



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

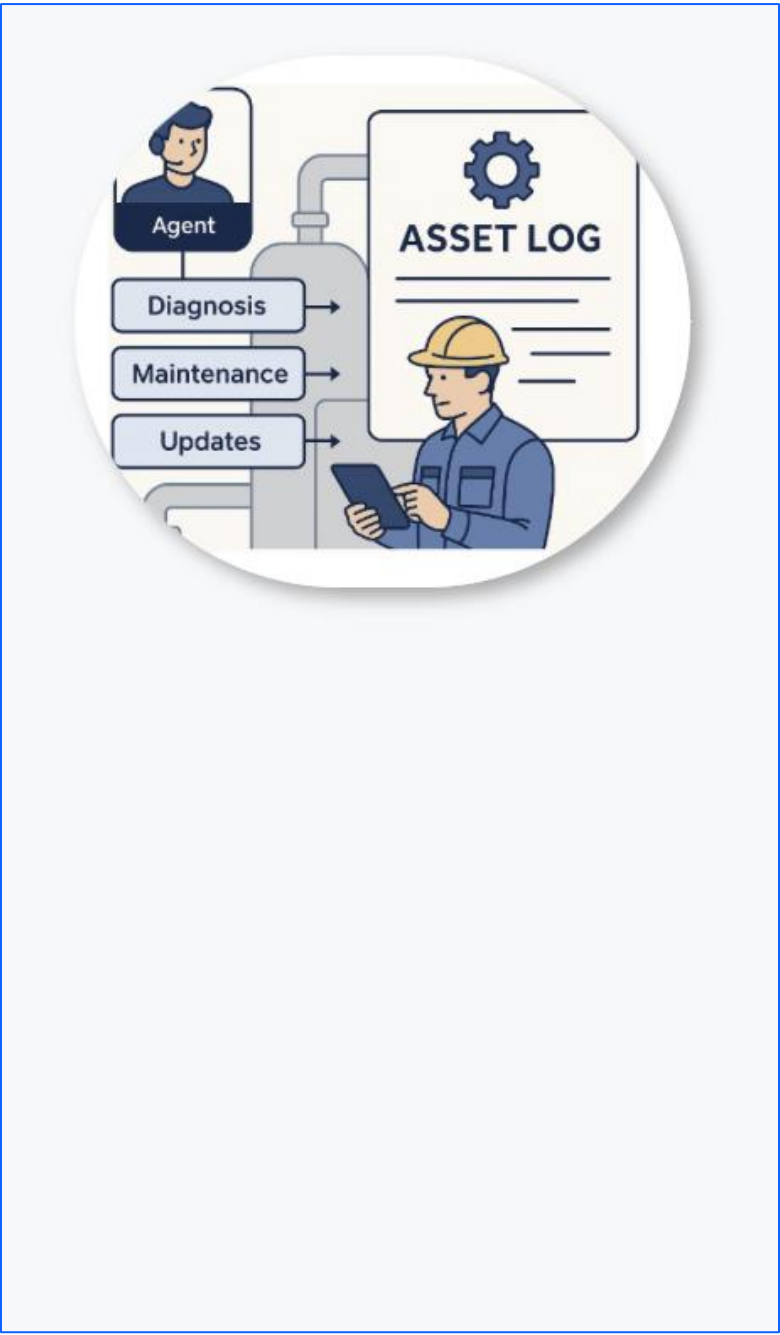
Dhaval Patel
Shuxin Lin

NeurIPS 2025 IBM Booth



Demonstration Links

AI Agent Challenges: Autonomous Industrial Agents
<https://www.codabench.org/competitions/10206/>



Get Started

Phases

My Submissions

Results

Forum

?

