

# AssetOpsBench:

## Benchmarking AI Agents for Task Automation in Industrial Asset Operation and Maintenance

Dr. Dhaval Patel  
IBM Research  
<https://github.com/IBM/AssetOpsBench>



# Rise of Enterprise Benchmark

<https://www.kaggle.com/benchmarks>

Special Thanks to Kaggle and IBM Research Team

The screenshot shows the Kaggle Benchmarks page. On the left, there's a sidebar with navigation links: Home, Competitions, Datasets, Models, **Benchmarks** (which is highlighted), Game Arena, Code, Discussions, Learn, More, Your Work, VIEWED, Enterprise Ops, FailureSensorIQ, FailureSensorIQ, ITBench, The 4th RePSS, and EDITED.

The main content area has a search bar at the top. Below it, a section titled "Benchmarks" says: "Discover open, rigorous benchmarks and model leaderboards from top AI labs and researchers in one place. Learn more in the [Documentation](#)". There's another search bar labeled "Search Benchmarks". Below that are filters for "All Benchmarks", "Type", "Task", and "Creator".

A large callout box highlights the "Featured Benchmarks" section. It contains two cards: "Enterprise Operations (EntOps) Bench" (curated by IBM Research) and "SimpleQA Verified" (curated by Google DeepMind). Both cards show current top 3 results with bar charts.

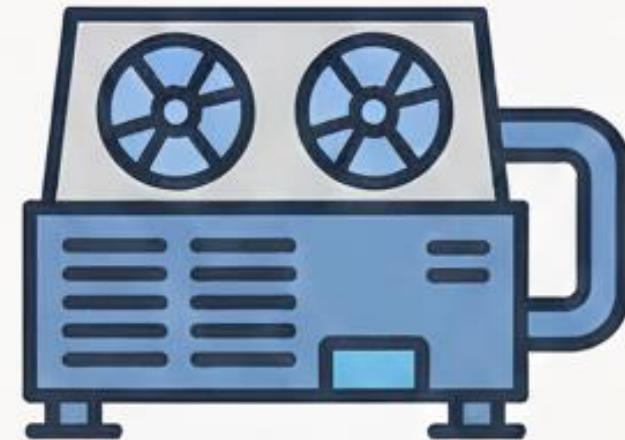
Other cards shown include "ICML 2025 Experts" (curated by Kaggle), "Chess" (curated by Kaggle), and "CURRENT TOP 3 (OF 11)" for various benchmarks.

A cartoon illustration of a person sitting at a desk, looking at a computer screen displaying a bar chart. There's a potted plant on the desk next to them.

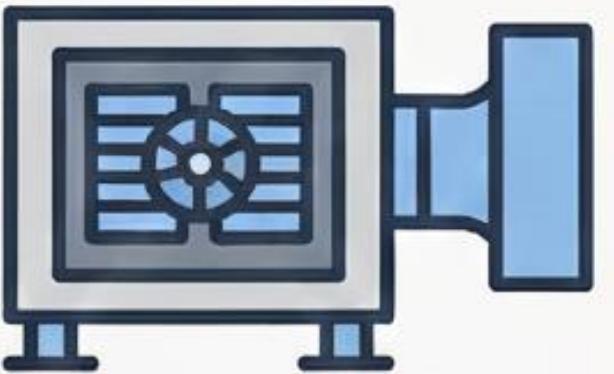
# Industrial Assets

We are all surrounded by industrial assets - often without noticing them. These systems form the backbone of critical infrastructure and everyday services:

- **Data Centers:** Chillers, air-handling units, standby generators
- **Hospitals:** Standby generators ensuring uninterrupted operations
- **Energy Generation:** Wind turbines powering renewable electricity
- ....



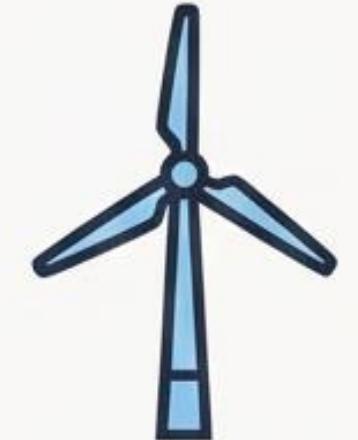
Chiller



Air-Handing Unit



Standby Generator



Wind Turbine



Transportation



Air Production Unit

# Monitoring and Maintenance Tasks

With the emergence of Industry 4.0 applications, these assets are now being monitored using 100s of sensors in real time:

