	Max
Total Duration (s)	23.10 ± 6.47	13.46	33.08
Total Token Usage	8164.90 ± 2718.89	5531	13991
Build Time (ms)	4.85 ± 1.98	3.50	10.49

Introducing sabench.com

LLMs for Software Architecture Tasks

- Collection of three tasks
  - ADR
  - Serverless function generation
  - Dynamic service generation
- Datasets and leaderboards
- You can contribute as well!
- **Live from today!**



ArchBench: LLMs for Software Architecture Tasks

Archbench is a leaderboard focused on software architecture benchmarking.
Our categories include:
1. Architecture Design Decisions
2. Architectural Component Generation (Serverless Functions)
3. Dynamic Service Generation
We provide a simple way to access data, methods and results through our leaderboards.

Total Submissions
7
Across all categories

Latest Submission
2025-02-02

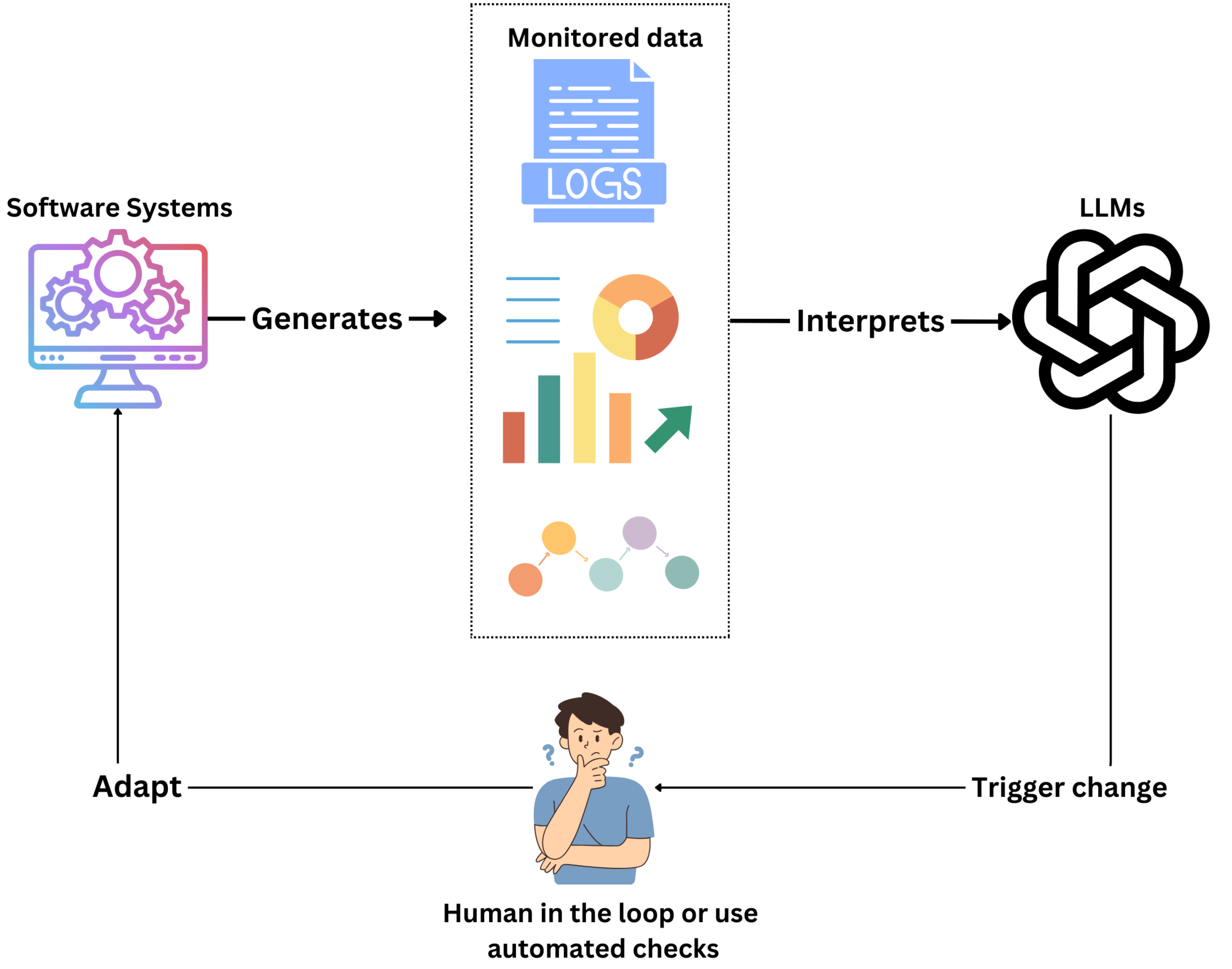
Categories
3
ADR, Serverless Architectural Component & Dynamic Service Generation

Looking to Contribute?