Introduction

Resources

Track 1: Task Planning

Track 2: Task Execution

Team Registration

Terms

AI Challenge: Autonomous Industrial Agents

This AI challenge invites participants to **learn, design, develop, and evaluate autonomous AI agents** that solve realistic industrial tasks across the full pipeline:

Sensing → **Reasoning** → **Actuation**

This AI Challenge is organized as a part of conference - [13th INTERNATIONAL CONFERENCE ON DATA SCIENCE - CODS](#)

Challenge Overview

Participants will work with a *curated set of scenarios* rooted in **Industry 4.0** applications such as:

- predictive maintenance
- fault diagnosis
- work-order generation
- root-cause analysis

These tasks demand both:

- **Strong single-agent intelligence**, and
- **Coordinated multi-agent behavior**

We enable LLM access hosted on **Watsonx.ai platform**. Please check [Forum](#) to get an access.

Example: Loading Scenarios

```
from datasets import load_dataset

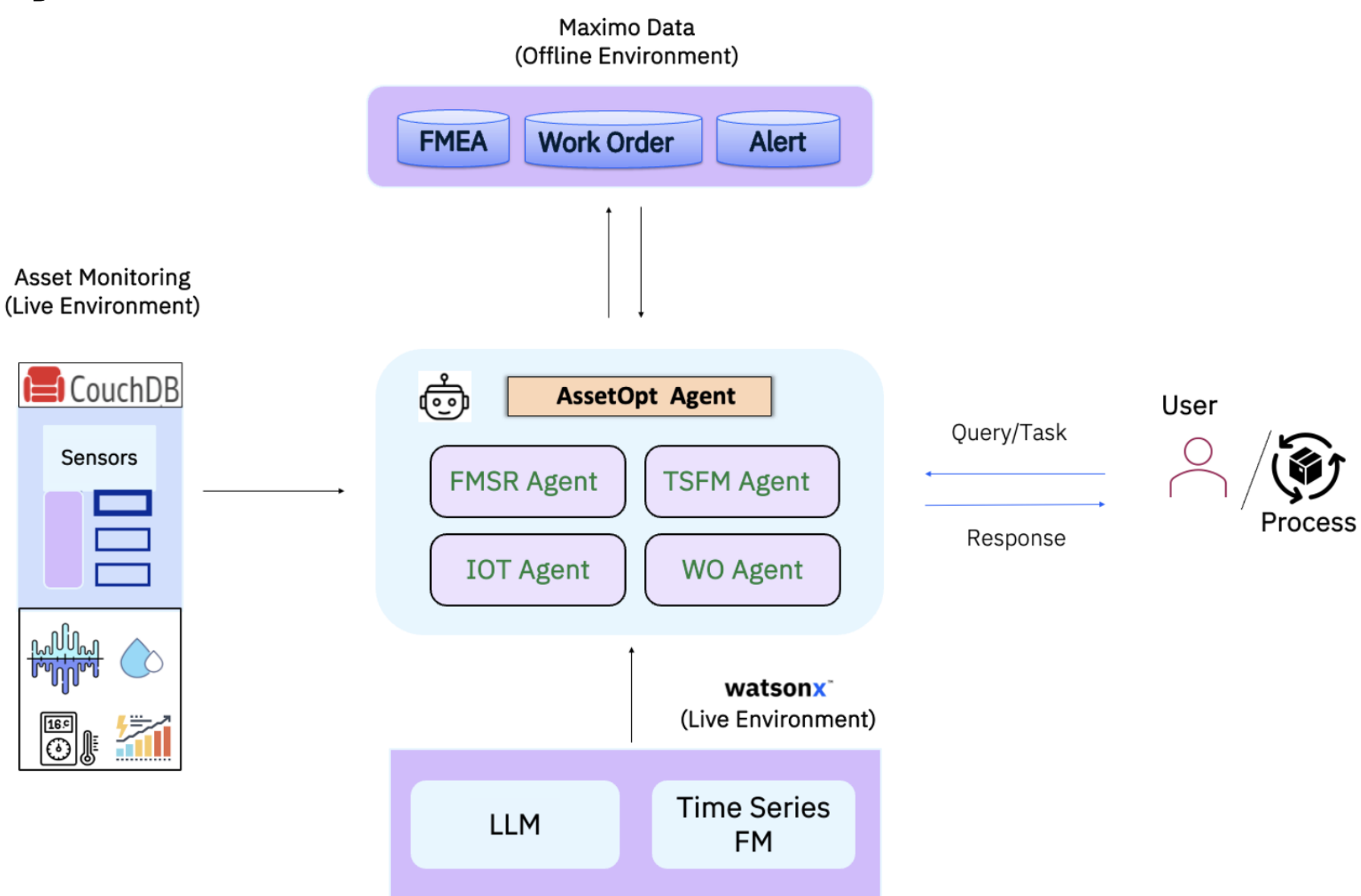
# Login using e.g. `huggingface-cli login` to access this dataset
ds = load_dataset("ibm-research/AssetOpsBench", "scenarios")
```