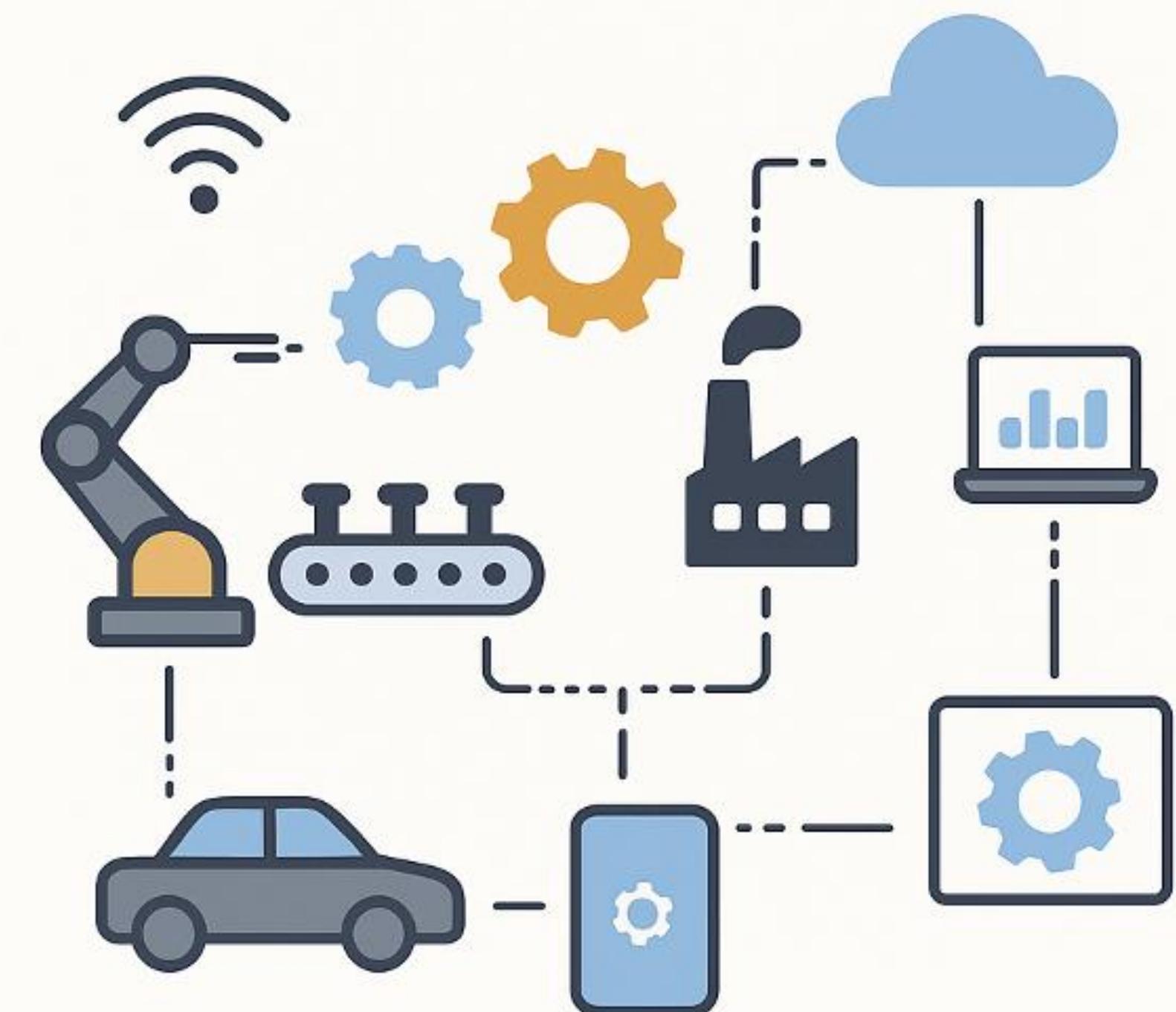
- Temperature,
- Pressure,
- Vibration,
- Power,
- Flow,
- Acoustic,
- Airflow,
- RPM,
- Thermal,
- Gas,
- Visual



*Is Chiller 11's compressor overheating? Generate a service request if needed.*

# Industry 4.0

The integration of physical machines with digital technologies like IoT, AI, cloud computing, robotics, and digital twins to create intelligent, automated, and interconnected industrial systems