Join our benchmark and help advance software architecture research

Extending beyond

Design time to run-time adaptation - Can LLMs help?



Monitored data

Software Systems

Generates

Interprets

LLMs

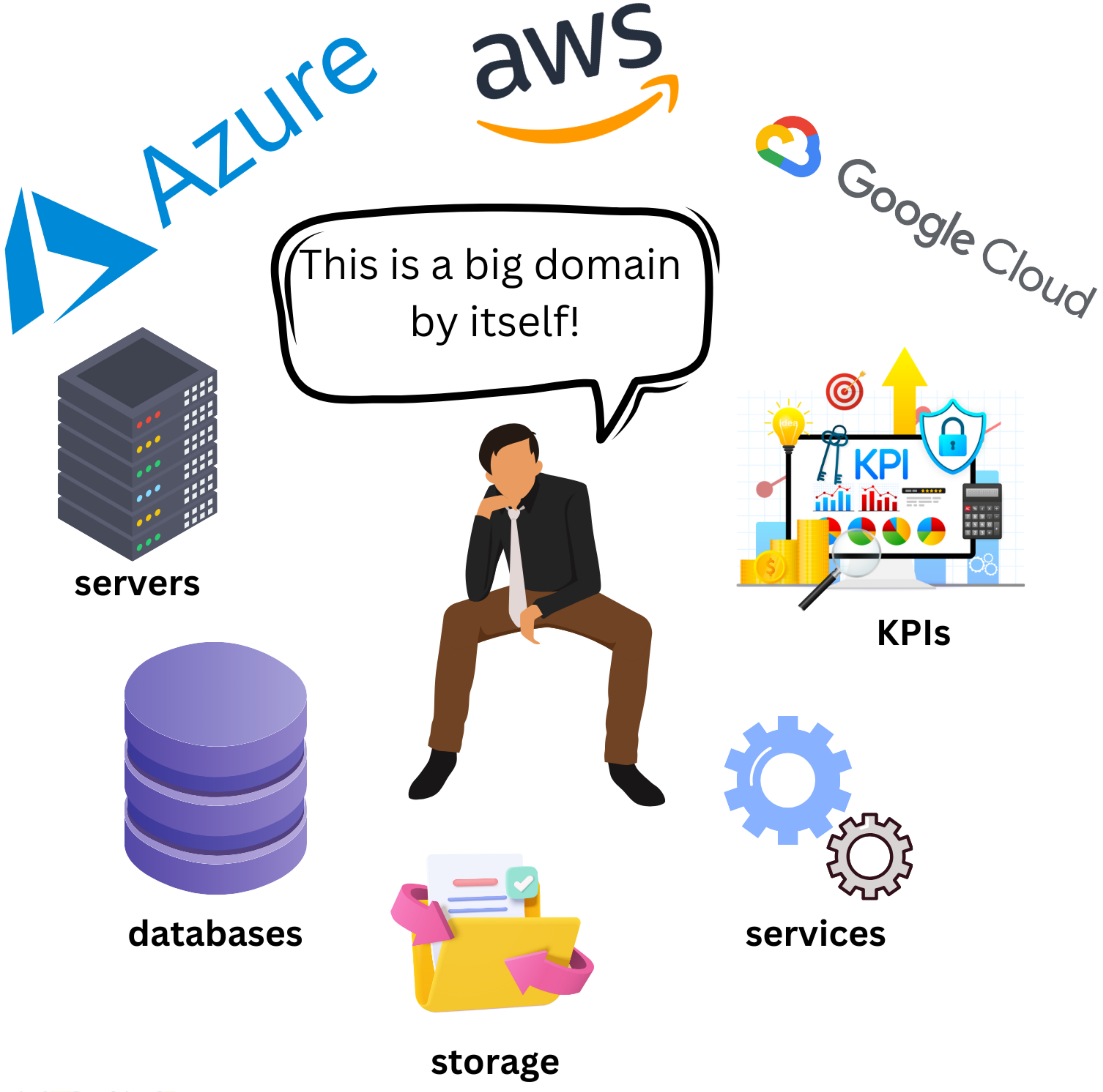
Adapt

Trigger change

Human in the loop or use
automated checks

R.Donakanti, P.Jain, S.Kulkarni, K.Vaidhyanathan, **Reimagining Self-adaptation in the age of Large Language Models,** ICSA 2024, Poster