© 2025 IBM Corporation

2

AssetOpsBench: Open-Source Benchmark for Industry 4.0

- Framework to assess Gen AI solutions' ability to solve Industry 4.0 Automation "Scenarios": **June GA**
- **Simulated** industrial environment, **9 multi-source data sets** (work orders, FMEAs, timeseries) and **4 agents** (IoT, data science, work order, failure mode to sensor mapping)
- **140+** human-authored natural language queries, grounded in **enterprise industrial scenarios**
- **Agent harness**: systematic procedure for automated discovery of emerging failure modes

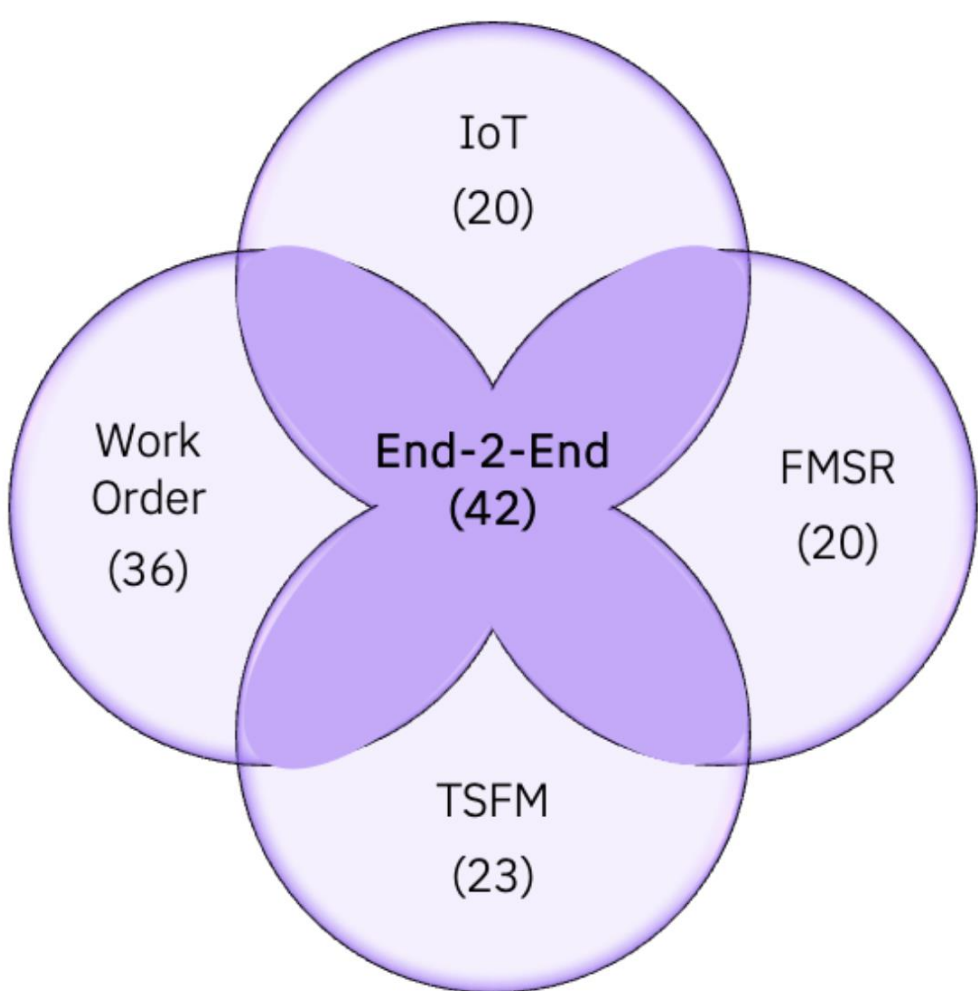


AssetOpsBench Live Environment

AssetOpsBench: Benchmarking AI Agents for Task Automation in Industrial Asset Operations and Maintenance