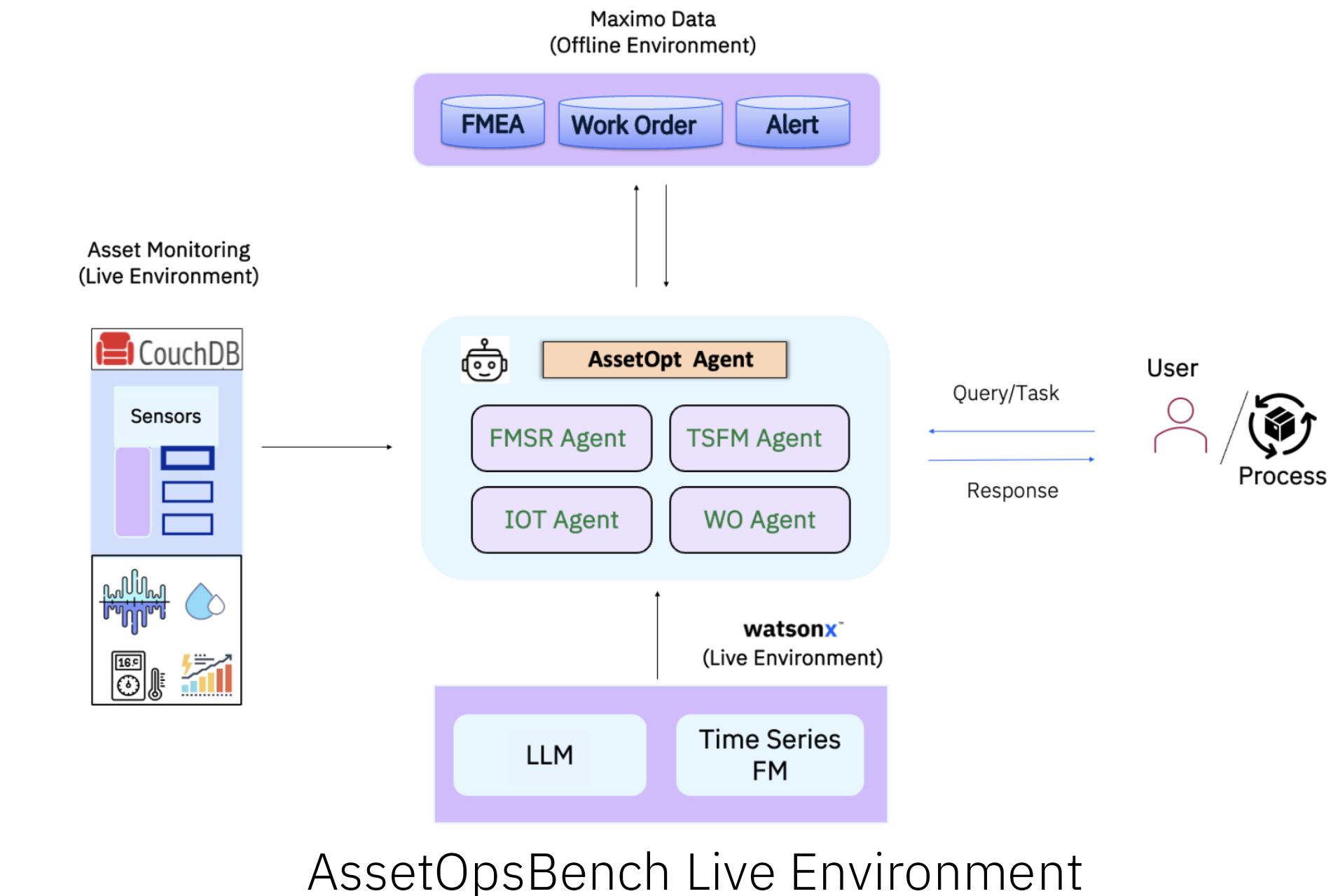


# AssetOpsBench : Open-Source Initiative

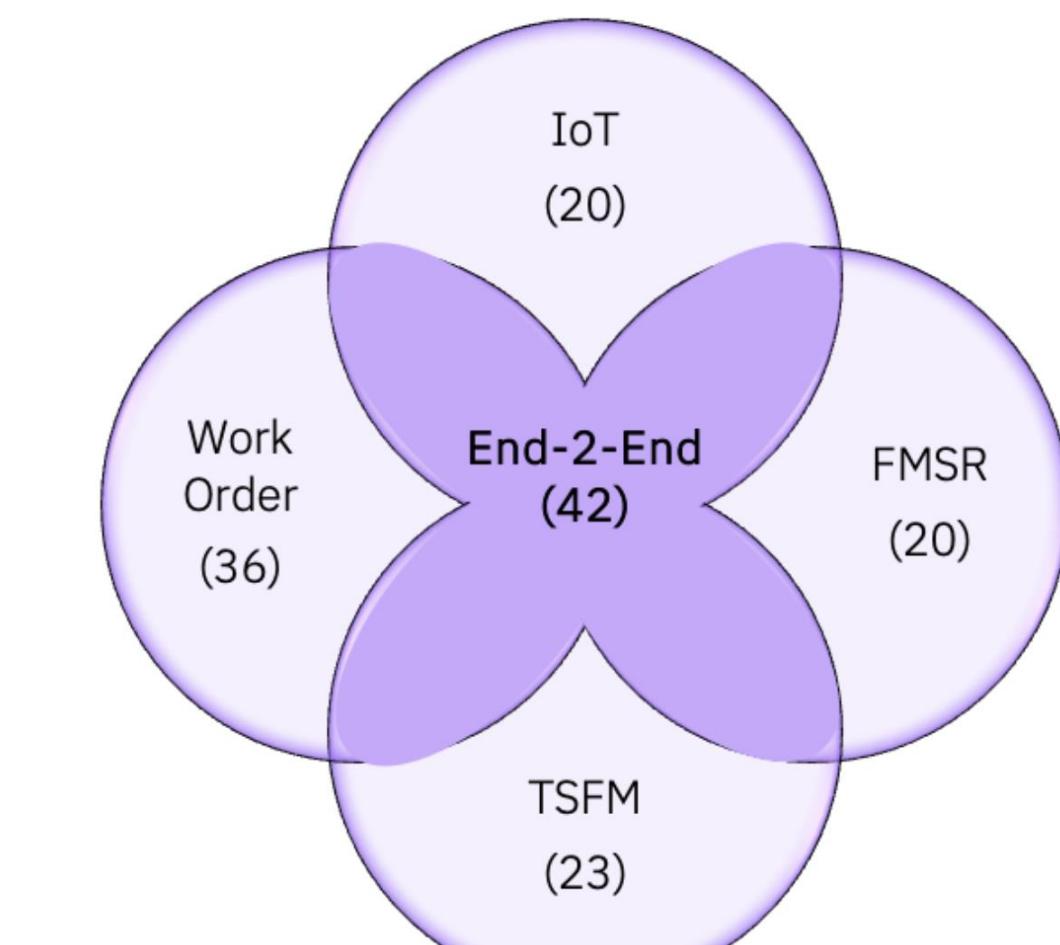


# AssetOpsBench: Open-Source Benchmark for Industry 4.0

- Framework to assess Gen AI solutions' ability to solve Industry 4.0 Automation "**Scenarios**"
- Simulated industrial environment, **9 multi-source data sets** (work orders, FMEAs, timeseries) and **4 domain-specific agents** (IoT, data science, work order, failure mode to sensor mapping)
- **140+** expert-authored natural language queries, grounded in **enterprise industrial scenarios**
- Two Multi-Agent Orchestration Recipes
  - **Agent-As-Tool**
  - **Plan-Execute**
- **LLM-as-Judge** for Rubric-based Agent Evaluation and **Reference-based** Scoring for Semantic Evaluation
- **Agent harness:** systematic procedure for automated discovery of emerging failure modes