Making it more Concrete - CloudOps domain

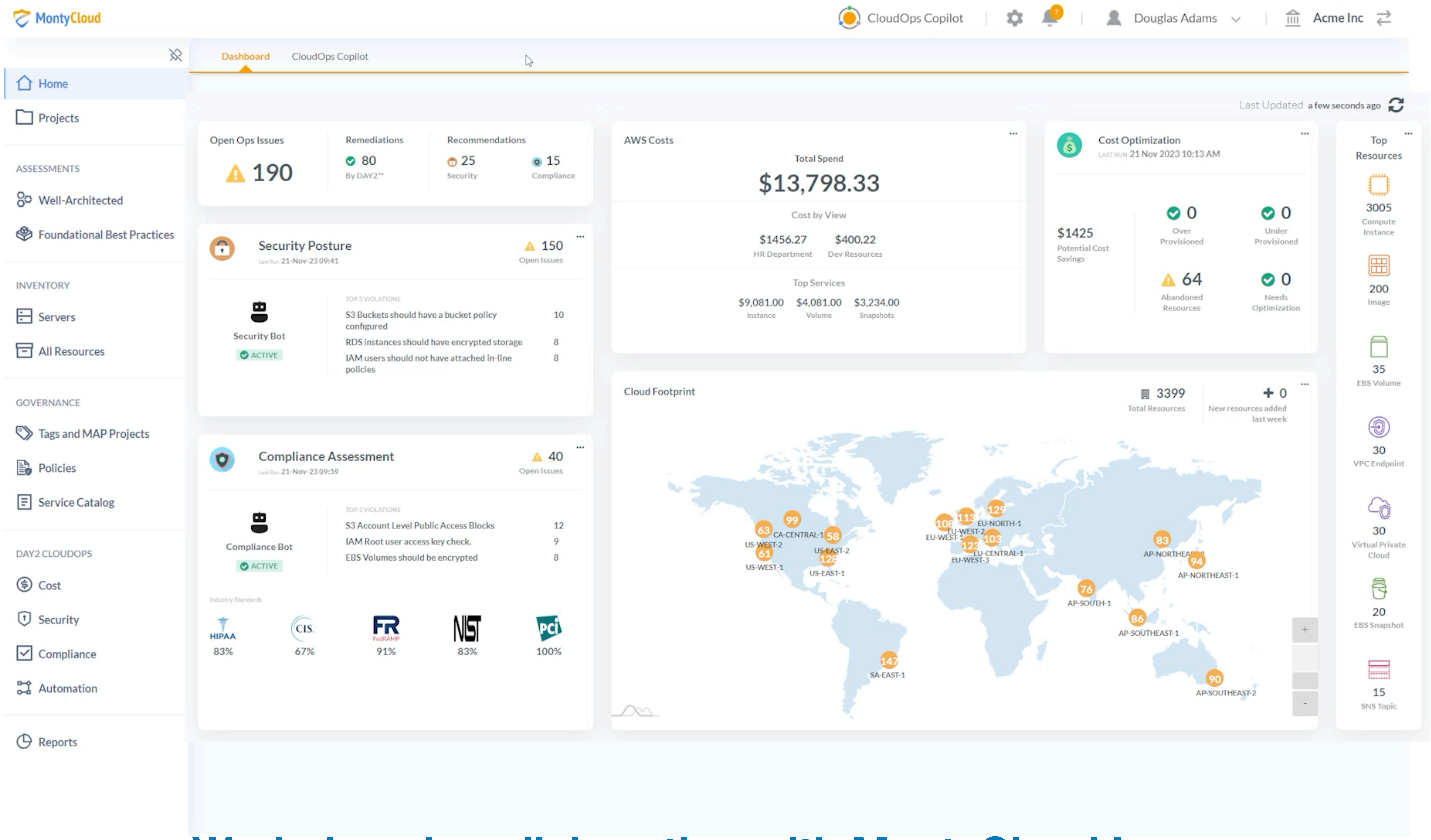


AWS Well Architected Framework

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



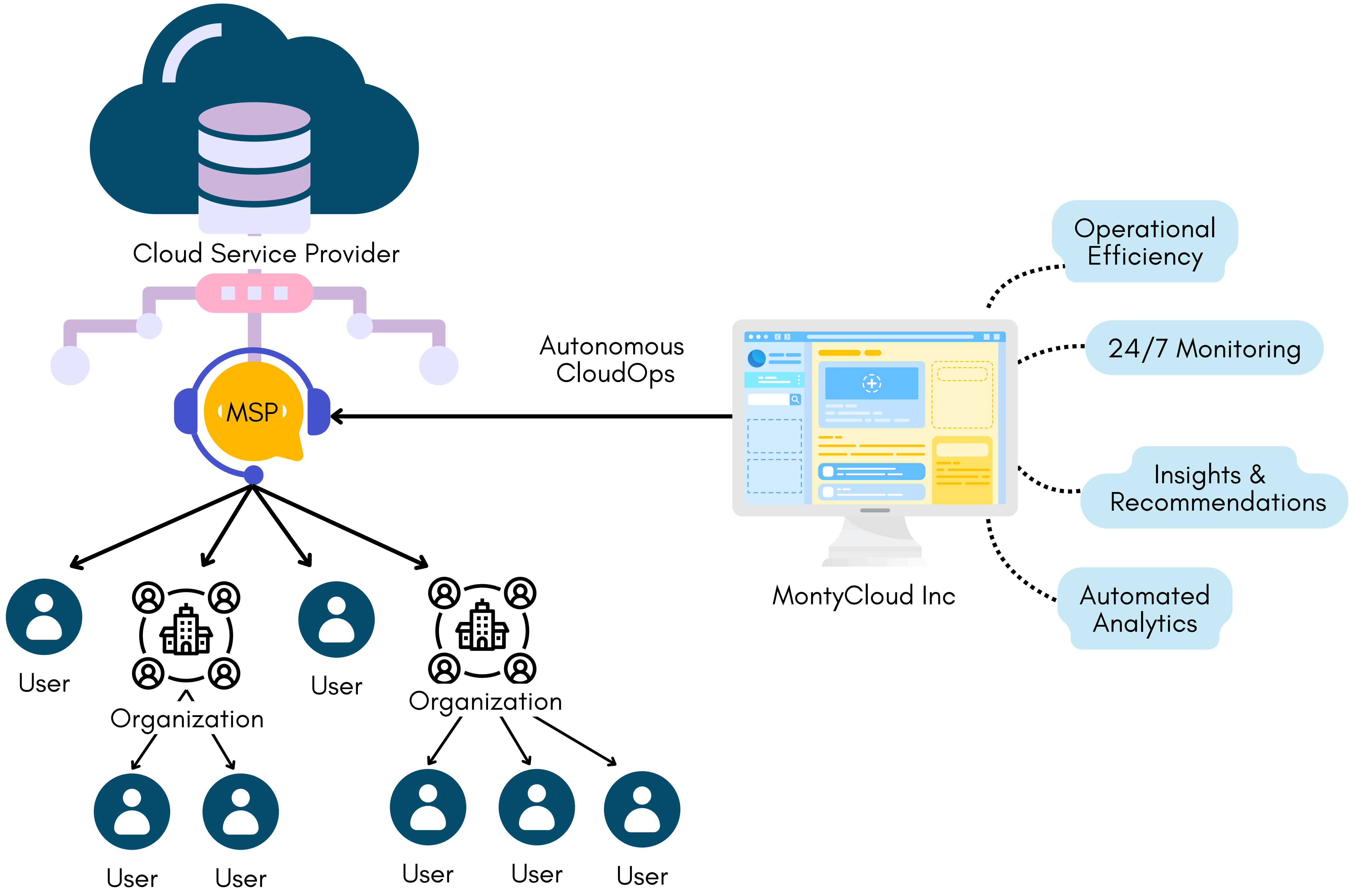
S3 Buckets should have a bucket policy configured	10
RDS instances should have encrypted storage	8
IAM users should not have attached in-line policies	8

S3 Account Level Public Access Blocks	12
IAM Root user access key check.	9
EBS Volumes should be encrypted	8