Dhaval Patel^{1*} Shuxin Lin^{1*} James Rayfield^{1*} Nianjun Zhou^{1*}
Roman Vaculin¹ Natalia Martinez¹ Fearghal O'donncha² Jayant Kalagnanam¹
¹IBM Research - Yorktown ²IBM Research - Ireland
pateldha@us.ibm.com, shuxin.lin@ibm.com, jtray@ibm.com, jzhou@us.ibm.com,
vaculin@us.ibm.com, Natalia.Martinez.Gil@ibm.com, feardonn@ie.ibm.com,
jayant@us.ibm.com
*Equal contribution

AssetOpsBench Project <https://github.com/IBM/AssetOpsBench>



141 Utterance Distribution

AssetOpsBench: Hugging face Dataset

```
from datasets import load_dataset

# Login using e.g. 'huggingface-cli login' to access this dataset
ds = load_dataset("ibm-research/AssetOpsBench", "scenarios")
```

Datasets: ibm-research/AssetOpsBench

like 1

Follow IBM Research 327

Dataset card

Data Studio

Files and versions

Community 2

Set

Subset (2)
scenarios · 141 rows

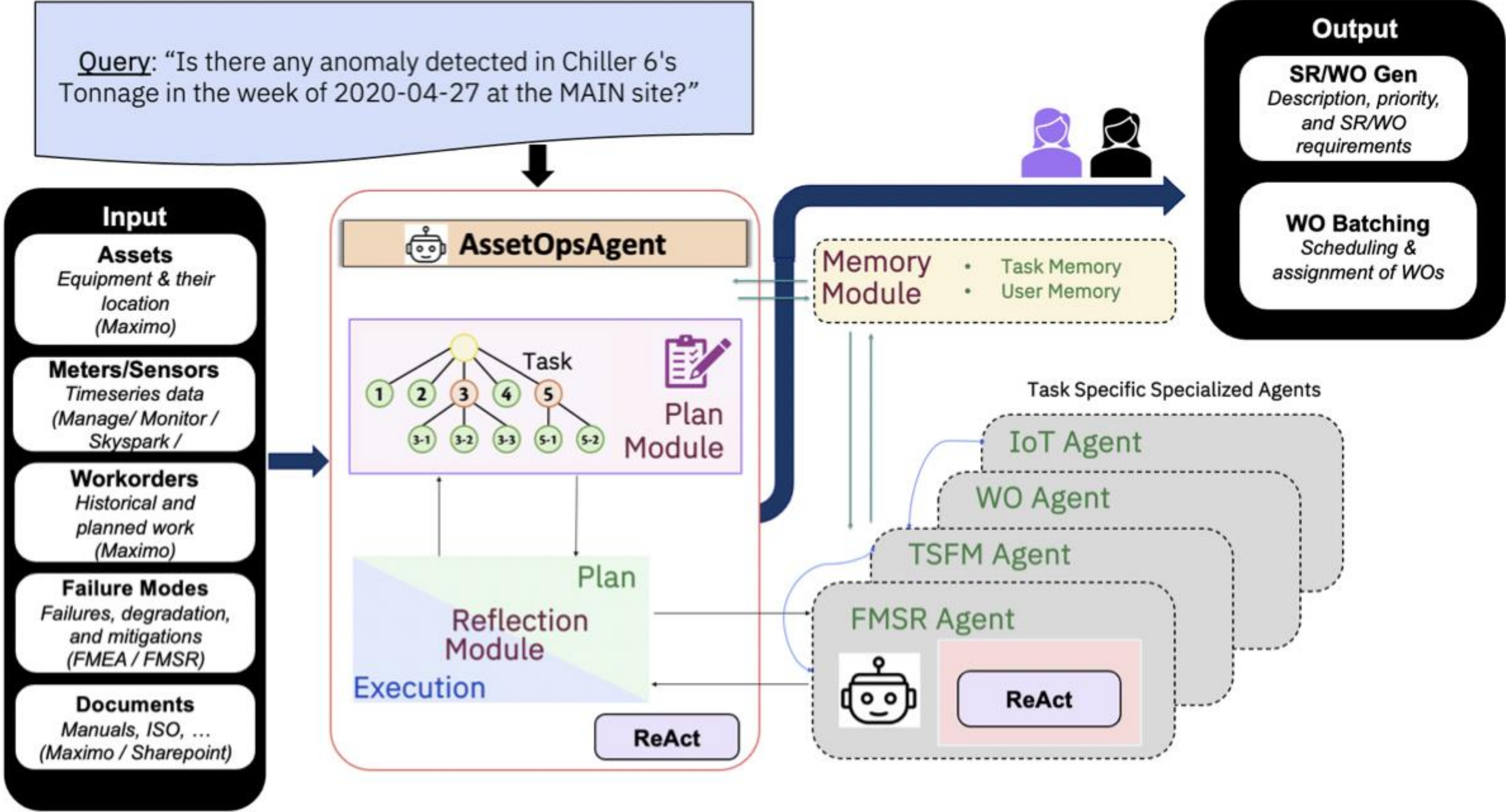
Split (1)
train · 141 rows

Search this dataset

| id | type | text | category | characteristic_form | deterministic | note |
|-------|------------------|--|-------------------------|---|---------------|----------------|
| int64 | string · classes | string · lengths | string · classes | string · lengths | bool | string · class |
| | | | | | | |
| 1 | 4 values | 27 | 6 values | 49 | 2 classes | 2 values |
| 622 | | 468 | | 936 | | |
| | | | | | | |
| 220 | TSFM | Finetune a forecasting model for 'Chiller 9 Condenser Water Flow'... | Tuning Query | The finetuned forecasting model is saved in save_model_dir=tunedmodels with result stored in... | null | |
| 221 | TSFM | Finetune a forecasting model for 'Chiller 9 Condenser Water Flow'... | Tuning Query | The finetuned forecasting model is saved in save_model_dir=tunedmodels with result stored in... | null | |
| 222 | TSFM | I need to perform Time Series anomaly detection of 'Chiller 9... | Anomaly Detection Query | The anomaly detection results are stored in file data/tsfm_test_data/tsad_conformal.csv | null | |
| 223 | TSFM | Find and run several methods to analyze data sensor 'Chiller 9... | Complex Query | The forecasting results for 'Chiller 9 Condenser Water Flow' using data in... | null | |
| 400 | Workorder | Get the work order of equipment CWC04013 for year 2017. | null | There will be 33 records. The expected response should retrieve all work orders for equipment... | true | |
| 401 | Workorder | I would like to check the work order distribution for the equipment... | null | Work order with primary Code MT010 occurred 3 times and code MT013 occurred once. The expected respons... | true | |

