

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>

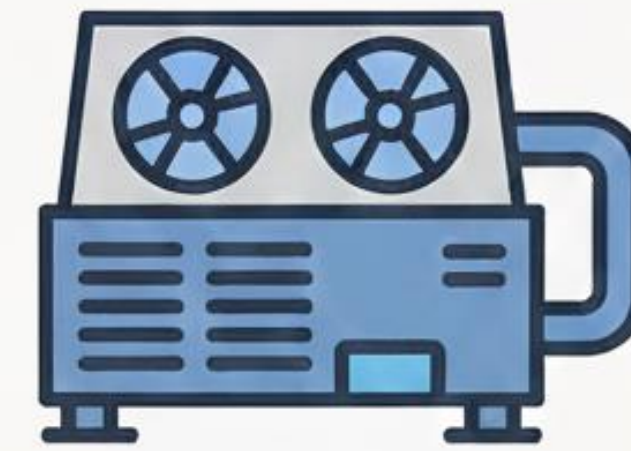
Special Thanks to Kaggle and IBM Research Team

[illegible]

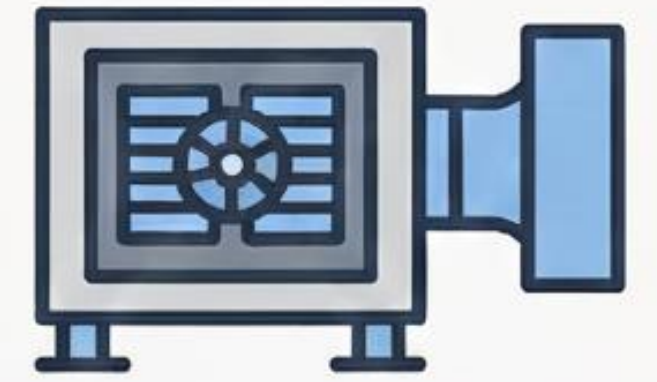
Industrial Assets

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-Handling 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:

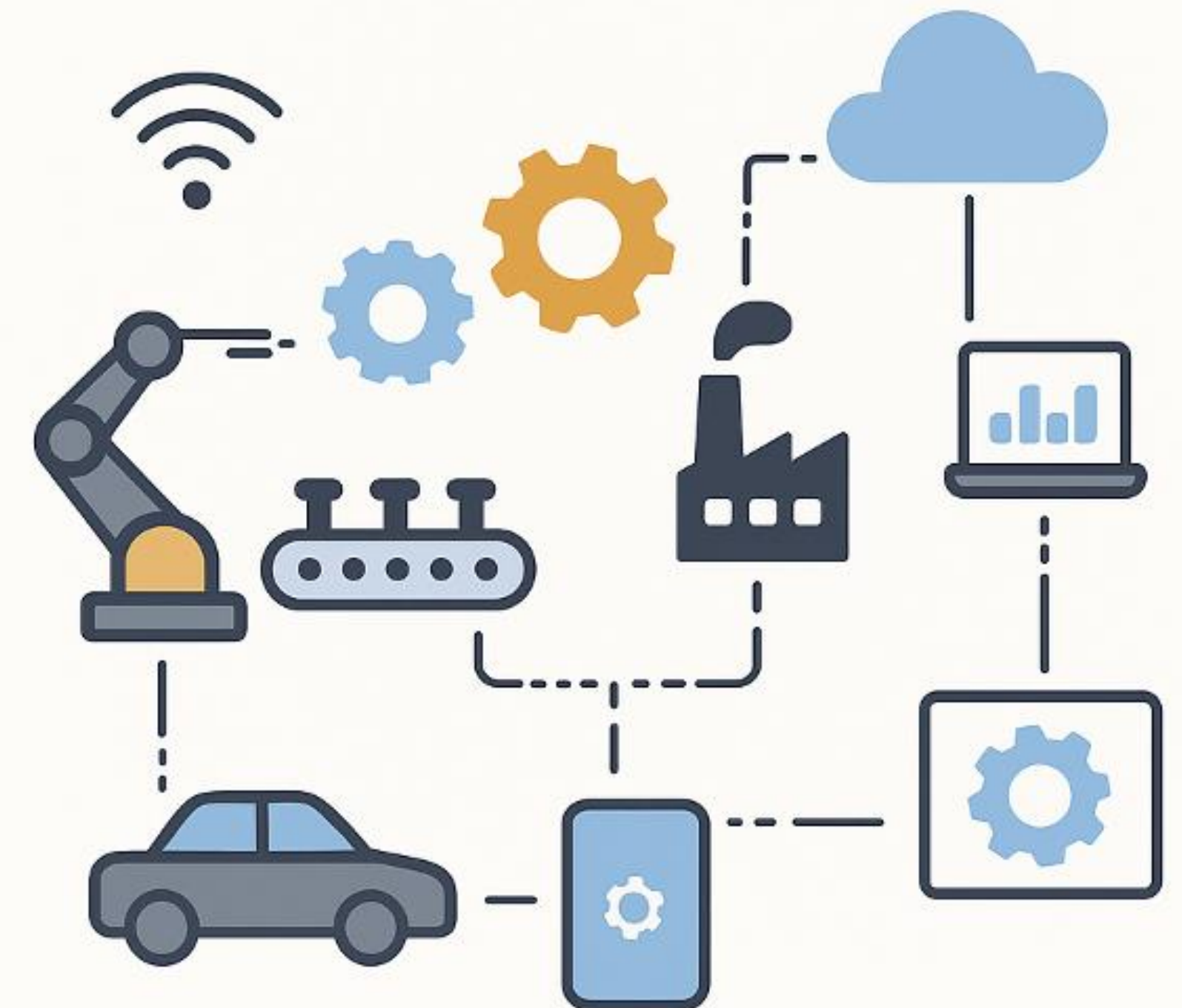
- 🌡️ Temperature,
- 🔵 Pressure,
- 🔊 Vibration,
- ⚡ Power,
- 💧 Flow,
- 🔊 Acoustic,
- 🌬️ Airflow,
- 📊 RPM,
- 🔥 Thermal,
- 🧪 Gas,
- 🔍 Visual