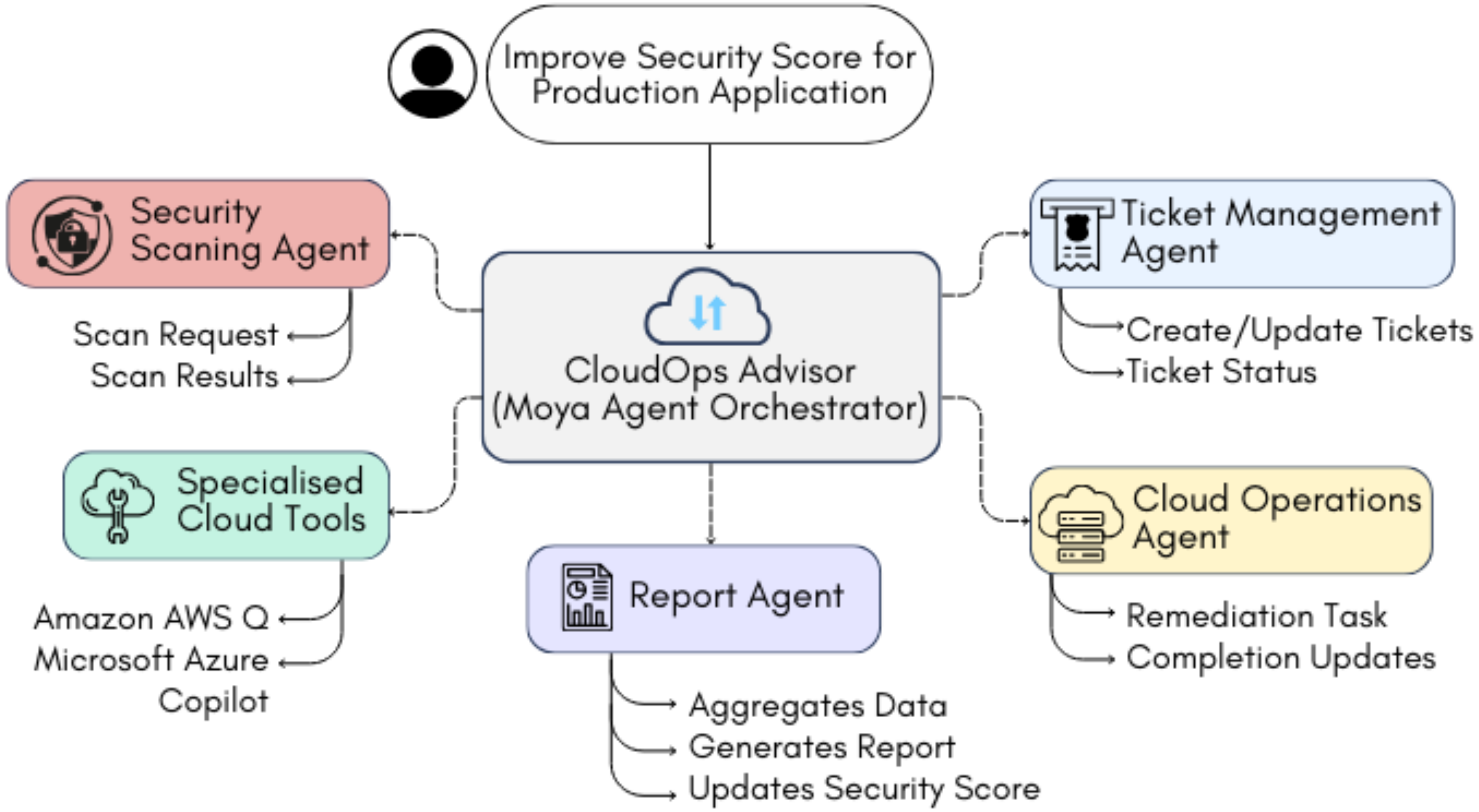
Work done in collaboration with MontyCloud Inc.