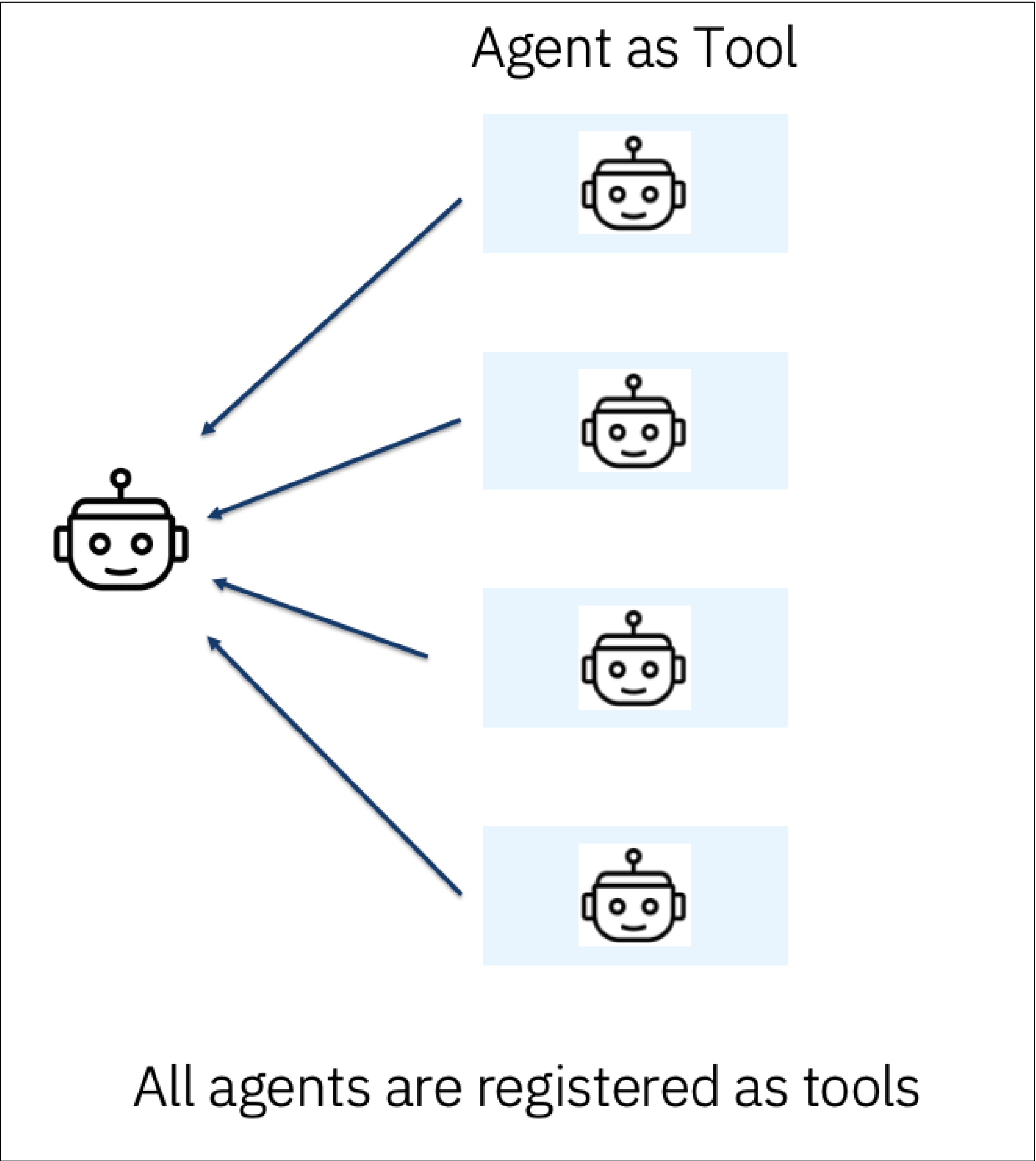
<https://huggingface.co/datasets/ibm-research/AssetOpsBench>

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

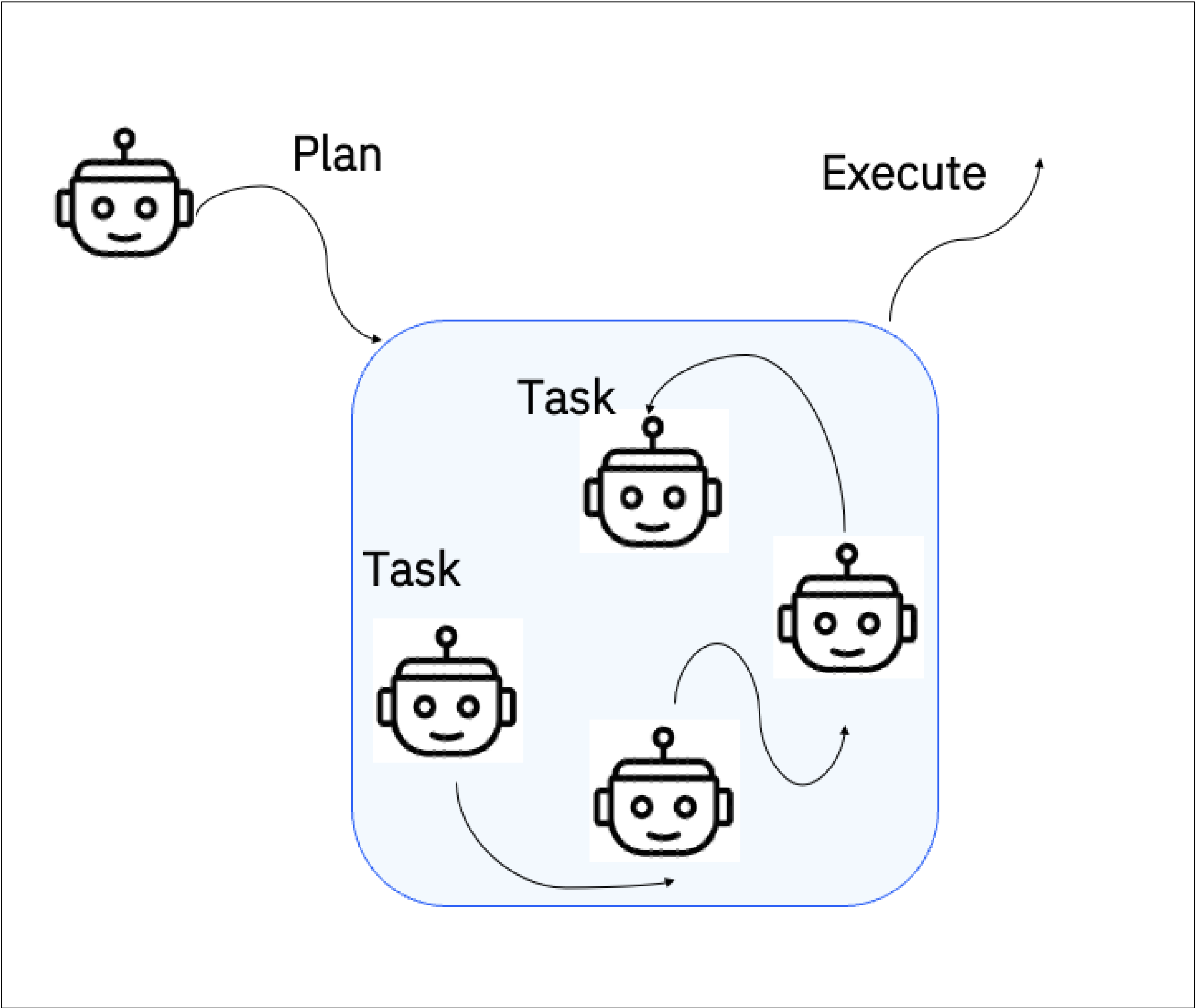


Architecture of the Multi-Agent System: Time Series (TSFM) Agent, Failure Mode Sensor Relations (FMSR) Agent, Work Order (WO) Agent, Time Series foundation model (TSFM) Agent

AssetOpsBench: Multi-Agent Implementation Strategy

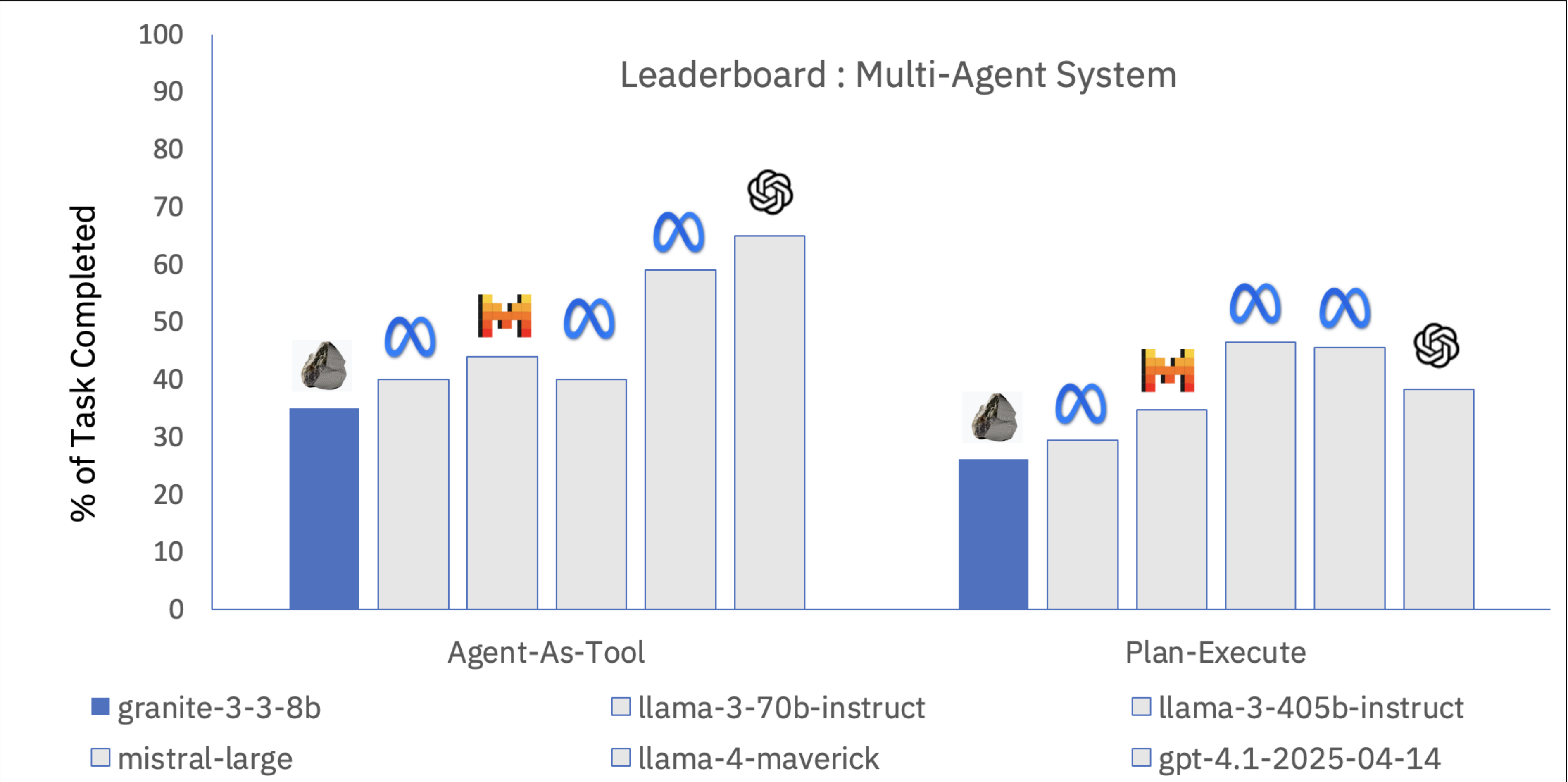


Agent-As-Tool Approach



Plan-Execute Approach

AssetOpsBench Leaderboard: Open Source V1, June 2025



Extensive Evaluation of two different paradigm for Multi-Agent System

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"... |

IBM Confidential

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



Event Announcement:

- We will be presenting FailureSensorIQ on Dec 4th, 11 a.m. - 2 p.m. PST in Exhibit Hall C,D,E #1515, NeurIPS 2025 @ San Diego Convention Center.
- We will be presenting Lightning Talk in NeurIPS 2025 Socials “*Evaluating Agentic Systems: Bridging Research Benchmarks and Real-World Impact*” on Dec 4th, NeurIPS 2025 @ San Diego Convention Center. <https://luma.com/mkyvypm>
- 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!](#)