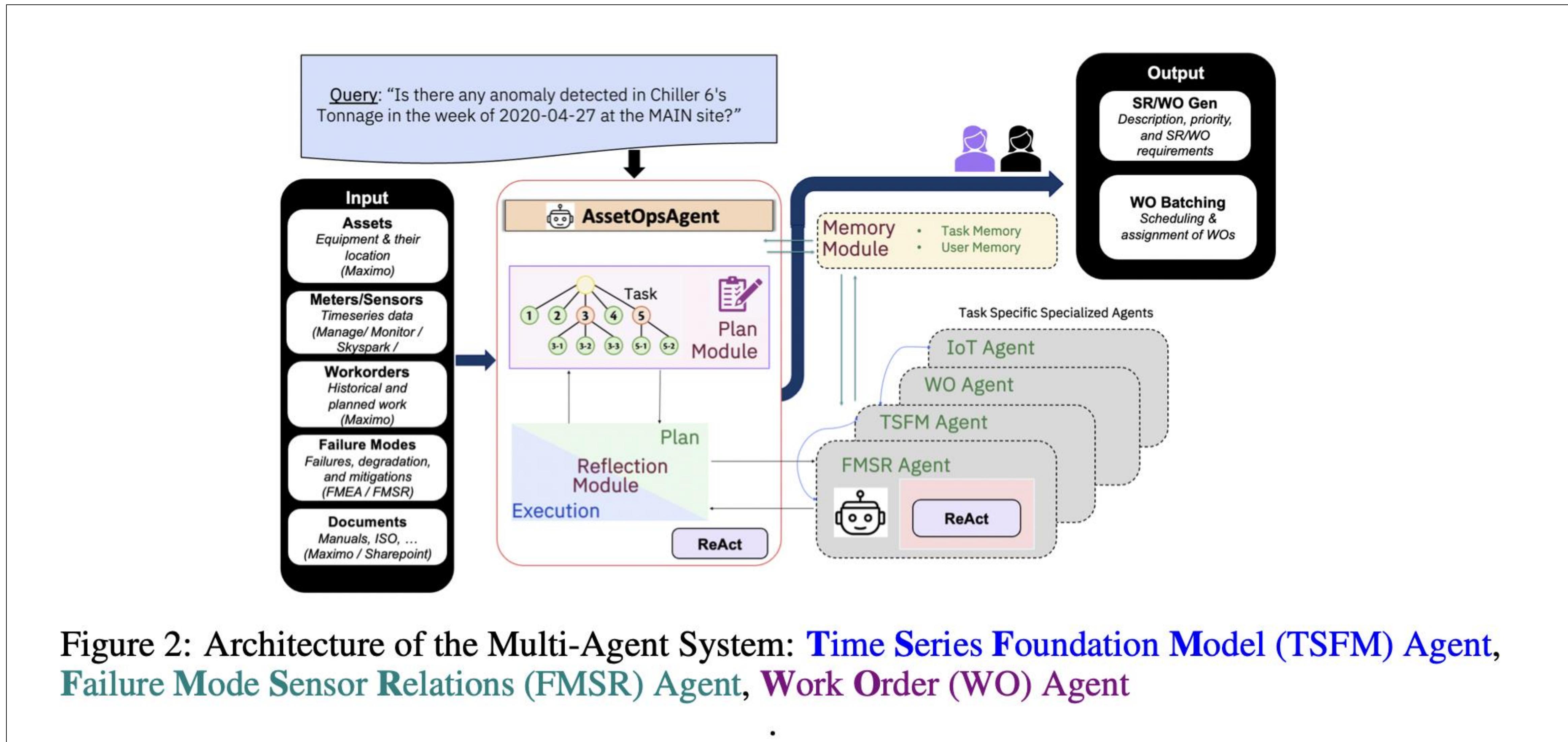


AssetOpsBench Live Environment

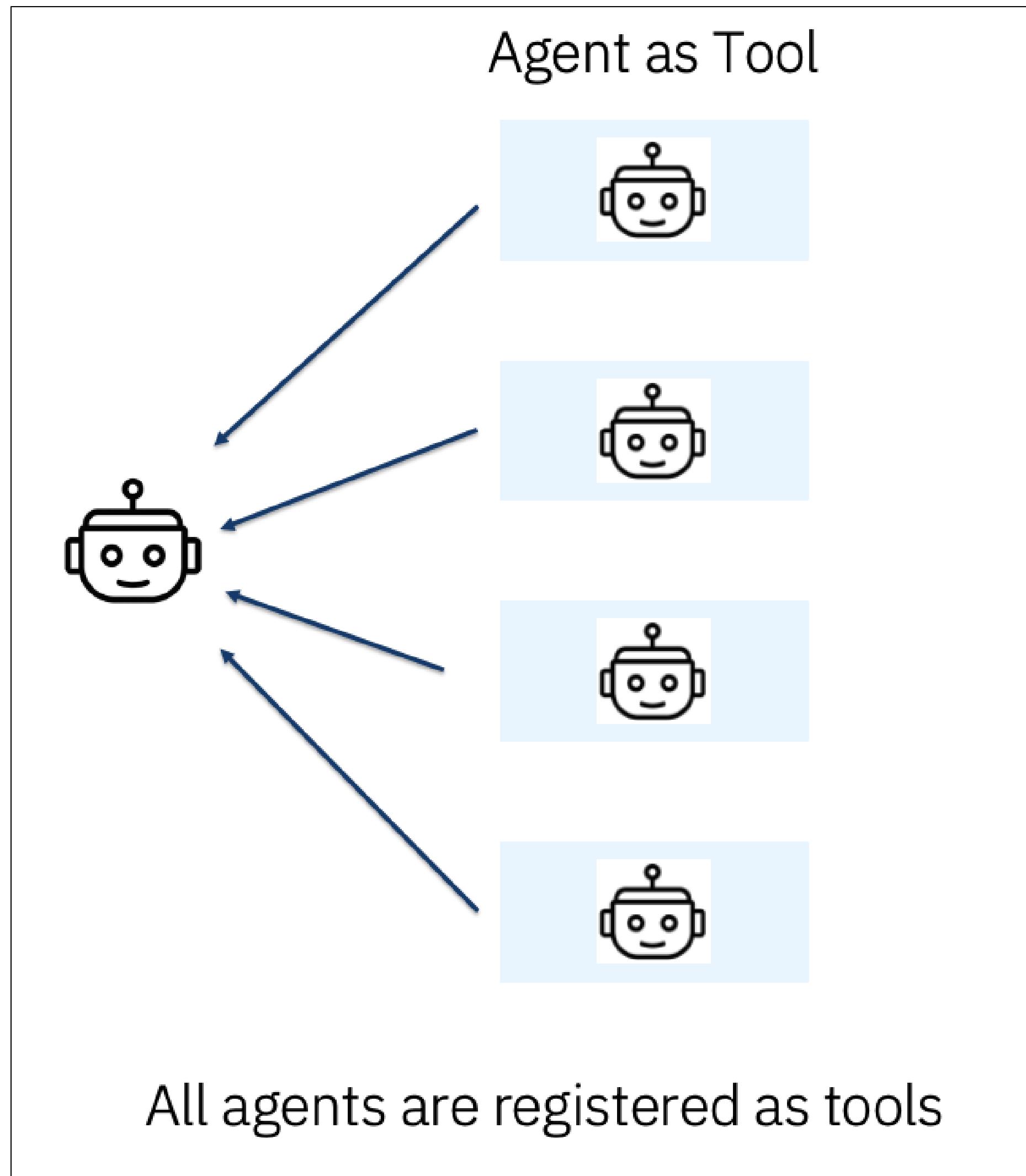