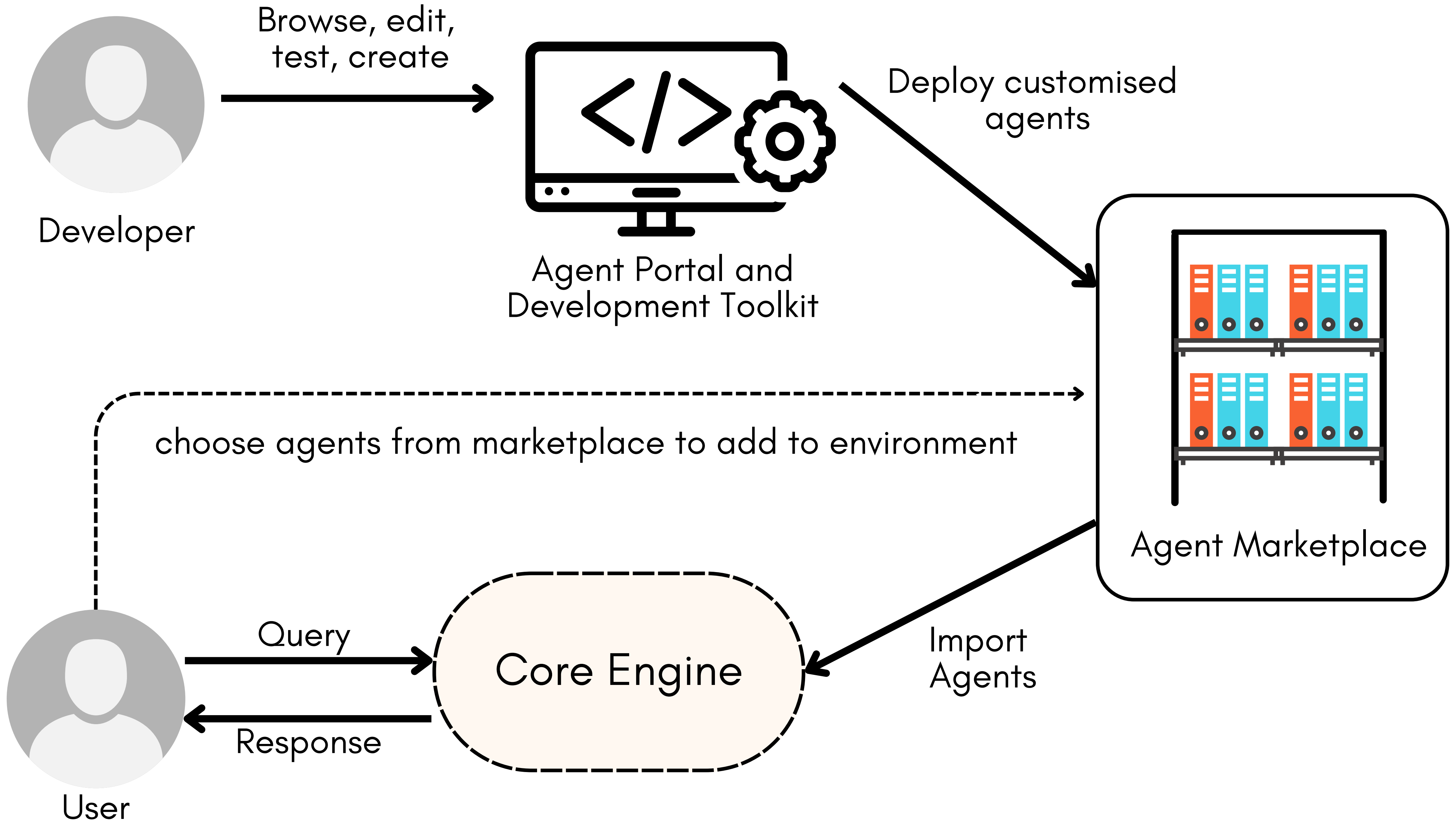
Engineering Challenges (SE for GenAI)

- **Managing Distributed Data**
  - Diverse data sources
- **Maintainability**
  - Large code base, time for updates
- **Extensibility and Modularity**
  - Single vendor, ease of extensions!



Can we go Multi-agent?