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

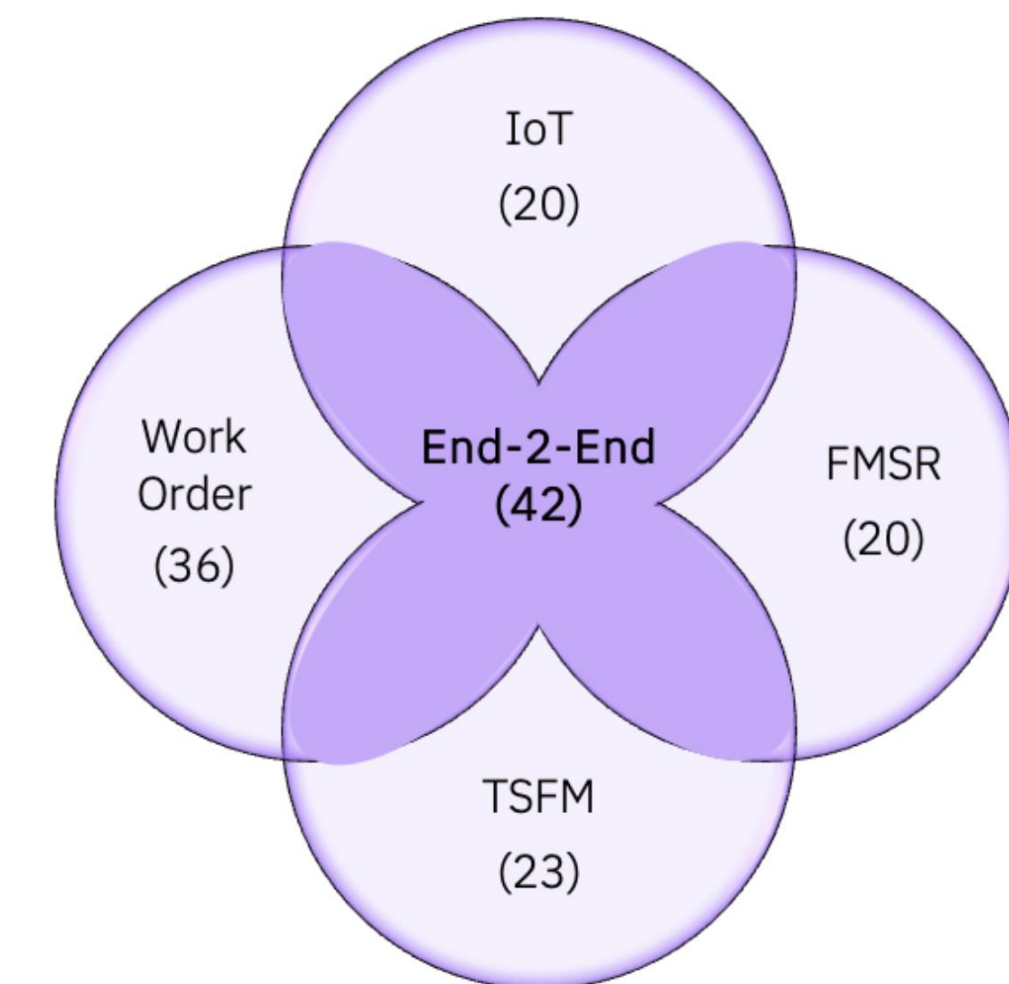
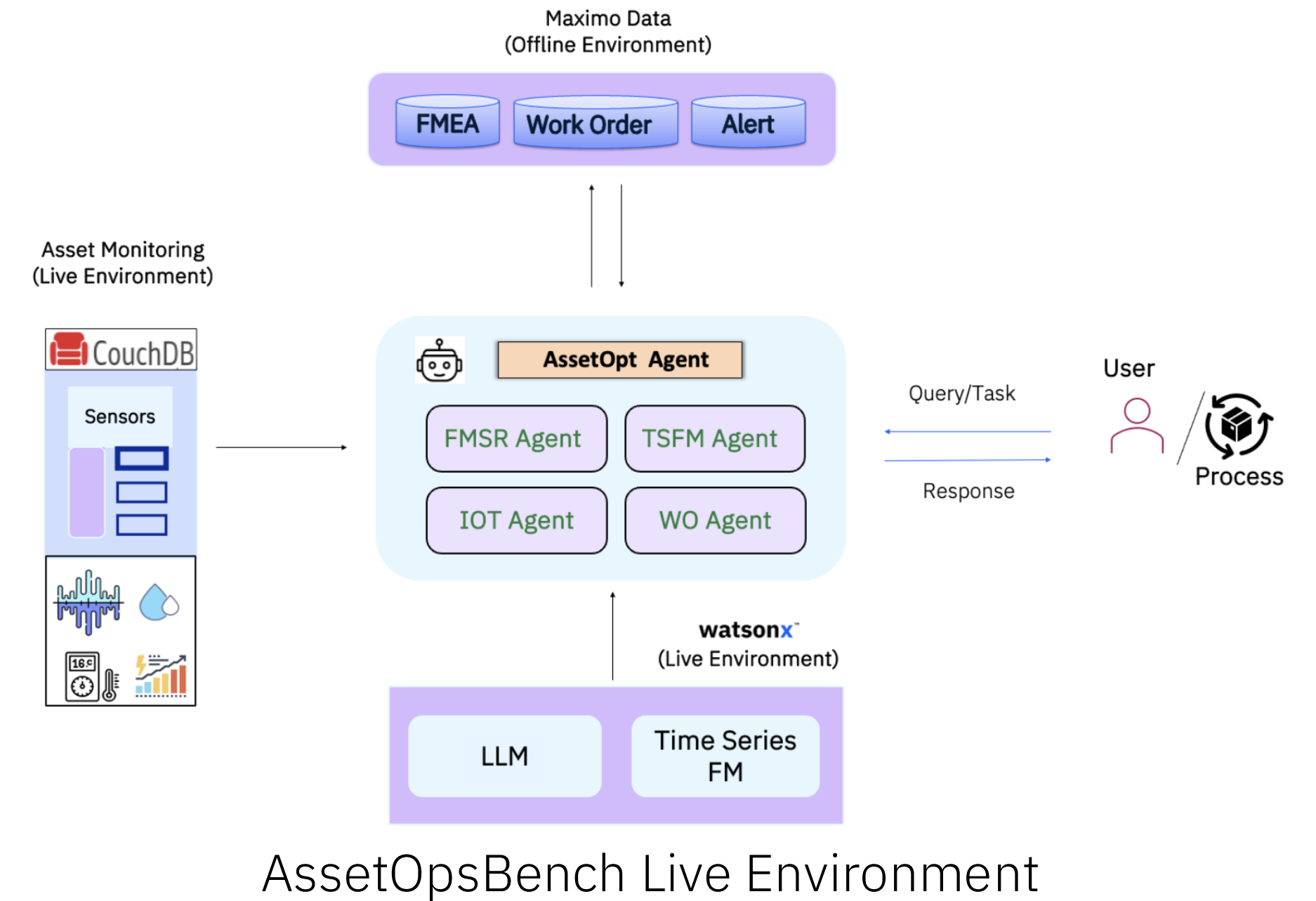