141 Utterance Distribution

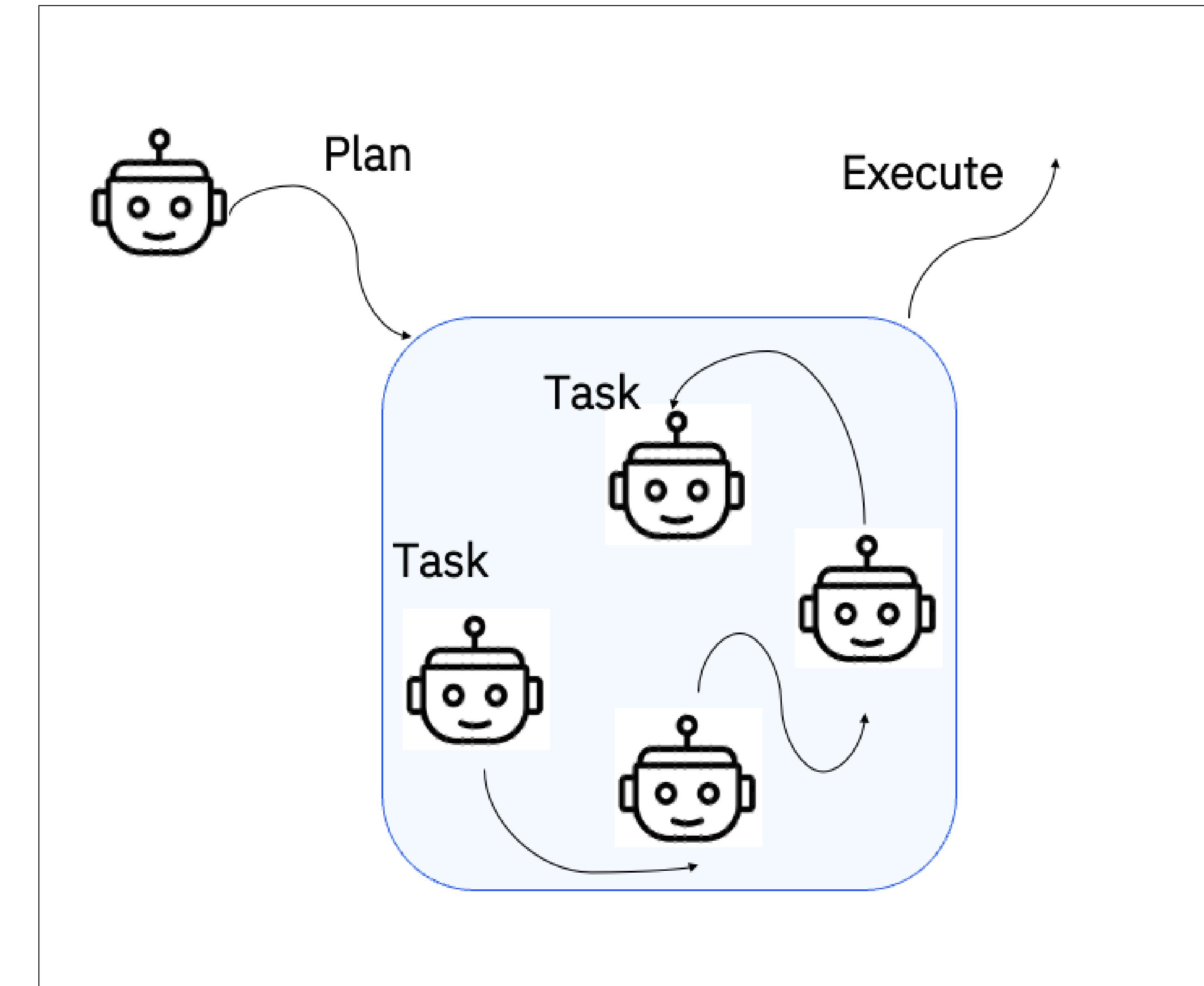
# AssetOpsBench : A Multi-Agent System (MAS) is at the core