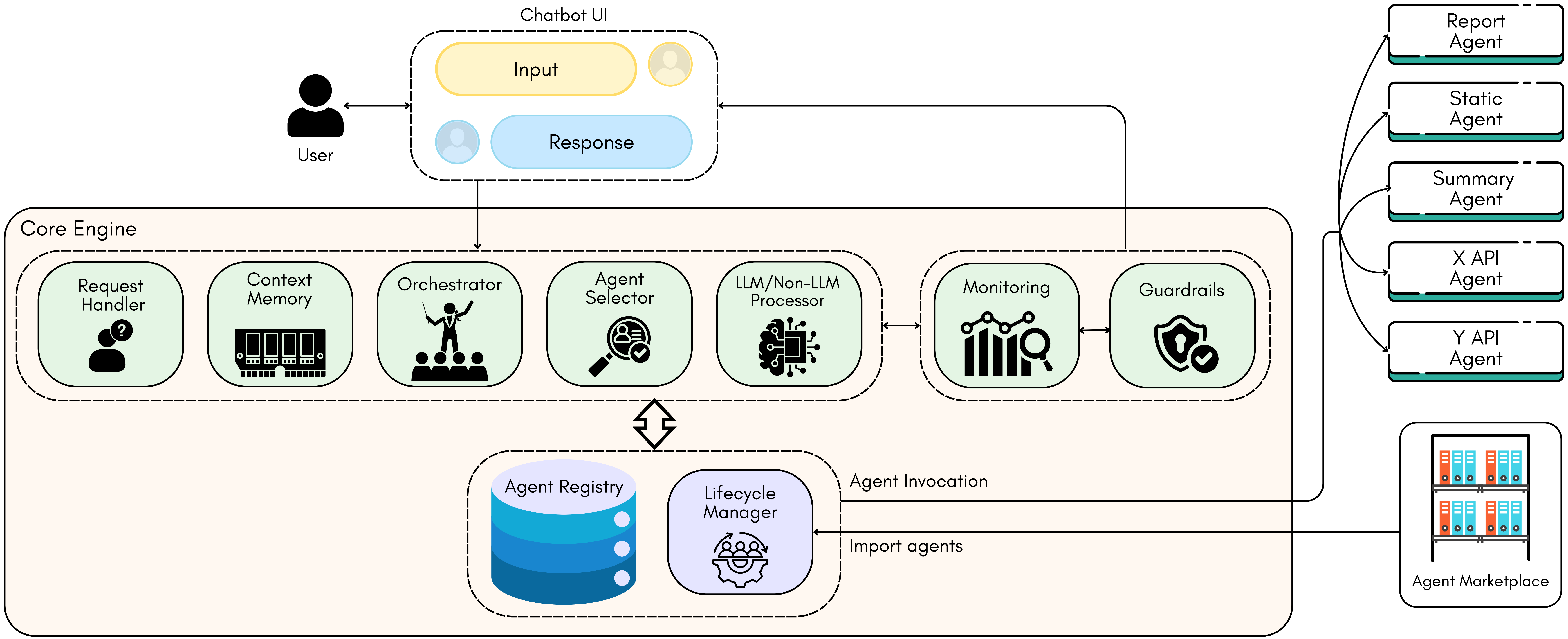



Developer
Browse, edit, test, create
Agent Portal and Development Toolkit
Deploy customised agents
Agent Marketplace
choose agents from marketplace to add to environment
Import Agents
User
Query
Response
Core Engine

Meta orchestration Framework

The MOYA Multi-agent Framework

MOYA repo



Chatbot UI
Input
Response
User
Core Engine
Request Handler
Context Memory
Orchestrator
Agent Selector
LLM/Non-LLM Processor
Monitoring
Guardrails
Agent Registry
Lifecycle Manager
Agent Invocation
Import agents
Report Agent
Static Agent
Summary Agent
X API Agent
Y API Agent
Agent Marketplace

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. To Appear

Key Takeaways

LLMs can be a best friend to the architect if used wisely!

- LLM presents a great opportunity for effective AKM!
- We need to have an effort to have qualitative data on architecture - **ArchBench is a step!**
- Domain specific LLMs which are smaller shall be the way forward - collection of SLMs!
- Need for better ways to architect/engineer systems around LLMs (SA4LLM-enabled Systems)
- LLMs are not here to replace but to support!

SERC Software Engineering Research Centre

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

The Future is here

- **Context is needed** - Capturing organizational aspects needs work, code can help! (Ongoing)
- **LLMs will hallucinate** - No stopping that but we can reduce it - better engineering!
- **Multiple agents collaborating** together to help architects
- **Large action models (LAMs)** for self-adaptation, task generation
- Lot of potential for support in **architecture migration**
- **SE Process will also need upgrades!!**