AssetOpsBench : Open-Source Initiative



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



141 Utterance Distribution

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

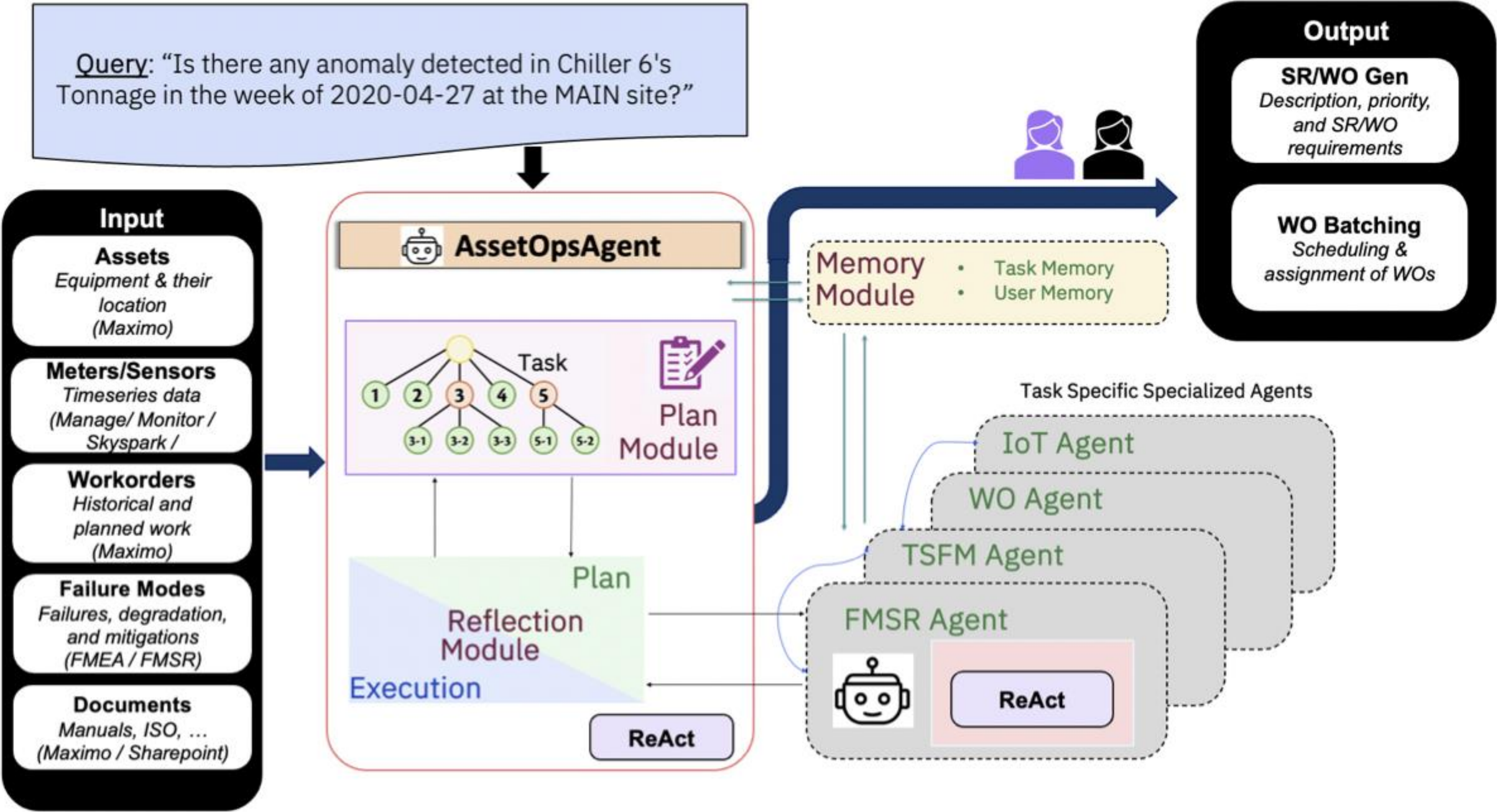
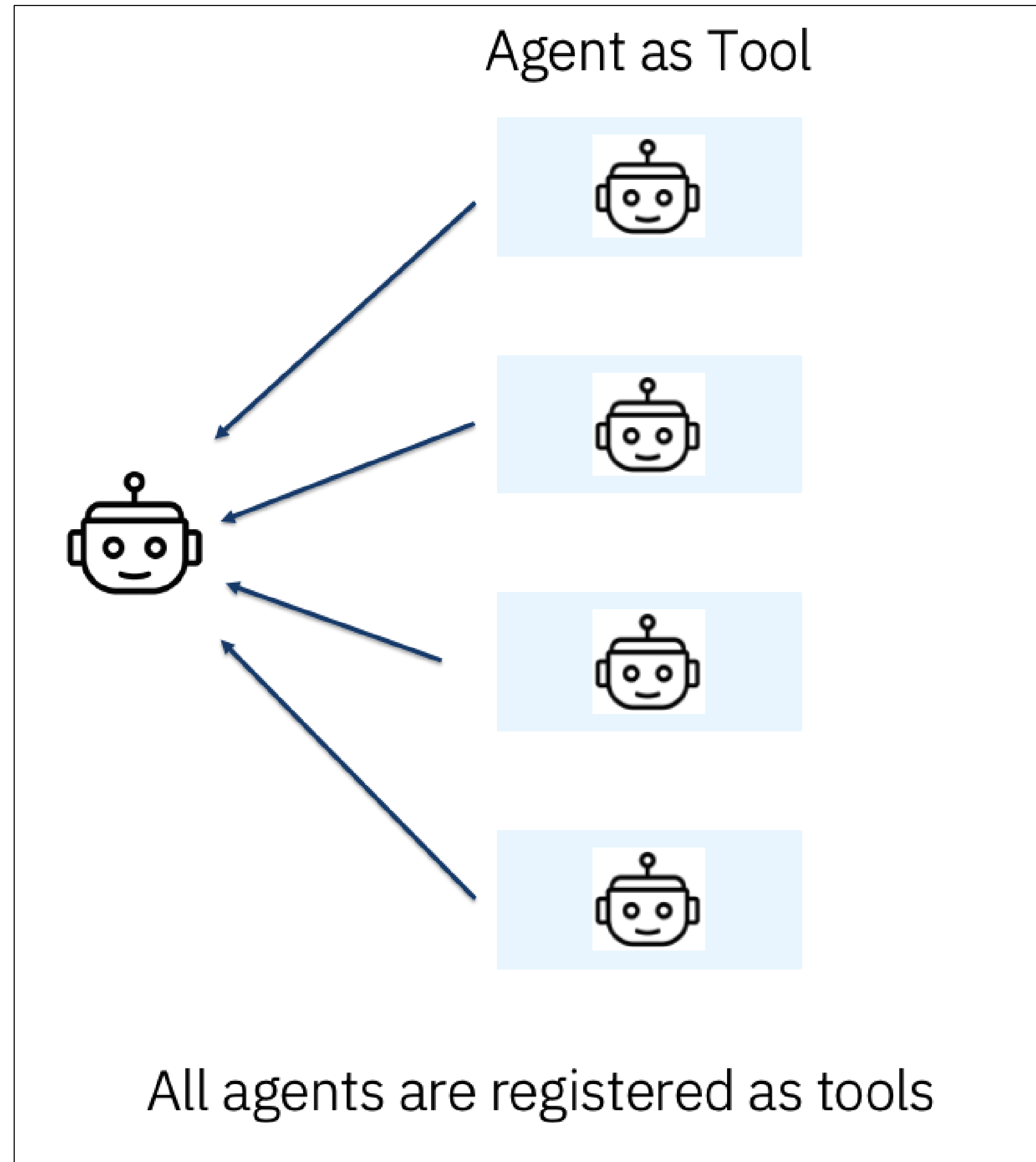
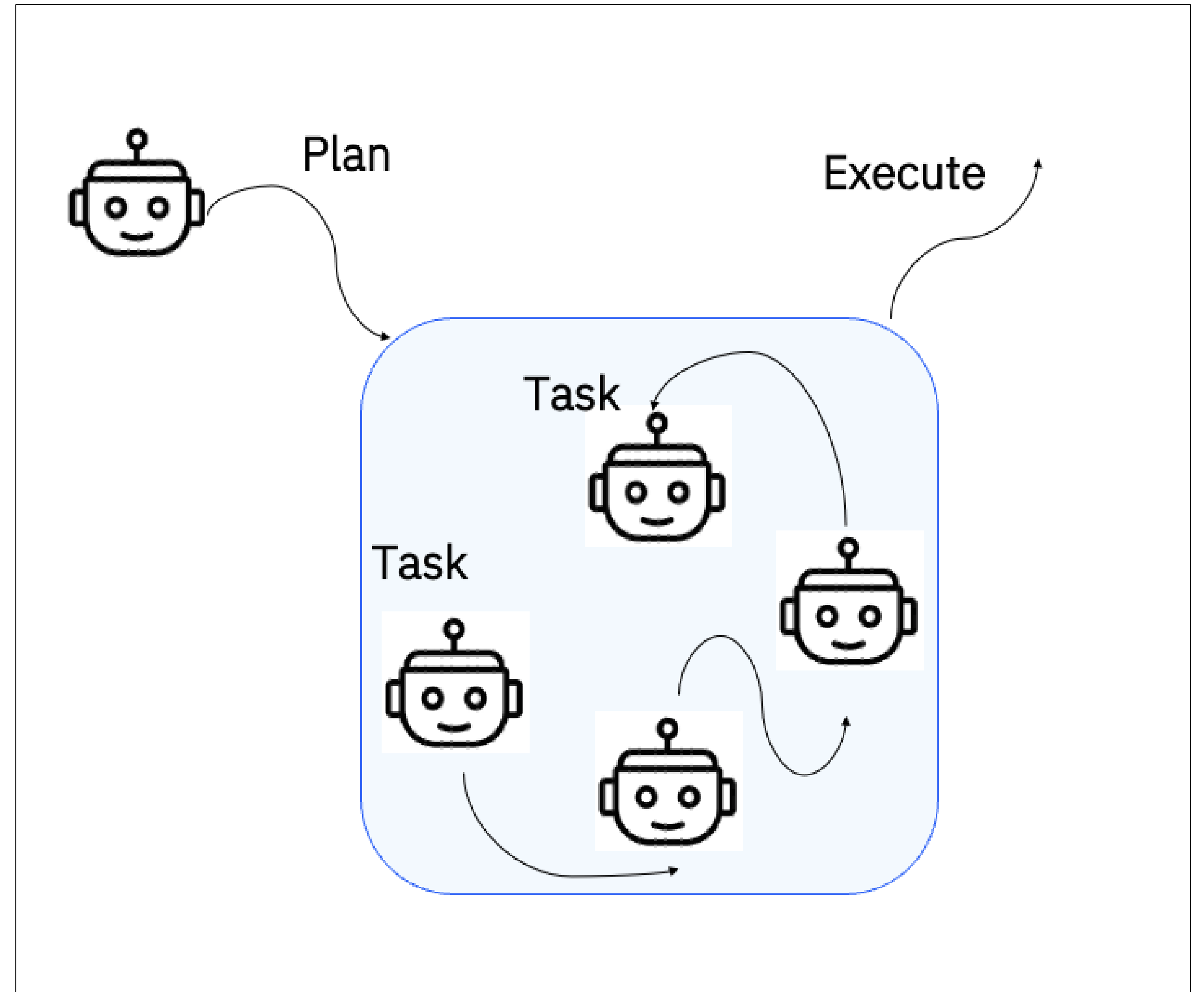


Figure 2: Architecture of the Multi-Agent System: **Time Series Foundation Model (TSFM) Agent**, **Failure Mode Sensor Relations (FMSR) Agent**, **Work Order (WO) Agent**

AssetOpsBench : Multi-Agent Implementation Strategy



Agent-As-Tool Approach



Plan-Execute Approach

AssetOpsBench : Evaluation



Automatic Evaluation of Agentic Workflow

Ground Truth Preparation

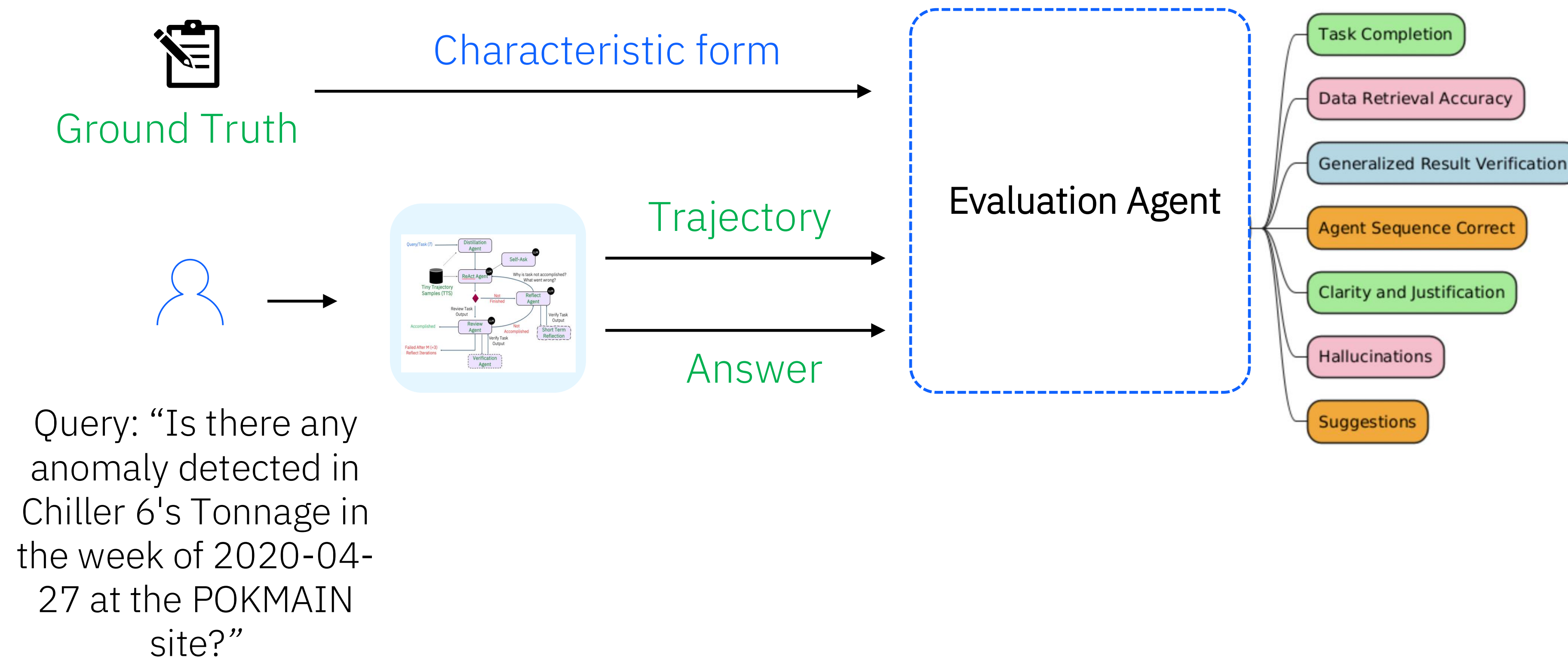
- **Characteristic form** describe the final output along with the process/step to be conducted. As an example:

“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

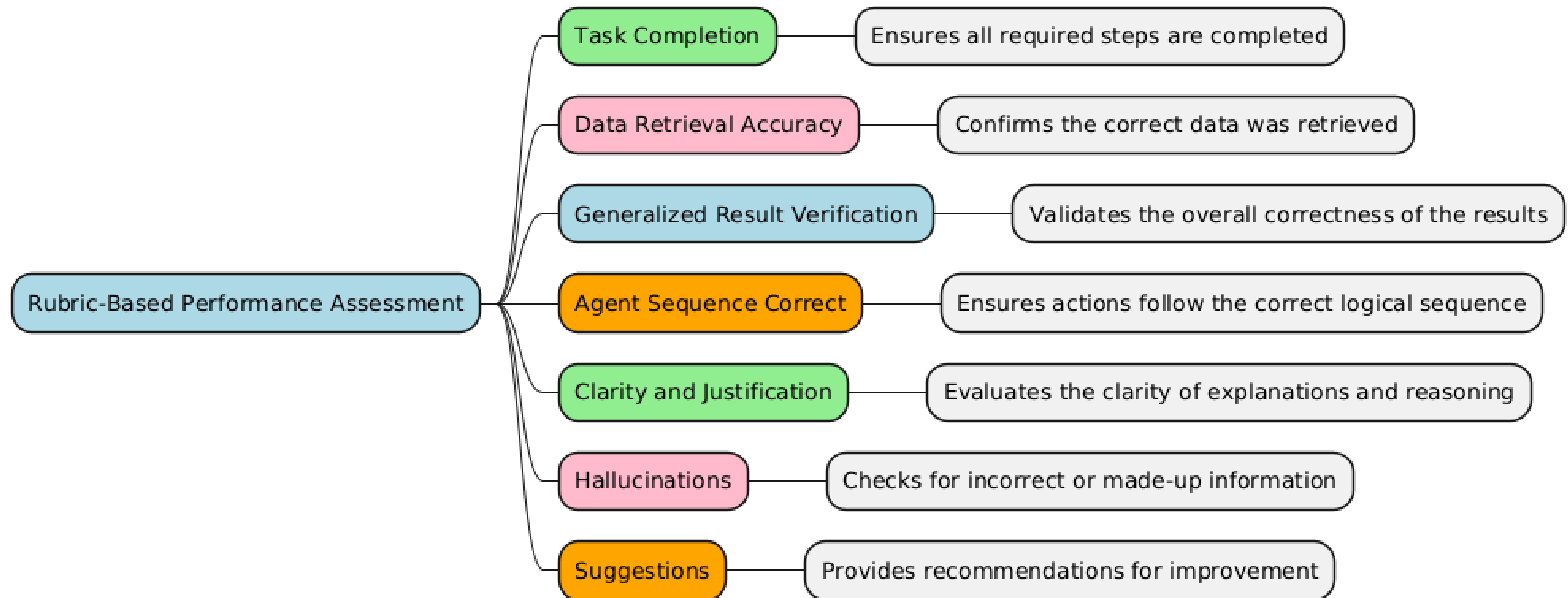
Evaluation Agent: Rubric-based Performance Assessment



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.



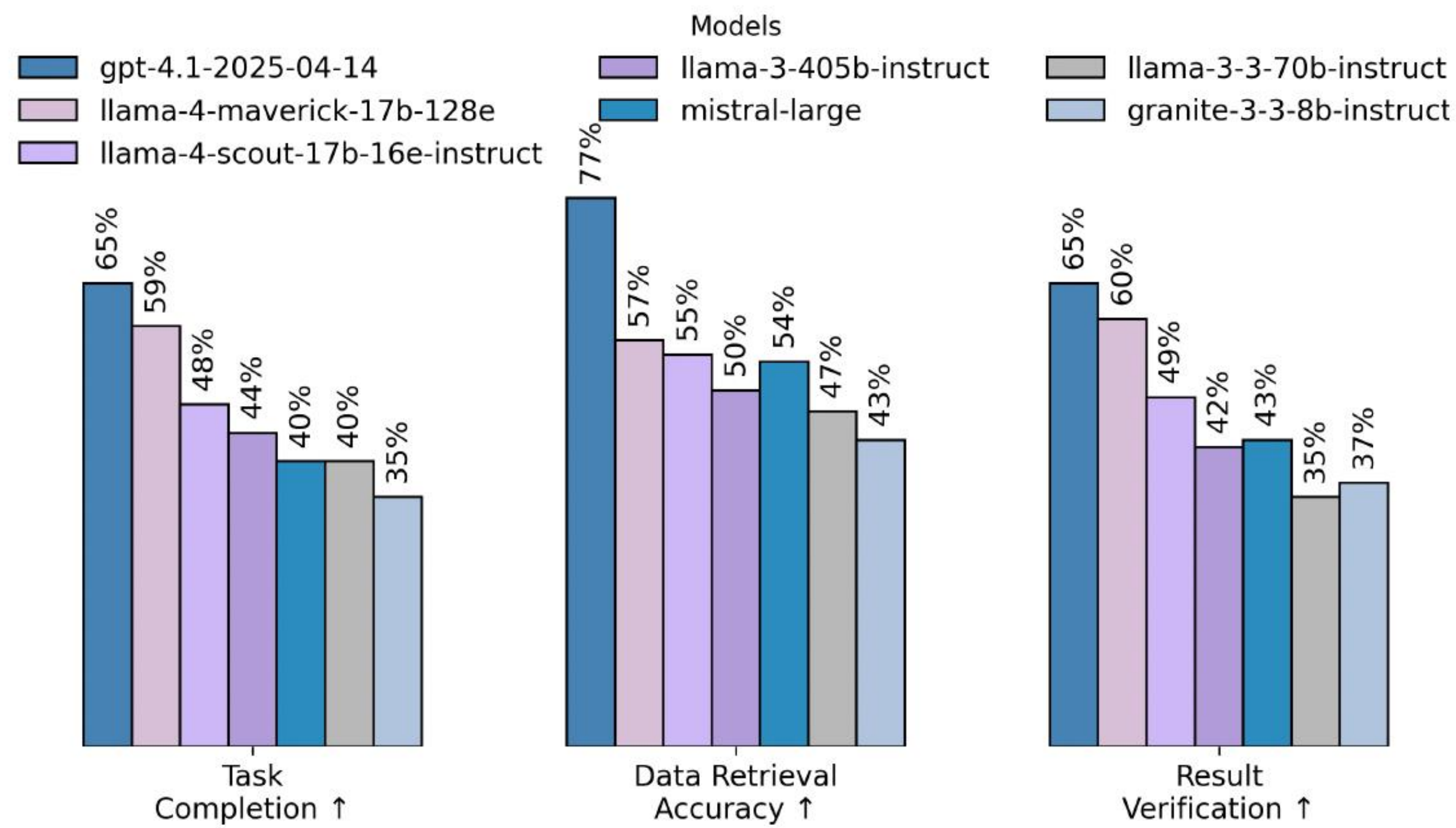
AssetOpsBench : Summary of Result



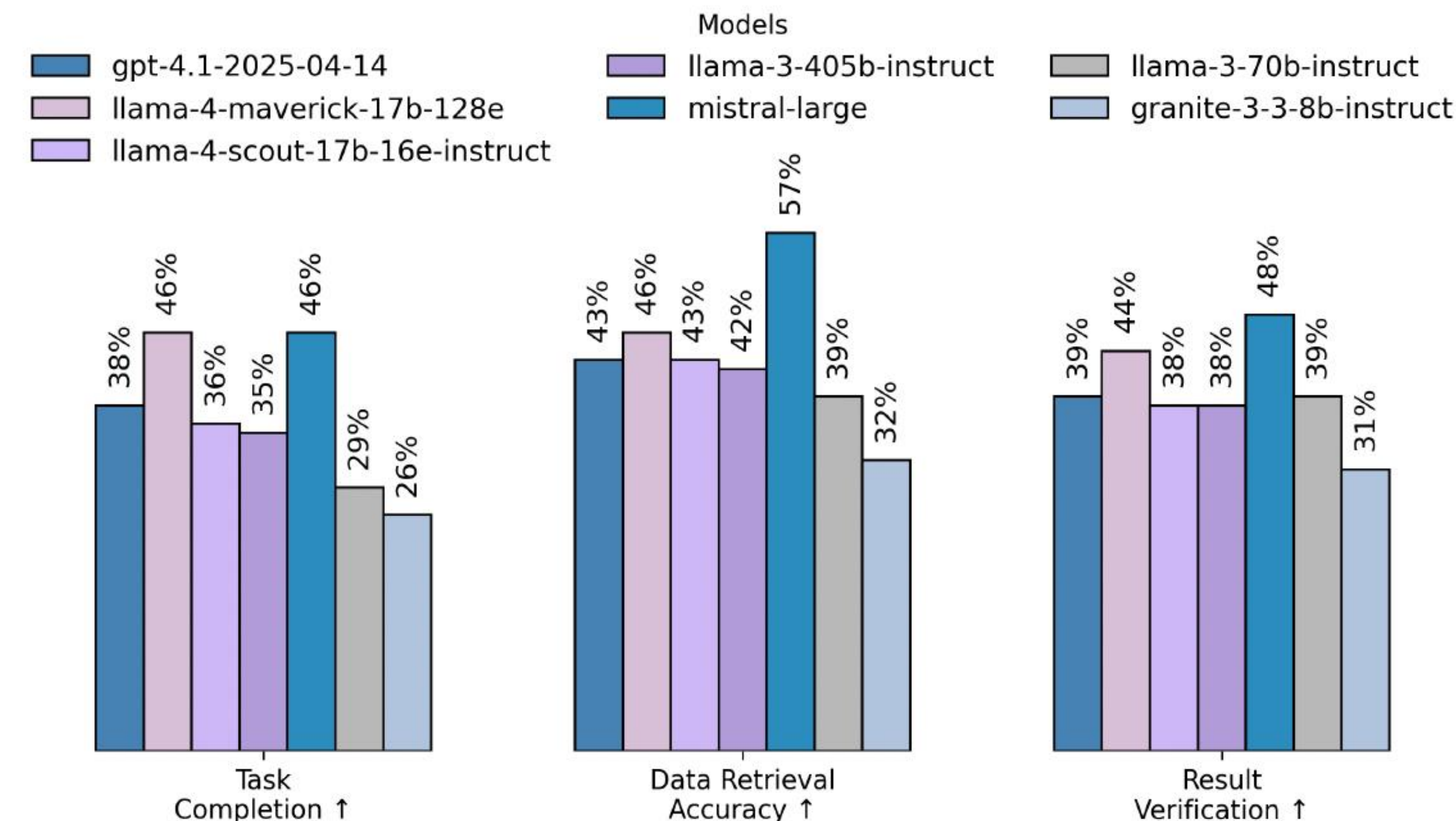
Leaderboard

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

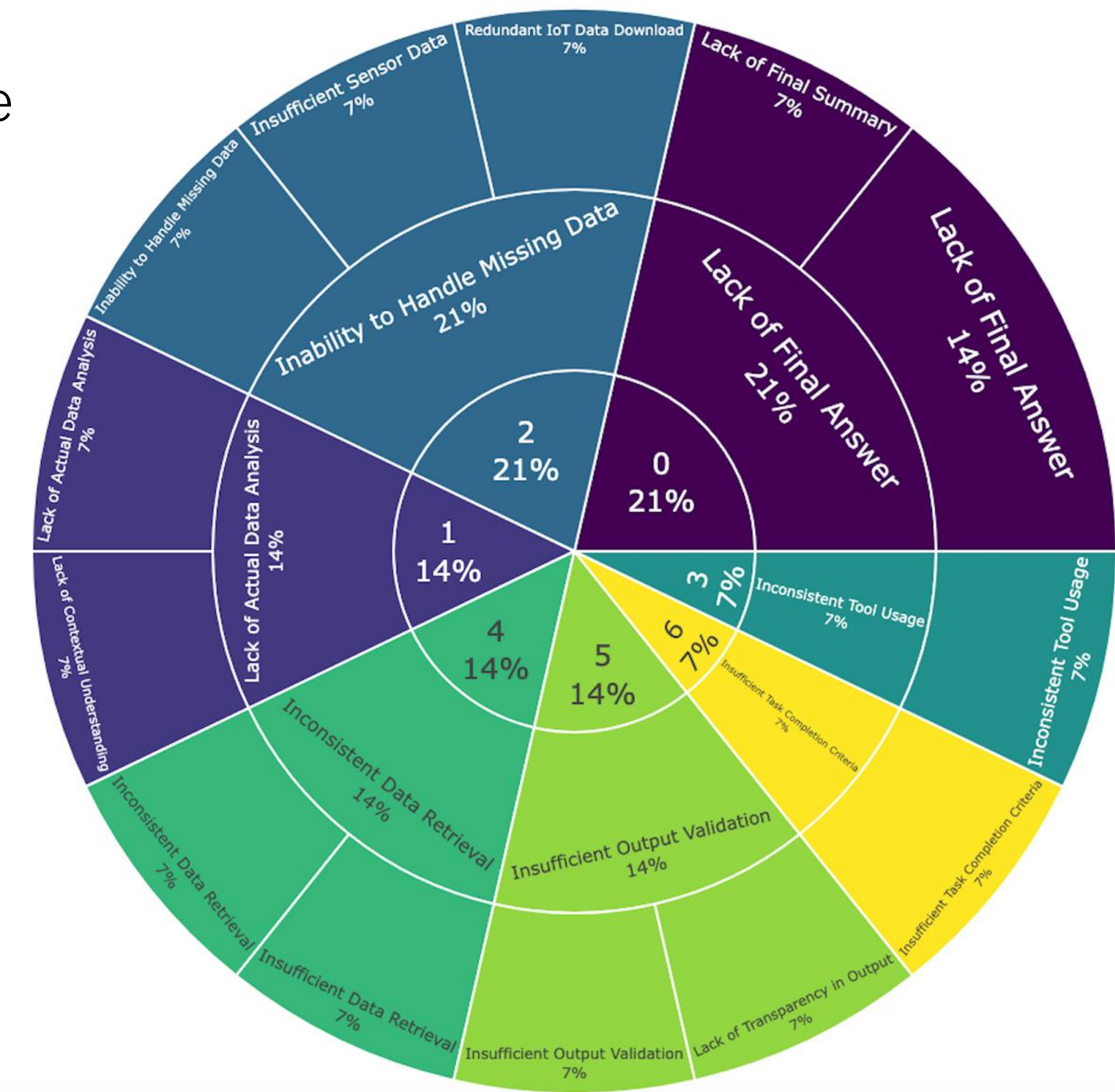
Emergent Failure Mode Discovery

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.

Datasets:

ibm-research

FailureSensorIQ

like

3

Follow

IBM Research

390

Tasks:

Question Answering

Modalities:

Text

Formats:

json

Languages:

English

Size:

1K - 10K

ArXiv:

Libraries:

Datasets

pandas

Croissant

+ 1

License:

apache-2.0

Dataset card

Data Studio

Files and versions

xet

Community

9

Dataset Viewer

Auto-converted to Parquet

API

Embed

Data Studio

Subset (2)

multi_true_multi_choice_qa · 5.63k rows

Split (1)

train · 5.63k rows

Search this dataset

subject	id	question	options	option_ids
string · classes	int64	string · classes	list · lengths	list · lengths
1 value	1	5.63k	327 values	5
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

IBM RESEARCH · BENCHMARK · V1

0

Share

FailureSensorIQ

A Multi-Choice QA (MCQA) dataset that explores the relationships between sensors and failure modes for 10 industrial assets.

Leaderboard

Discussion (0)

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

Results independently reproduced by Kaggle. Learn more

Last updated November 23, 2025

#	Model	Score	Consistency	F-Score	Elimination Accuracy
1	Gemini-3-Pro-Preview	69.1%	63.8%	69.1%	78.8%
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	67.8%	57.5%	67.8%	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%	65.8%	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

Shuxin Lin
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!](#)