# AssetOpsBench : Multi-Agent Implementation Strategy



Agent-As-Tool Approach



Plan-Execute Approach

# AssetOpsBench : Evaluation



# Automatic Evaluation of Agentic Workflow

## Ground Truth Preparation

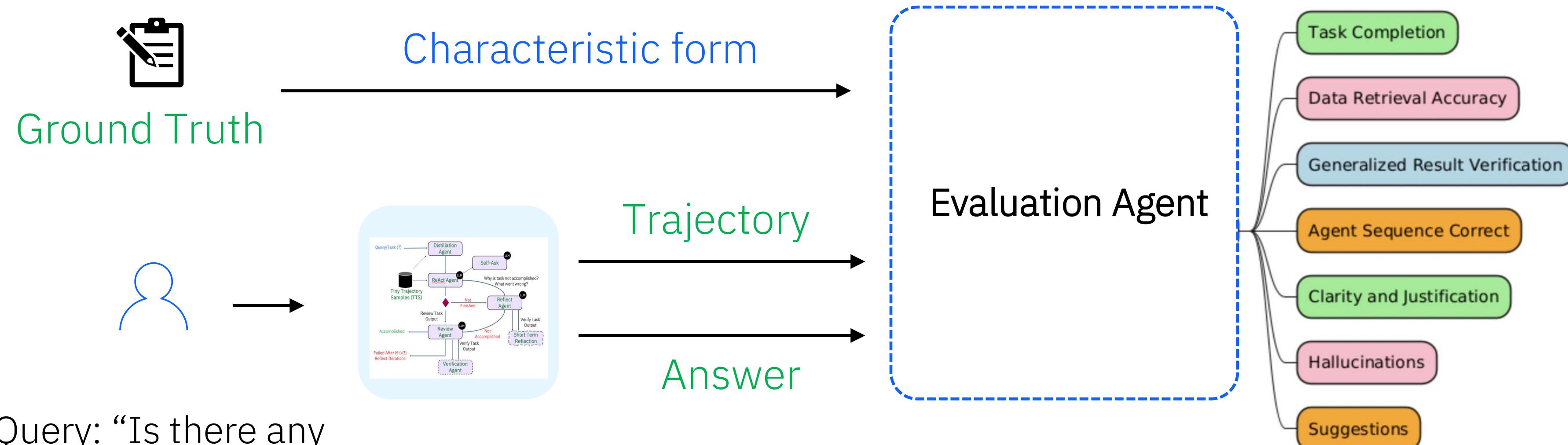
- **Characteristic form** describe the final output along with the process/step to be conducted. As an example:

“The expected response should confirm the successful execution of all required actions while ensuring the correct variables, including the asset (Chiller 6), location (POKMAIN), and time range (week of 2020-04-27), were used for data retrieval and analysis. It should specify that [IoTAgent](#) was called to request and download the data, and [TSFMAgent](#) was properly utilized to perform Time Series anomaly detection on the Tonnage parameter. The response must also verify that the data was accurately stored in the designated file location, and that the analysis results were saved to a new file. Additionally, the response should explicitly confirm the detection of anomalies in Chiller 6's Tonnage during the specified timeframe at the POKMAIN site, as these anomalies were anticipated.”

This characteristic form serves as the ground truth for evaluating responses using a **rubric-based performance assessment**.

# Automatic Evaluation of Agentic Workflow

## Evaluation Agent: Rubric-based Performance Assessment

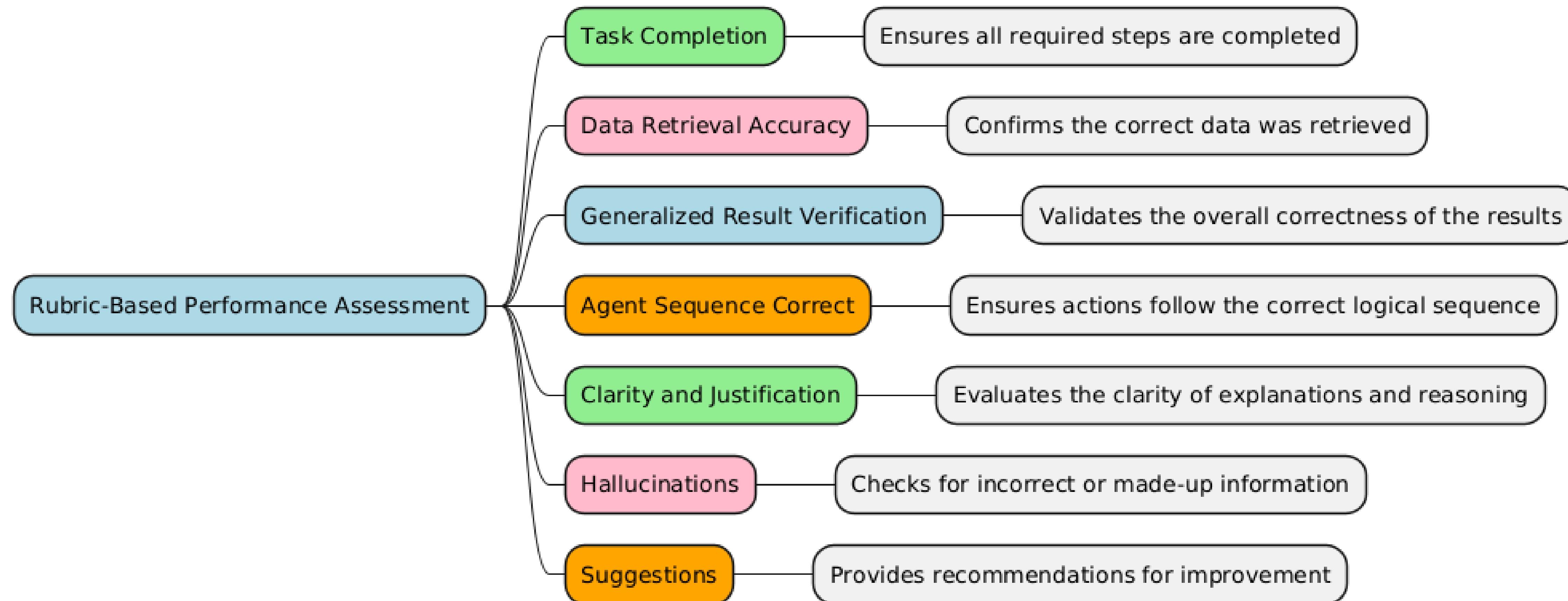


Query: “Is there any anomaly detected in Chiller 6's Tonnage in the week of 2020-04-27 at the POKMAIN site?”

# Automatic Evaluation of Agentic Workflow

## Evaluation Agent: Rubric-based Performance Assessment

We captured performance across 6 metrics to better understand the failure mode of agents.



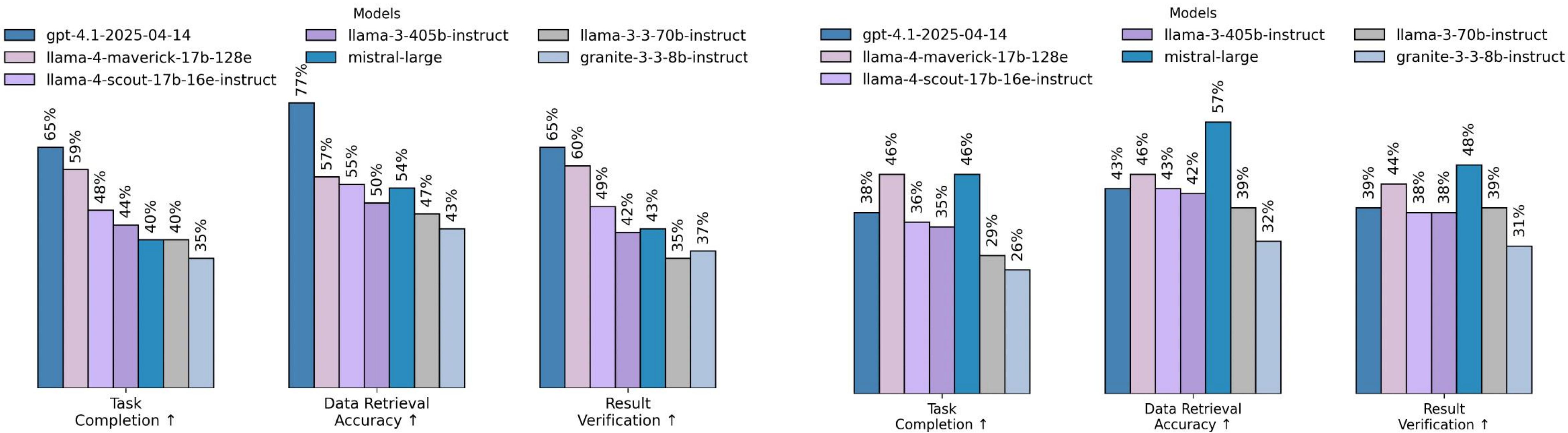
# AssetOpsBench : Summary of Result



# Leaderboard

We conducted an extensive benchmark to compare the two orchestration recipes:

- Agent-As-Tool consistently outperforms Plan-Execute



((a)) **Agent-As-Tool Approach**

((b)) **Plan-Execute Approach**

# Emergent Failure Mode Discovery

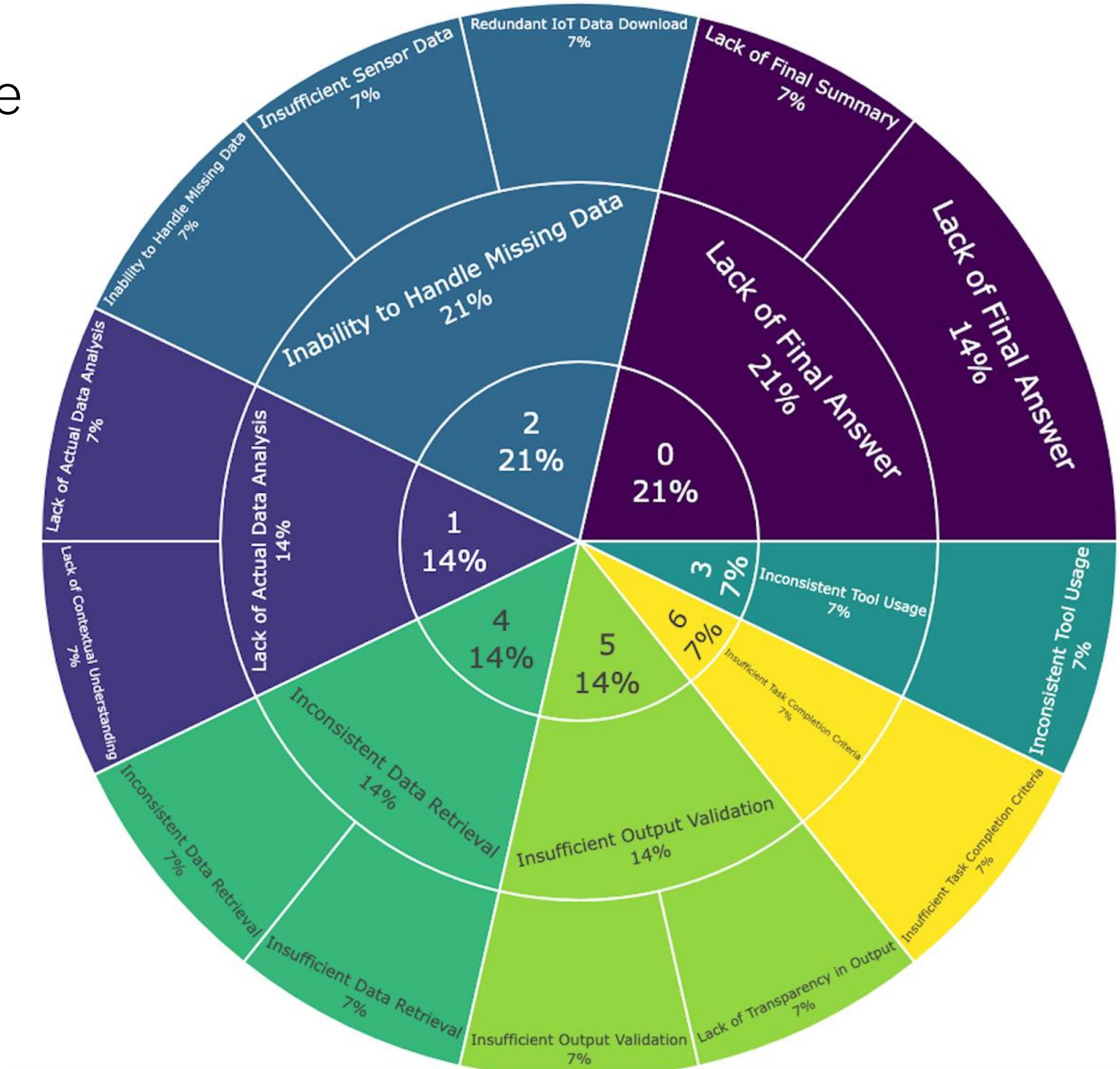
We have an automated way of trajectory introspection to discover novel, emerging failure modes automatically

## Agent Coordination (Total 27.52%)

Conversation Reset	Execution: 0.00%
Fail to Ask for Clarification	Execution: 10.22%
Task Derailment	Execution: 4.34%
Information Withholding	Execution: 2.22%
Ignored Agent's Input	Execution: 2.06%
Action Mismatch	Execution: 8.68%

We added a new method called “Self-Ask” to demonstrate the performance improvement

Model	enable_agent_ask=True	enable_agent_ask=False
gpt-4.1-2025-04-14	63%	65%
lama-4-maverick	66%	59%
llama-3-405b-instruct	61%	44%
mistral-large	58%	40%
llama-3-3-70b-instruct	35%	40%
granite-3-3-8b-instruct	32%	35%



# AssetOpsBench Extensive Research: FailureSensorIQ

- FailureSensorIQ introduces a **dataset** and **benchmark** that tests whether LLMs can reason about sensors, assets, and failure modes beyond data-driven correlations. It benchmarks *sensor-failure relationships*, which is the primary capability targeted by the Failure Mode Sensor Relation (FMSR) agent in AssetOpBench.
- FailureSensorIQ is accepted in NeurIPS 2025.

The screenshot shows the HuggingFace Dataset page for FailureSensorIQ. At the top, it displays basic dataset statistics: Tasks: Question Answering, Modalities: Text, Formats: json, Languages: English, Size: 1K - 10K, and ArXiv: (link). Below this are sections for Libraries (Datasets, pandas, Croissant), License (apache-2.0), and navigation links (Dataset card, Data Studio, Files and versions, xet, Community).

The screenshot shows the Dataset Viewer for the FailureSensorIQ dataset. It displays a table with columns: subject, id, question, options, and option\_ids. The table has two rows under the subset "multi\_true\_multi\_choice\_qa". The first row corresponds to the "train" split, showing 5.63k rows. The second row corresponds to the "test" split, showing 5.63k rows. The "options" column contains lists of strings, and the "option\_ids" column contains lists of integers.

subject	id	question	options	option_ids
failure_mode_sensor_analysis	1	For electric motor, if a...	[ "oil debris", ... ]	[ "A", "B", "C", "D", "E" ... ]
failure_mode_sensor_analysis	2	For electric motor, if a...	[ "resistance", ... ]	[ "A", "B", "C", "D", "E" ... ]
failure_mode_sensor_analysis	3	For electric motor, if a...	[ "coast down time", ... ]	[ "A", "B", "C", "D", "E" ... ]
failure_mode_sensor_analysis	4	For electric motor, if a...	[ "partial discharge", ... ]	[ "A", "B", "C", "D", "E" ... ]
failure_mode_sensor_analysis	5	For electric motor, if a...	[ "temperature", ... ]	[ "A", "B", "C", "D", "E" ... ]

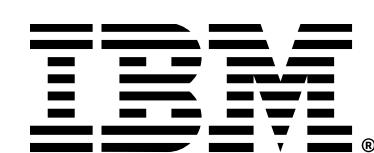
HuggingFace Dataset

The screenshot shows the Kaggle Benchmark page for FailureSensorIQ. At the top, it displays basic benchmark statistics: Leaderboard (0), Discussion (0), and a brief description: "FailureSensorIQ is a novel Multi-Choice Question-Answering benchmarking system designed to assess the ability of Large Language Models to reason and understand complex, domain-specific scenarios in Industry 4.0. Unlike traditional QA benchmarks, our system focuses on multiple aspects of reasoning through failure modes, sensor data, and the relationships between them across various industrial assets."

The screenshot shows the Leaderboard for the FailureSensorIQ benchmark. It lists eight models with their scores and other metrics. The models are: Gemini-3-Pro-Preview, O3-2025-04-16, Gpt-5-2025-08-07, Gemini-2.5-Pro, Gemini-2.5-Flash, Gpt-5-Mini-2025-08-07, O4-Mini-2025-04-16, and Grok-4.1-Fast-Reasoning. The table includes columns for Model, Score, Consistency, F-Score, and Elimination Accuracy.

#	Model	Score	Consistency	F-Score	Elimination Accuracy
1	Gemini-3-Pro-Preview	69.1%	63.8%	69.1%	70.8%
2	O3-2025-04-16	67.6%	59.1%	67.4%	69.4%
3	Gpt-5-2025-08-07	67.2%	59.2%	67.3%	69.4%
4	Gemini-2.5-Pro	67.0%	57.5%	67.0%	68.8%
5	Gemini-2.5-Flash	65.5%	56.1%	65.8%	68.3%
6	Gpt-5-Mini-2025-08-07	65.3%	56.8%	65.5%	68.2%
7	O4-Mini-2025-04-16	64.8%	56.7%	65.0%	67.4%
8	Grok-4.1-Fast-Reasoning	64.6%	57.9%	64.9%	66.4%

Kaggle Benchmark



## Learn More

### Contact Information:

Dhaval Patel  
[pateldha@us.ibm.com](mailto:pateldha@us.ibm.com)

Shuxin Lin  
[shuxin.lin@ibm.com](mailto:shuxin.lin@ibm.com)

AssetOpsBench  
Github Repo



AssetOpsBench  
HuggingFace  
Dataset



FailureSensorIQ  
HuggingFace  
Dataset



AssetOpsBench  
Arxiv Paper



AssetOpsBench  
Codabench  
Competitions



FailureSensorIQ  
Kaggle  
Benchmark



### Upcoming event announcement:

- Please come to IBM Booth : NeurIPS 2025 (10 AM - Dec 5, 2025)
- We will be presenting AssetOpsbench in AAAI 2026 Lab “*From Inception to Productization: Hands-on Lab for the Lifecycle of Multimodal Agentic AI in Industry 4.0*” on Jan 21st, 2026 in AAAI 2026 @ Singapore Expo.

[Find Us in NeurIPS 2025, San Diego and AAAI 2026, Singapore!](#)