SA4S@SERC

Rudra Dhar

Akhila Matathammal

Hiya Bhatt

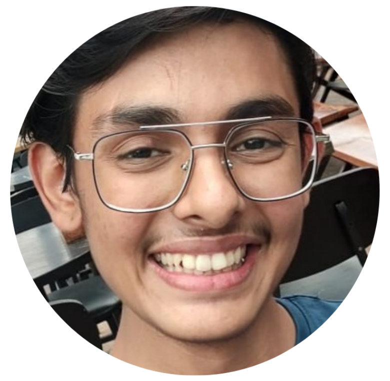
Chandrasekar S

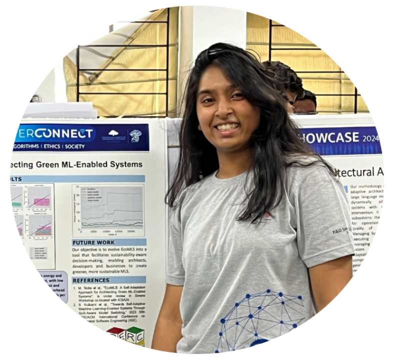
Shubham Kulkarni

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

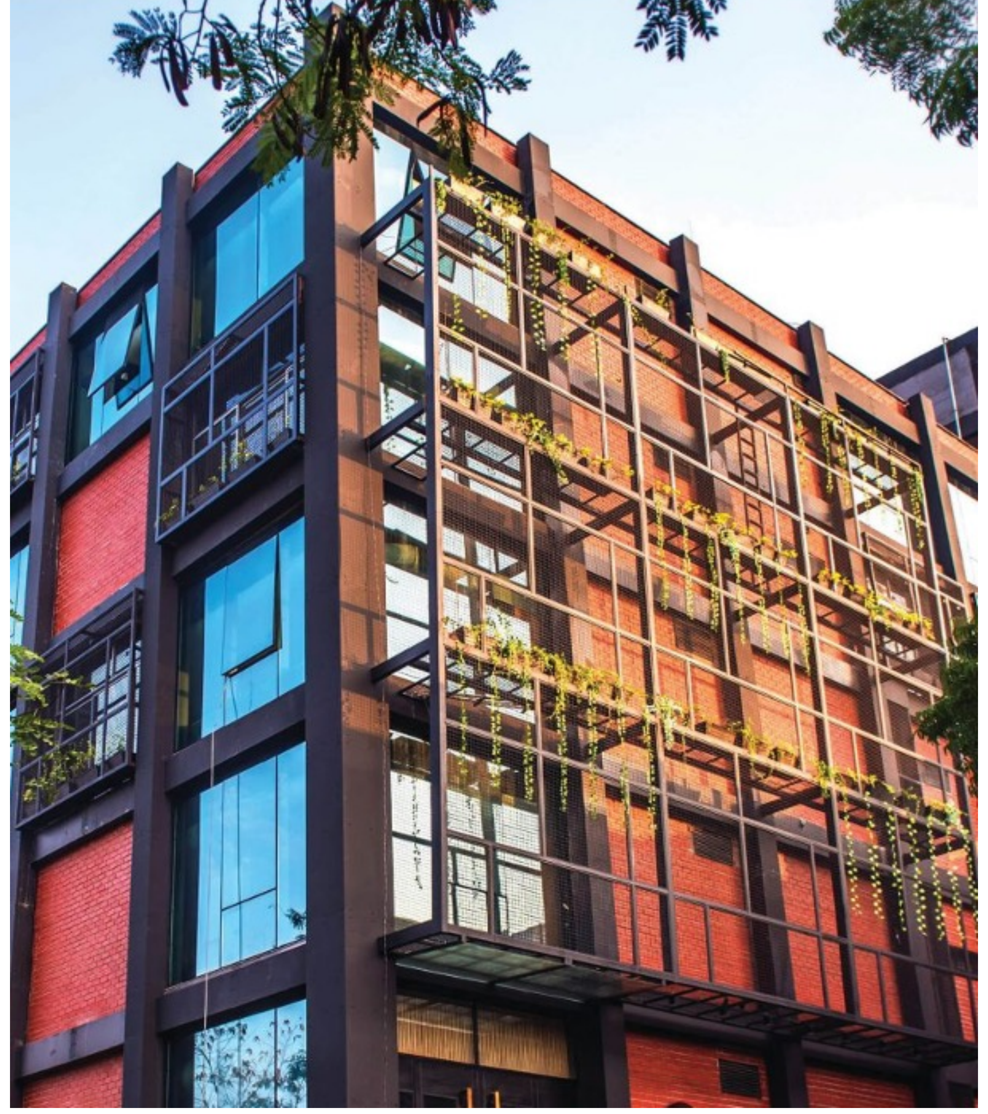
Sreemaee Akshathala

Divyansh Pandey

Maddireddy Kritin

Santosh Kotekal

Vyakhya Gupta



https://serc.iiit.ac.in

Team SA4S

https://sa4s-serc.github.io

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

INTERNATIONAL ARCHITECTURE
4th International Workshop on Software Architecture and Machine Learning
Call for Papers
ODENSE
HANS CHRISTIAN ANDERS

https://sa-ml.github.io/saml2025/

@ ICSA 2025, SAGAI 2025

ODENSE•DENMARK
ICSA 2025
31 MARCH - 4 APRIL

SCAN ME

Thank you

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

IEEE Software Magazine

SE

SE Radio Podcasts

SERC
Software Engineering Research Centre

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD