workflows community initiative

Oak Ridge National Laboratory

Dec 10, 2025 – Workflows Community Initiative

For Agents and with Agents: The Role of Provenance Data in Agentic Workflows

Presenter: Renan Souza, PhD

Research Scientist

National Center of Computational Sciences

Oak Ridge National Laboratory

U.S. DEPARTMENT of ENERGY

ORNL IS MANAGED BY UT-BATTELLE LLC
FOR THE US DEPARTMENT OF ENERGY

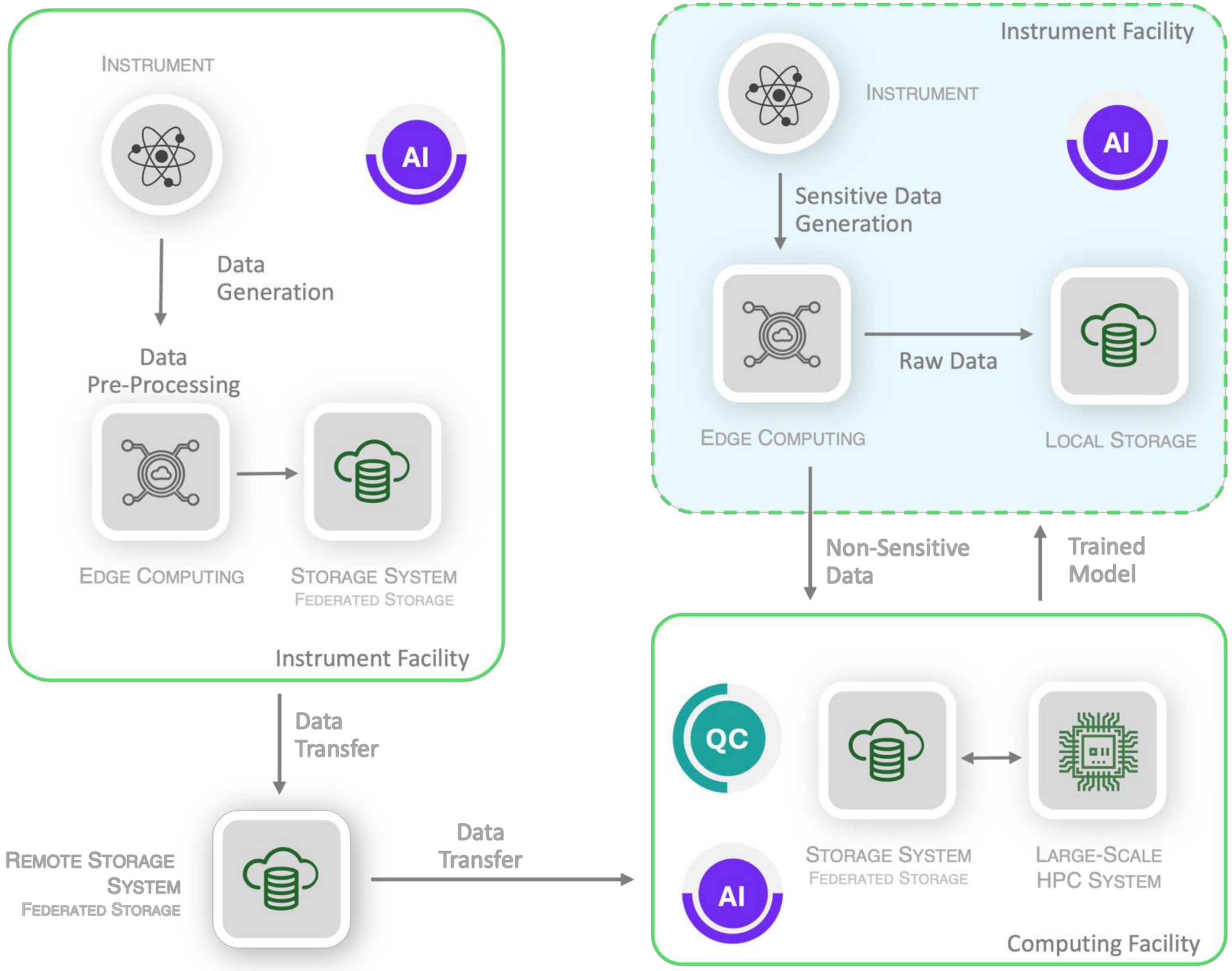
The current reality in Scientific Campaigns: DISCONNECTED DATA, WORKFLOWS, AND USERS

MANUAL ORCHESTRATION ACROSS FACILITIES

MANUAL AND AD-HOC DATA MOVEMENTS

AI + HPC + INSTRUMENTS CREATING COMPLEXITY

SCIENTISTS AS SYSTEM AND DATA INTEGRATORS



CONVERGENCE LANDSCAPE

Edge – Cloud – HPC – Instruments – AI

AI MODELS + SIMULATIONS + EXPERIMENTS

Enable intelligent prediction, automation, and integration

MULTI-FACILITY INTEGRATION

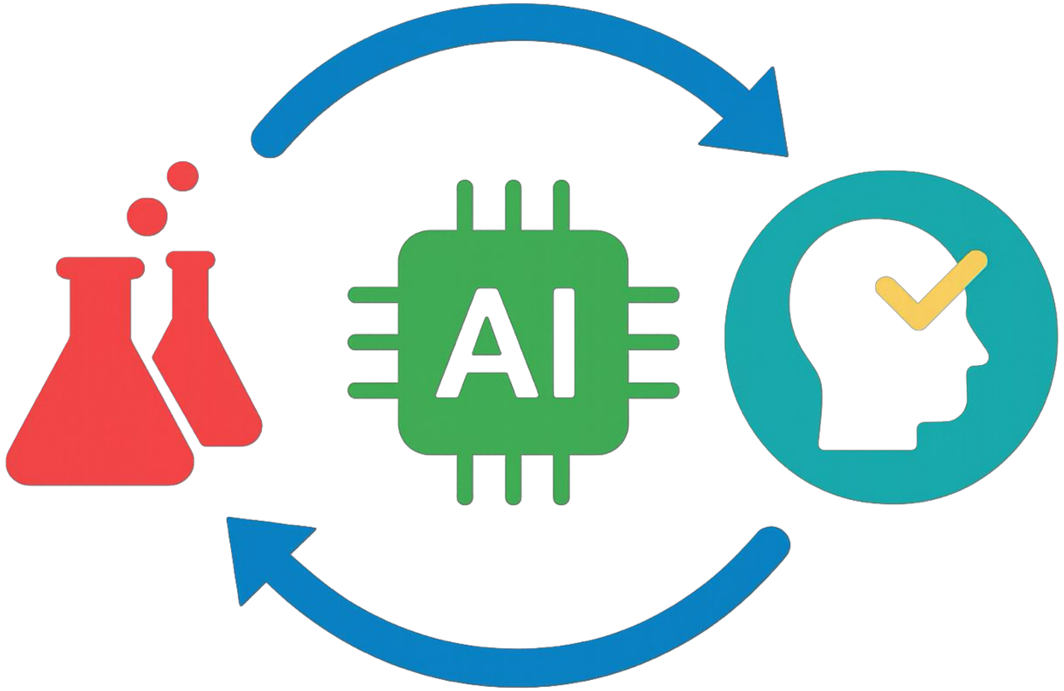
Aggregate and stage distributed data, and execute large-scale simulations and analytics

INSTRUMENT AND EDGE

Capture and stream experimental data

CONTINUOUS DECISION LOOPS

Drive continuous adaptive decision-making

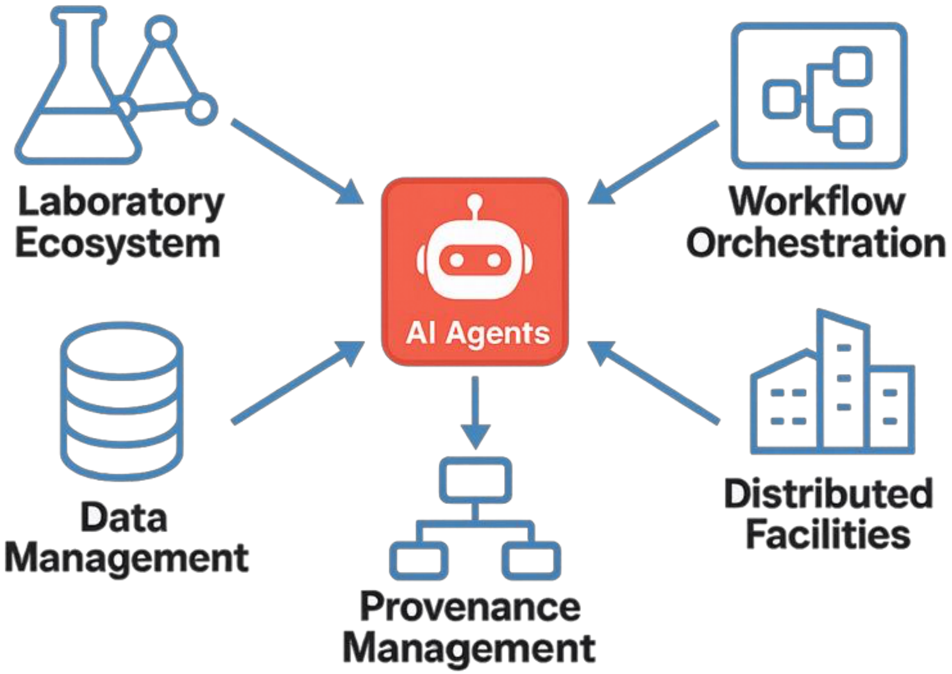


TOWARDS AUTONOMOUS SCIENCE AT SCALE

From static workflows to intelligent, agent-driven systems

Continuous, closed-loop experimentation integrating HPC, AI, and instruments

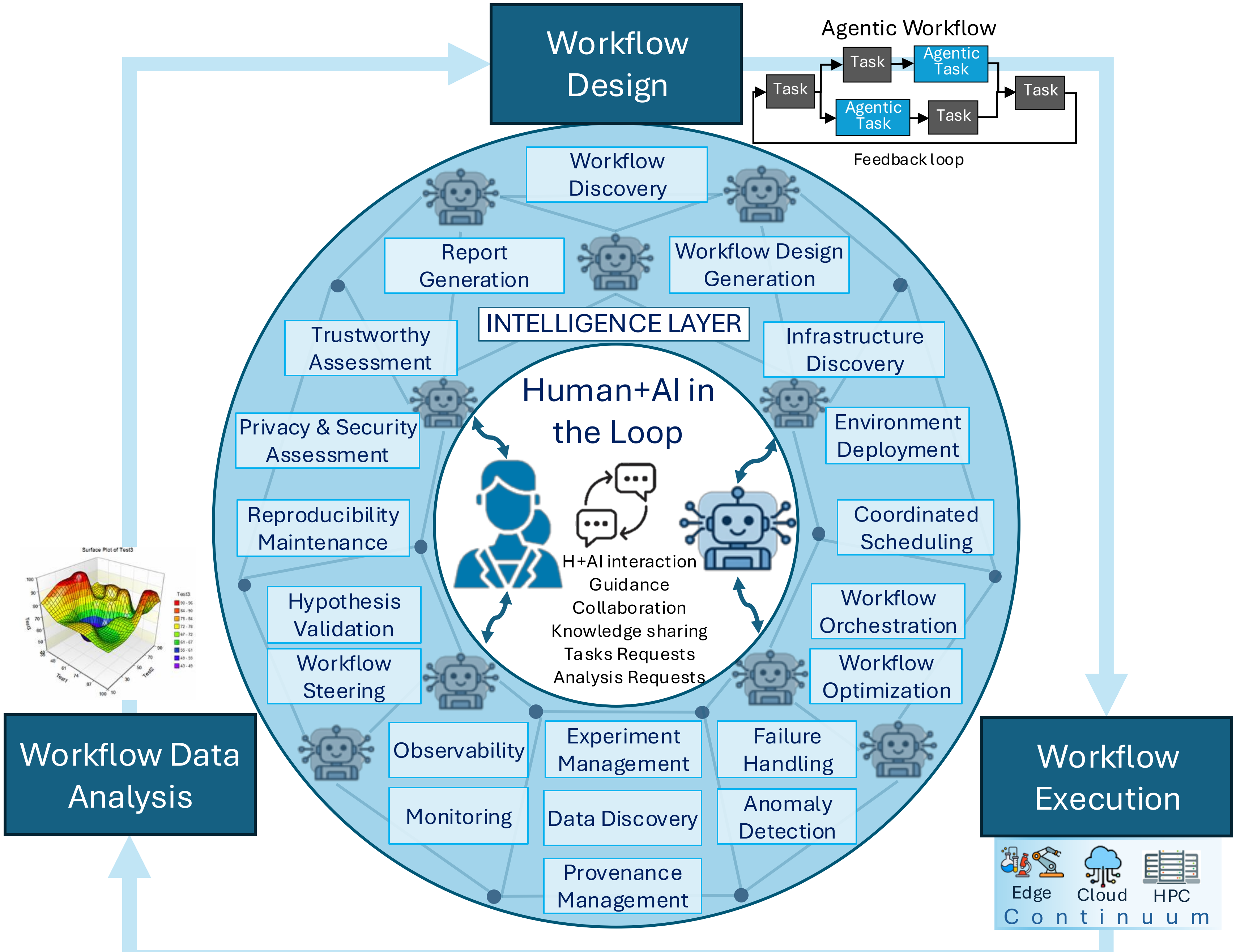
Accelerate discovery cycles from years to months while ensuring reproducibility and openness



Agentic Scientific Workflows

From single-agent actions to collaborative, autonomous workflows

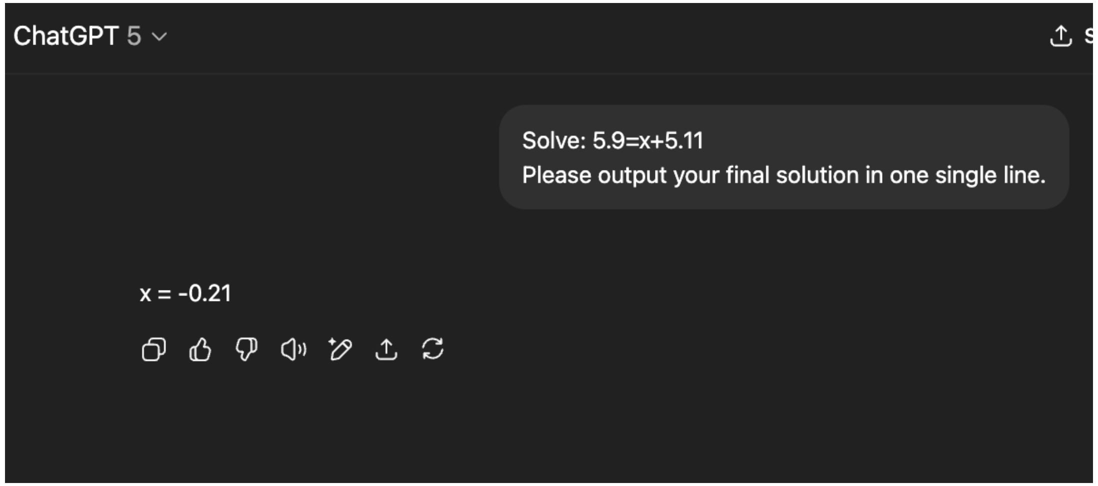
- Goal-driven, adaptive, autonomous decision-making
- Dynamic workflow tasks definition and execution
- Multiple specialized collaborating agents
- Planning, Orchestration, and Coordination between:
  - Agentic and non-agentic tasks
  - Humans
  - Federated systems

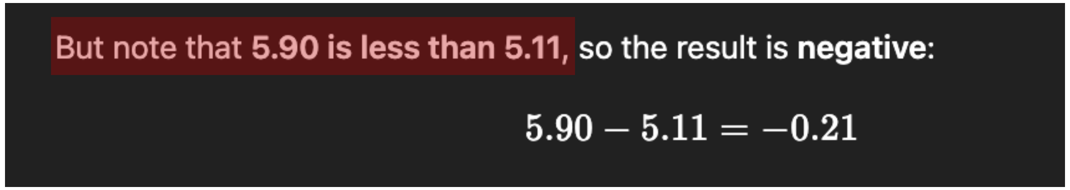


W. Shin, R. Souza, D. Rosendo, et al. The (R)evolution of Scientific Workflows in the Agentic AI Era: Towards Autonomous Science. Best paper on WORKS co-located with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2025.

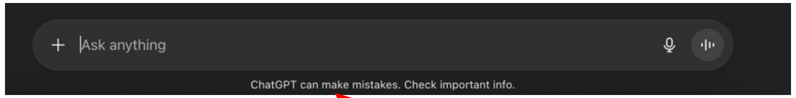
Oak Ridge National Laboratory

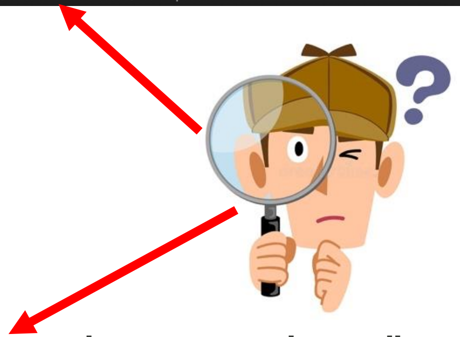
All very exciting and promising, but…





ChatGPT5 on 8/19/2025

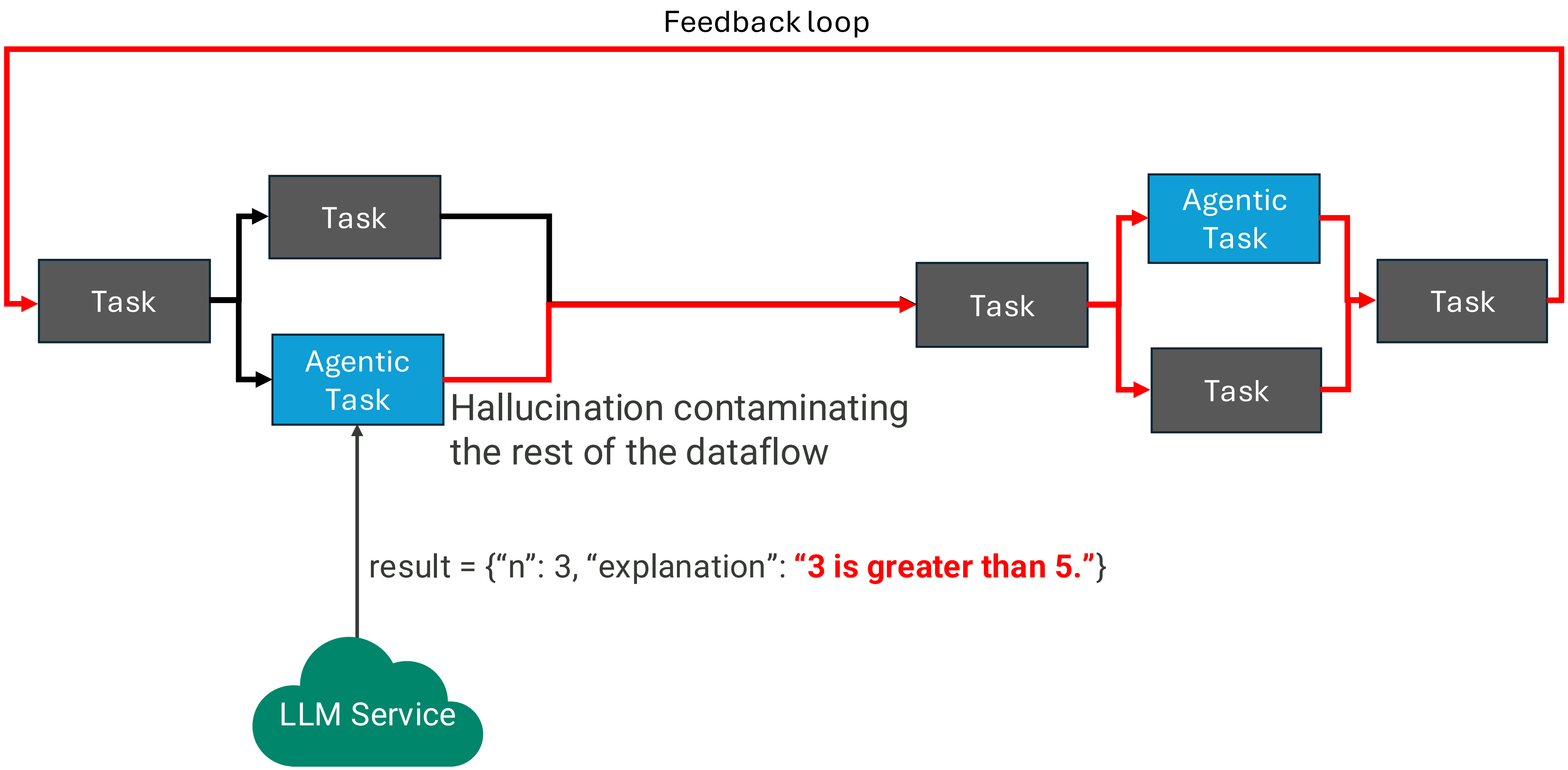




"ChatGPT can make mistakes."

In fact, **GenAI** is non-deterministic in nature and **hallucinations are very common**, even in state-of-the-art models.

Dataflow Contamination



Oak Ridge National Laboratory

How to make Agentic Scientific Workflows more reliable?

How to enable:

- **Reproducibility,**
- **[Agentic] Accountability,**
- **Transparency,** and

other critical principles in science?

How to **keep track of hallucinations**, **mitigate, and remediate their downstream impact** in agentic workflows?

*Spoiler alert: **provenance data management** can help!*

Provenance Data to Support Large-scale Workflows

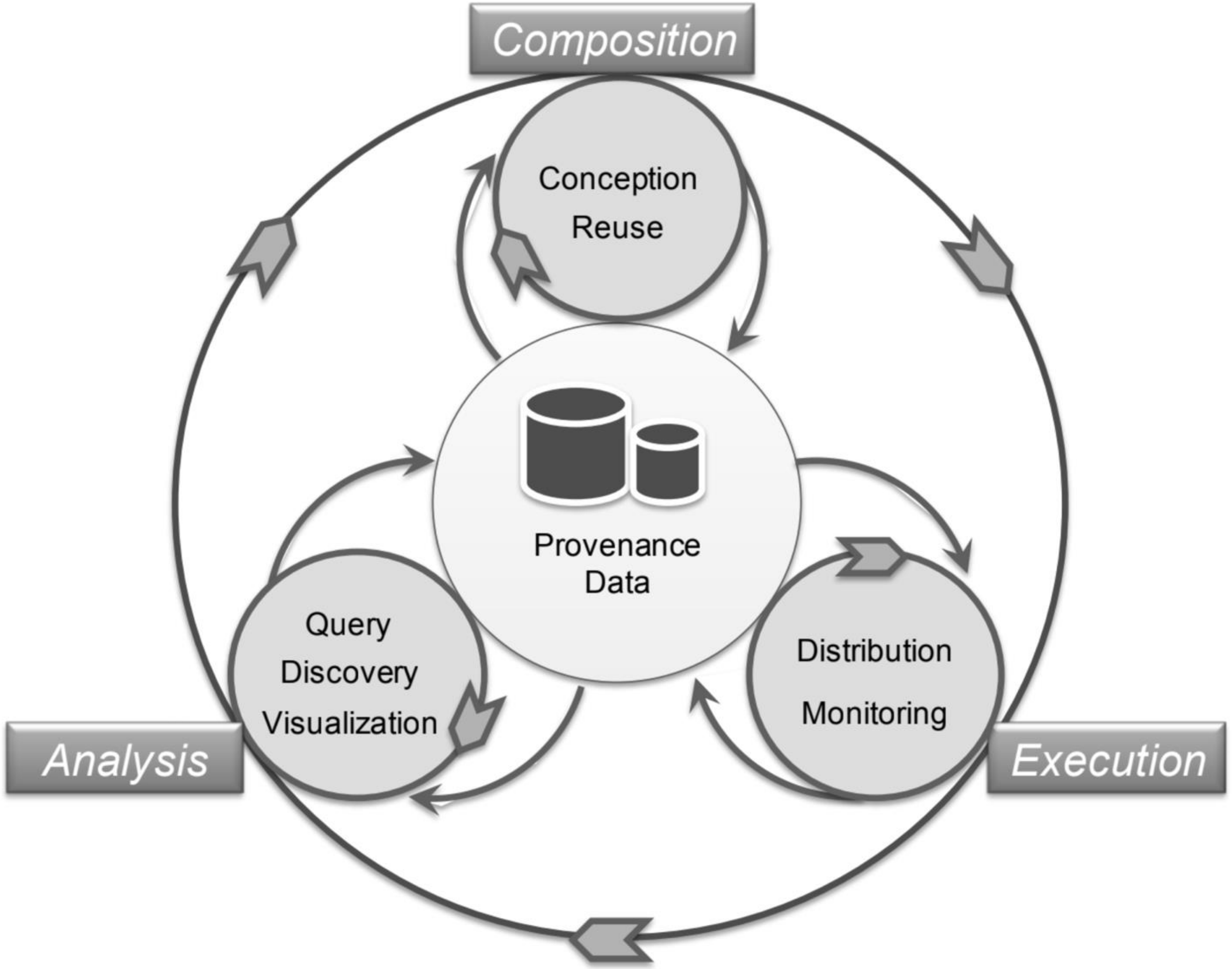
Workflow Provenance:

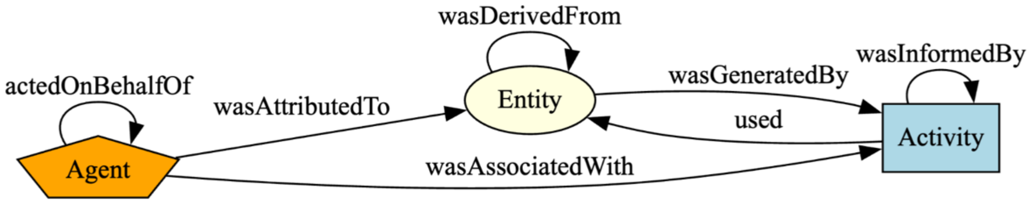
Structured record of workflow execution

Captures input and output data in a coherent dataflow

Captures task, related (non-AI) agents, and execution environment metadata

Marta Mattoso et al. "Towards supporting the life cycle of large scale scientific experiments." *International Journal of Business Process Integration and Management* (2010).





The W3C PROV model: foundational constructs behind a provenance database

The Role of Provenance in Agentic Scientific Workflows

Provenance *of* Agents

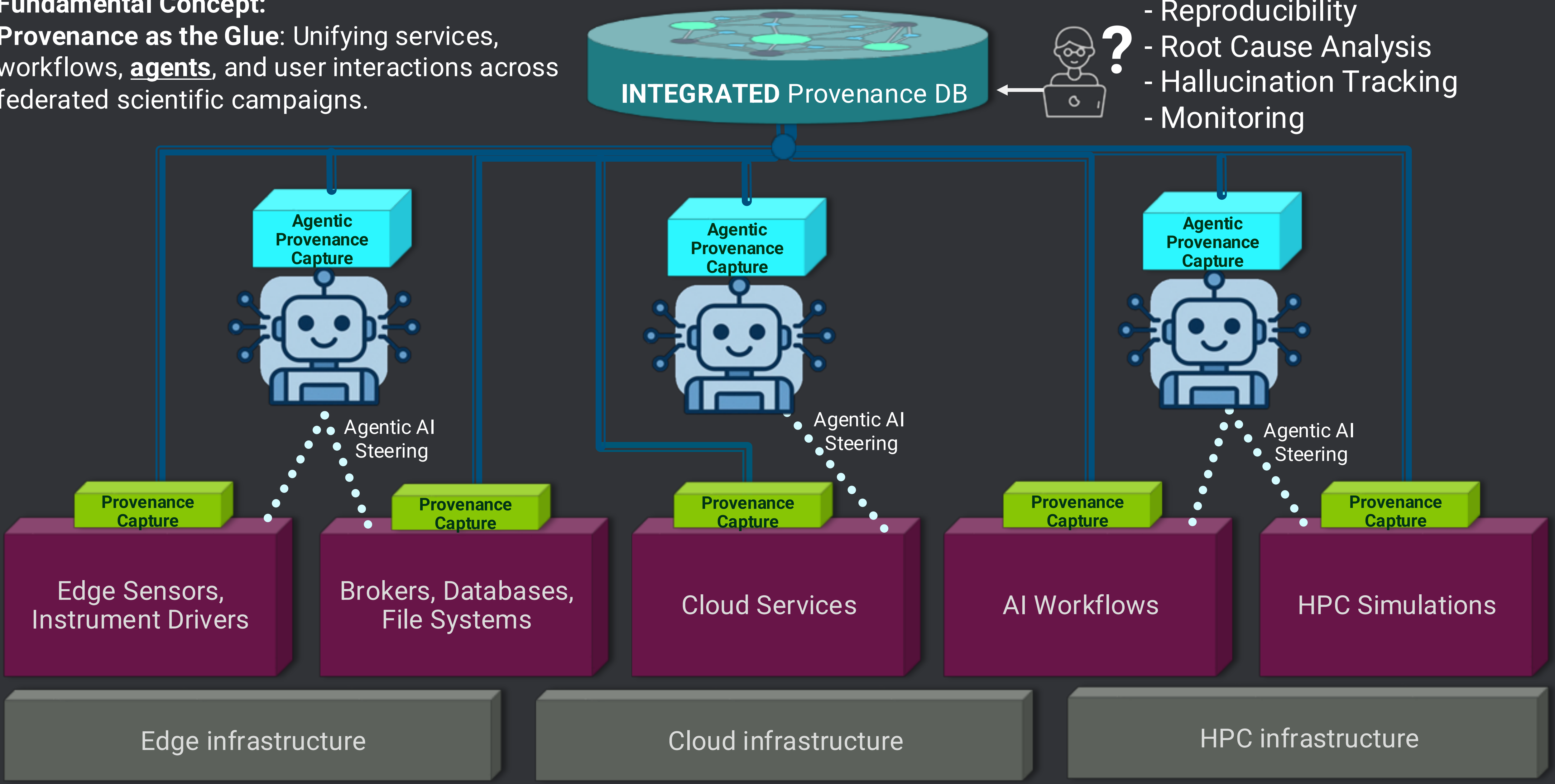
Capturing and contextualizing an agent’s decisions, actions, and interactions within a workflow to enable traceability, accountability, and impact assessment.

Provenance *with* Agents

Leveraging agentic AI as a natural language interface to provenance databases, enabling scientists to query and explore complex provenance data more easily and interactively.

Provenance keeps agents accountable, while agents make provenance accessible.

Fundamental Concept:
Provenance as the Glue: Unifying services, workflows, **agents**, and user interactions across federated scientific campaigns.



Edge-Cloud-HPC Continuum

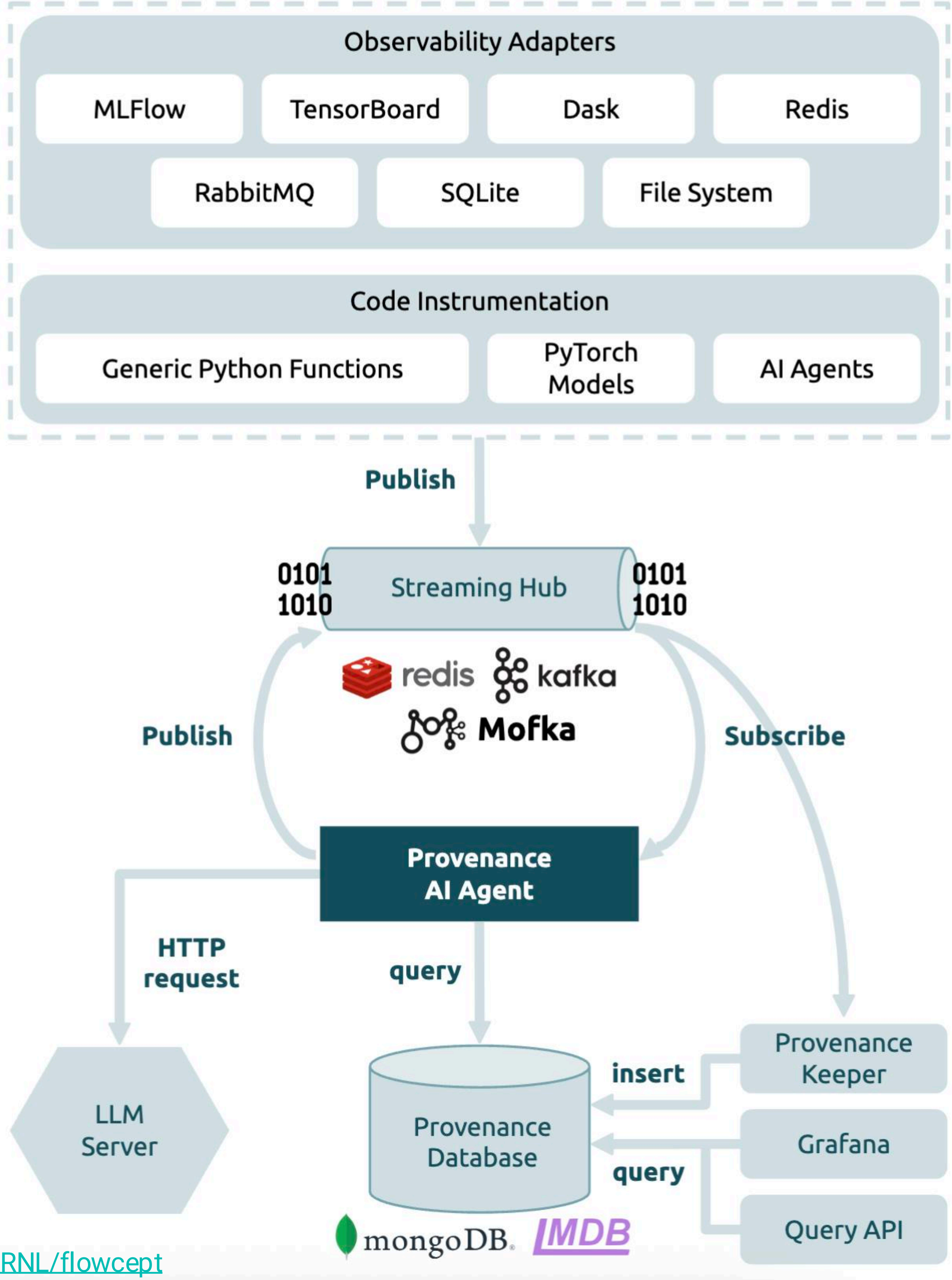
Workflow Provenance in the Computing Continuum for Responsible, Trustworthy, and Energy-Efficient AI R. Souza, S. Caino-Lores, M. Mattoso, R. Silva et al. IEEE International Conference on e-Science, 2024.

Flowcept: Unified Workflow Provenance Data Management

Distributed, Loosely-coupled, Flexible, Stream-Based Architecture

Provenance + Metadata + Energy indicators Capture via Code Instrumentation and Data Observability

Unified Runtime Data Access, from Science Labs to Supercomputers



R. Souza et al. Towards Lightweight Data Integration using Multi-workflow Provenance and Data Observability. 19th IEEE International Conference on e-Science, 2023.

https://github.com/ORNL/flowcept

Why Provenance WITH Agents?

- **Workflow Provenance:**

- Essential for reproducibility, anomaly diagnosis, experiment understanding
- Especially in **federated ECH** workflows
- Even more in **Agentic AI** workflows

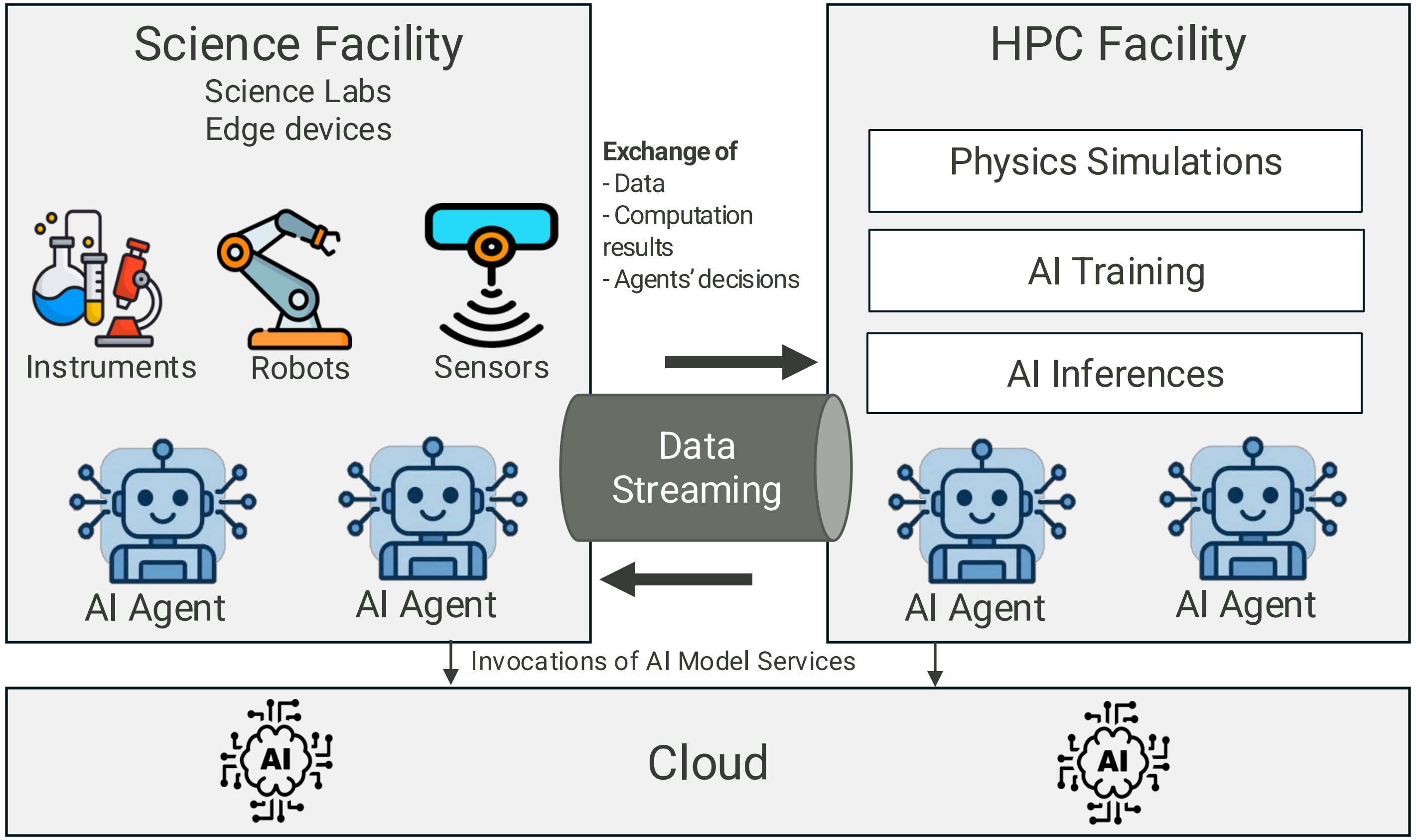
- **Problem:**

- At scale, provenance data are hard to analyze
- Current tools rely on custom scripts, structured queries, dashboards, and complex graph vis. tools

- **Goal**:

Bring scientists closer to runtime provenance data through natural language interaction.

Computing Continuum: Edge-Cloud-HPC (ECH) Workflows



OAK RIDGE National Laboratory

Problem and Challenges

- **Needs:**

- Democratize access to large prov. data via natural language ***during*** and ***after*** runs

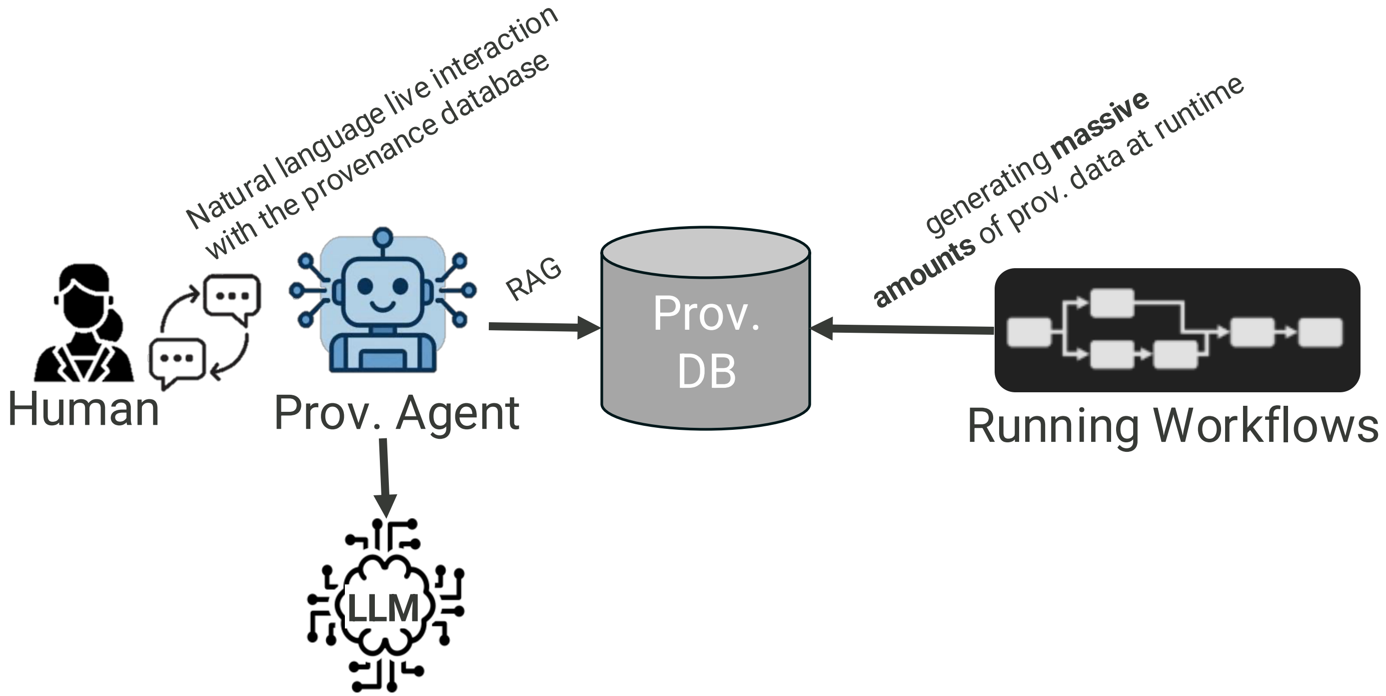
- **Challenges:**

- **Heterogeneous** data in multiple workflows
- Designing **a domain-agnostic system** that generalizes to multiple domains
- Large **streaming data** across **ECH**
- GenAI **hallucinations**
- **Limited LLM context windows**, even with the edge models

- **Solution**:

- **Metadata** and schema driven LLM agent that generates structured provenance queries
- **Modular architecture** and evaluation methodology to evaluate generalization

Basic Idea



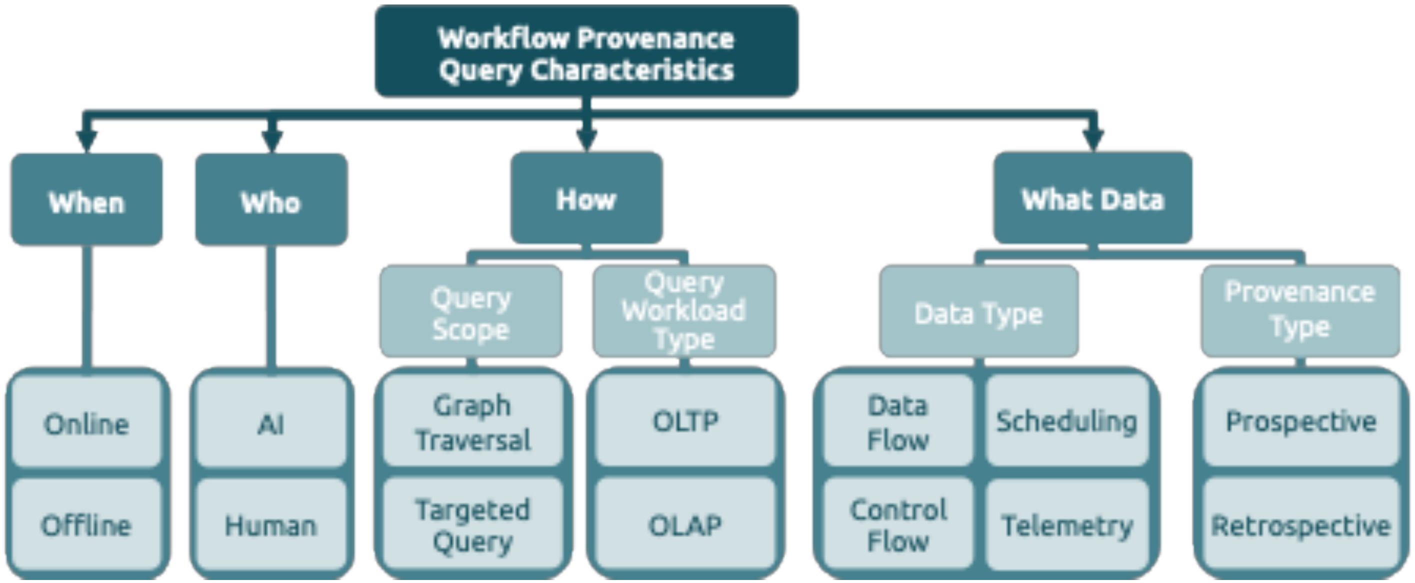
R. Souza, T. Poteet, B. Etz, D. Rosendo, A. Gueroudji, et al. LLM Agents for Interactive Workflow Provenance: Reference Architecture and Evaluation Methodology. Workflows in Support of Large-Scale Science (WORKS) co-located with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2025.

Prov. Query Taxonomy and Evaluation Methodology

- Use taxonomy to build a balanced ground-truth dataset (query + ideal answer per class)

Systematically tune prompts / RAG across diverse query classes, avoiding ad-hoc, overfitted fine-tunes

- Building domain-agnostic dataset aiming at generalization across various domains



Reusable recipe to evaluate and iteratively improve prompts for provenance queries

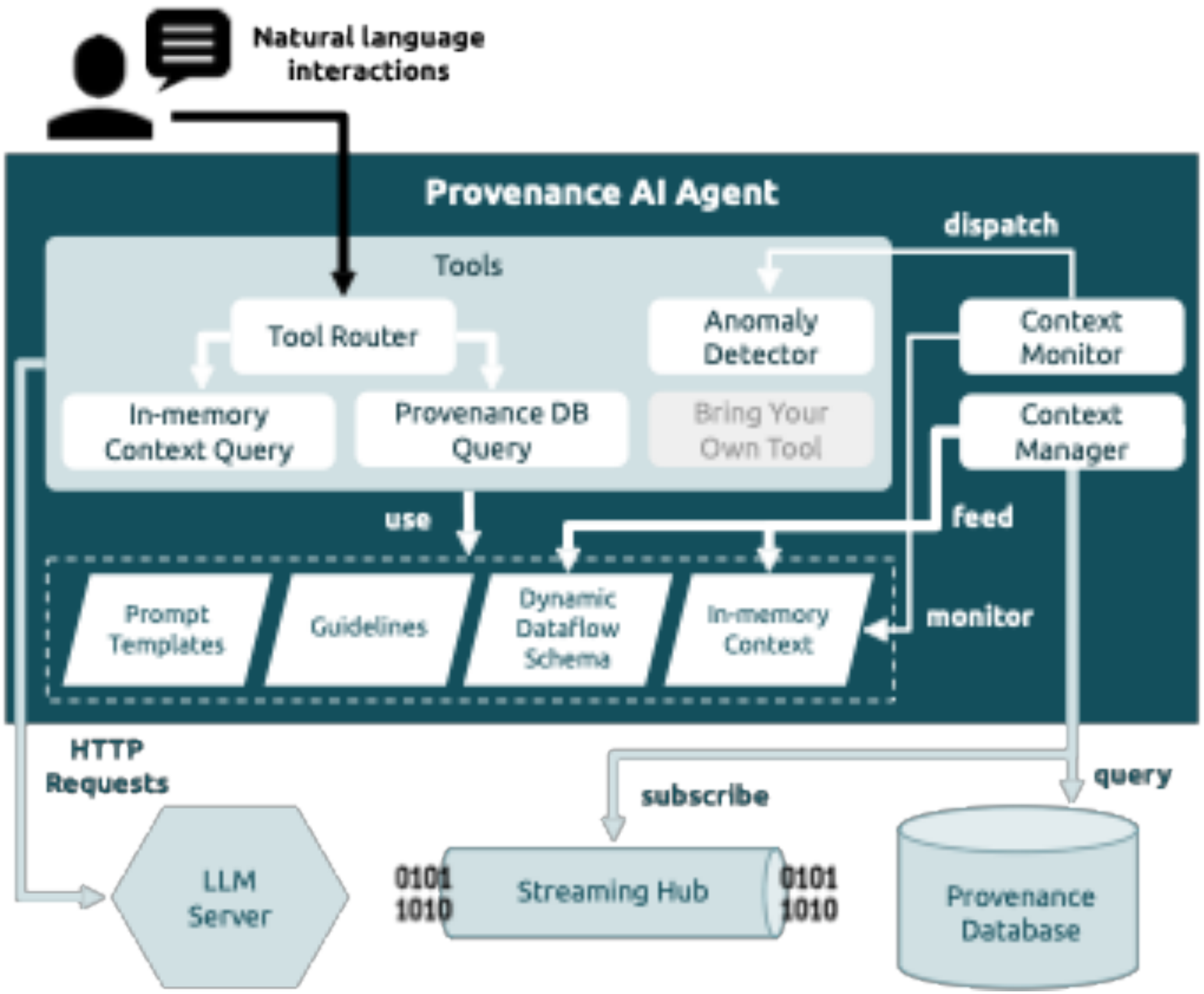
Oak Ridge National Laboratory

Prov AI Agent Architecture

Context Manager: connects to the streaming hub and builds the internal context (dynamic schema + values)

Tool Router & Monitors: route NL queries, detect anomalies; some tools run without LLM

All tools and LLM calls are **recorded** as provenance data



Metadata, Dynamic Schema, Query Driven

No LLM fine tuning, only prompt eng. + RAG

The LLM only knows a compact, token-efficient schema, not the actual prov. data

Prompt Structure:
- Schema = {**Static** common fields + **Dynamic** domain fields}
- Guidelines + Custom Guidance + FS

High-scalability: Scales with workflow structure (#activities, #parameters), **independent of #workflow tasks**, and supports sensitive data (no data export)

Incoming Raw Prov Messages -> Dynamic Schema Building (in agent context)


```
{
"task_id": "1753457858.952133_0_3_973",
"campaign_id": "0552ae57-1273-4ef8-a23b-c5ae6dd0c080",
"workflow_id": "4f2051b9-cfa3-4ef5-b632-907a3be06899",
"activity_id": "run_individual_bde",
"used": {
  "e0": -155.033799510504,
  "frags": {
     ”label": "C-H_3",
     "fragment1": "[H]OC([H])([H])[C]([H])[H]",
      "fragment2": "[H]"
   },
   "z0": 0.08026498424723788
},
"generated": {
    "bond_id": "C-H_3",
     "bd_energy": 98.64865792890485,
    "bd_enthalpy": 100.22765792890056,
    "bd_free_energy": 92.39108332890055
 },
"started_at": 1753457858.952133,
"ended_at": 1753457859.009404,
"hostname": "frontier00084.frontier.olcf.ornl.gov",
"telemetry_at_start": {"cpu": ["percent": 23.4]},
"telemetry_at_end": {"cpu": ["percent": 53.8]},
"status": "FINISHED",
"type": "task"
}
```


Domain-specific fragments

- While prov. data are streamed into the agent, a background thread updates its context
- **Tracks domain fields as inputs (used) or outputs (generated) and stores a few sample values per field**

Oak Ridge National Laboratory

Why Provenance *OF* Agents

Agentic Provenance Definition: "Systematic capture and representation of agents' **decisions, actions, interactions, and their effects within workflows**, supporting that agent behavior is traceable, accountable, **and connected to the broader provenance of data and tasks**."

Capture: Each time an agent runs, its metadata are recorded and linked with the broader workflow provenance.

W3C PROV+MCP: extending PROV for agentic workflows

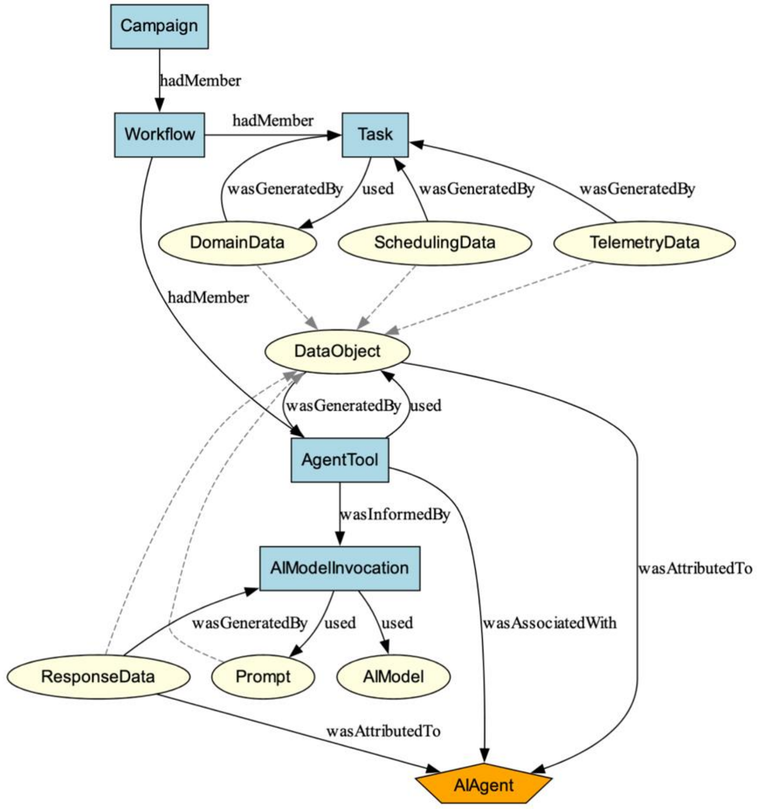


Fig. 3: PROV-AGENT: A W3C PROV Extension for Agentic Workflows. Dashed arrows represent *subClassOf*.

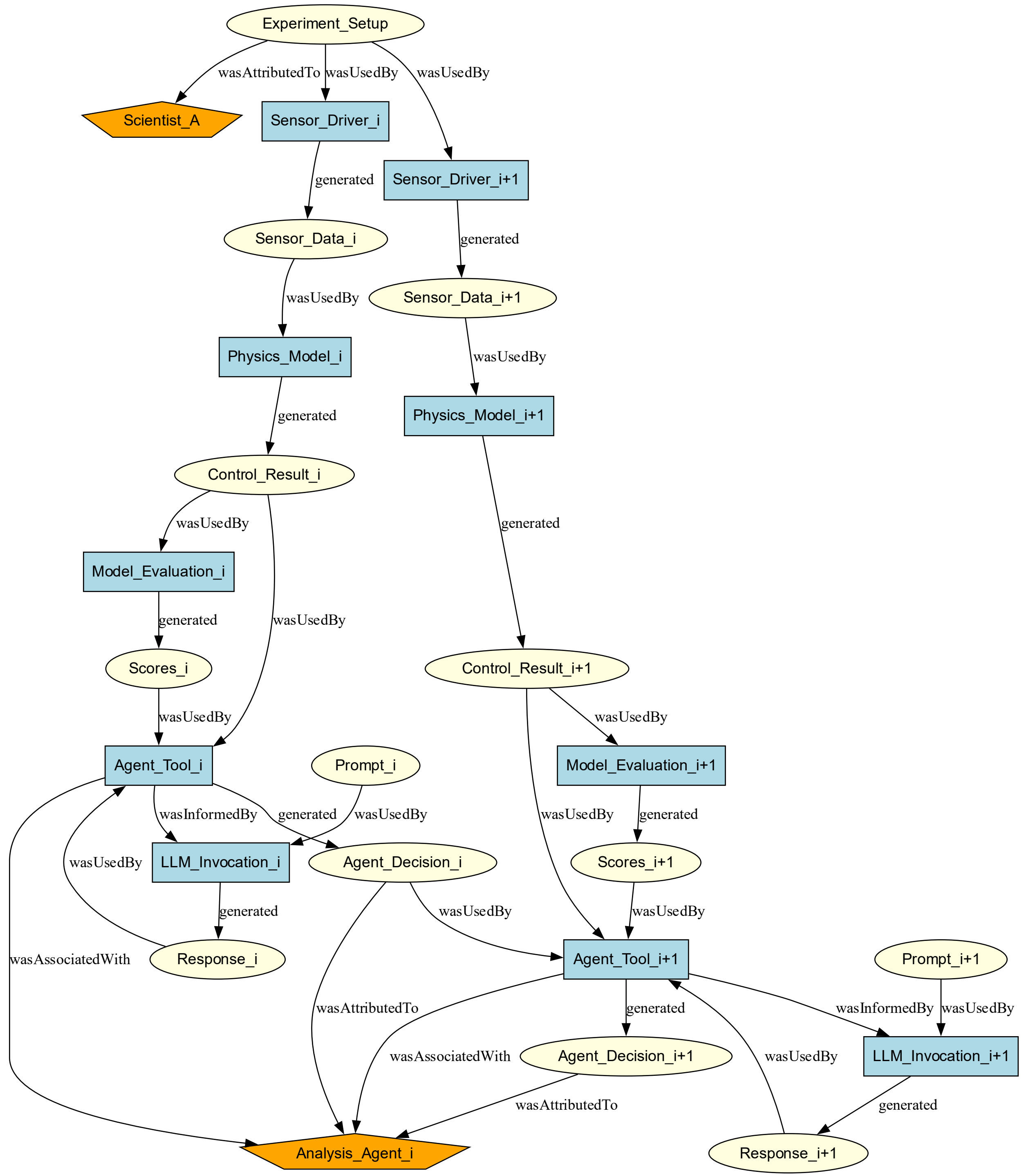
```python
from langchain_openai import ChatOpenAI
from flowcept import FlowceptLLM, flowcept_agent_tool

@mcp.tool()
@flowcept_agent_tool
def evaluate_scores(layer, result, scores):
    ...
    prompt = get_prompt(layer, result, scores)
    llm = FlowceptLLM(ChatOpenAI(model="gpt-4o"))
    response = llm.invoke(prompt)
    ...
    return ...
```

Fig. 4: MCP agent tool that invokes an LLM to assess physics model outputs. With the decorator `@agent_flowcept_task` and `FlowceptLLM` wrapper, agent tool and LLM invocation provenance are captured.

R. Souza *et al*. PROV-AGENT: Unified Provenance for Tracking AI Agent Interactions in Agentic Workflows. IEEE e-Science, 2025.

Agentic Provenance Graph

The resulting PROV-compliant data graph allows for tracking agents' decisions and assessing their downstream impact in the workflow, supporting accountability, transparency, and reproducibility



Oak Ridge National Laboratory

Experimental Evaluation

Experimental Evaluation

Ground truth Data Set: 20 curated natural language queries over synthetic, simple workflow.

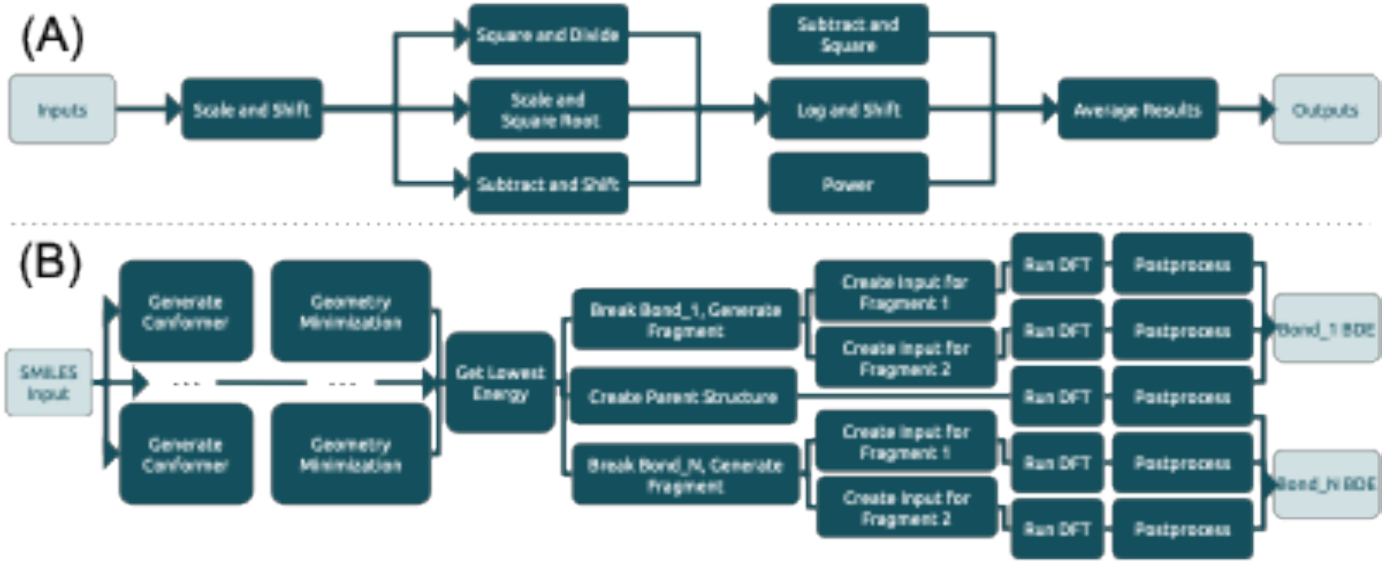
LLM as a judge evaluate generated query quality, not result set.

Goal is **not** to pick the “best” LLM, but to test if our approach generalizes across LLMs and domains

LLMs tested: LLaMA 3 (8B, 70B), GPT-4, Gemini 2.5 Flash Lite, Claude Opus 4.

Fine-tuned on a synthetic workflow and evaluated on other real workflows

Simple Synthetic Workflow



Real Computational Chemistry Workflow on the Frontier HPC System

OAK RIDGE National Laboratory

Main Findings Across LLMs

- LLM judges (GPT & Claude): similar results

- GPT & Claude strong performance

- LLaMA & Gemini higher variability

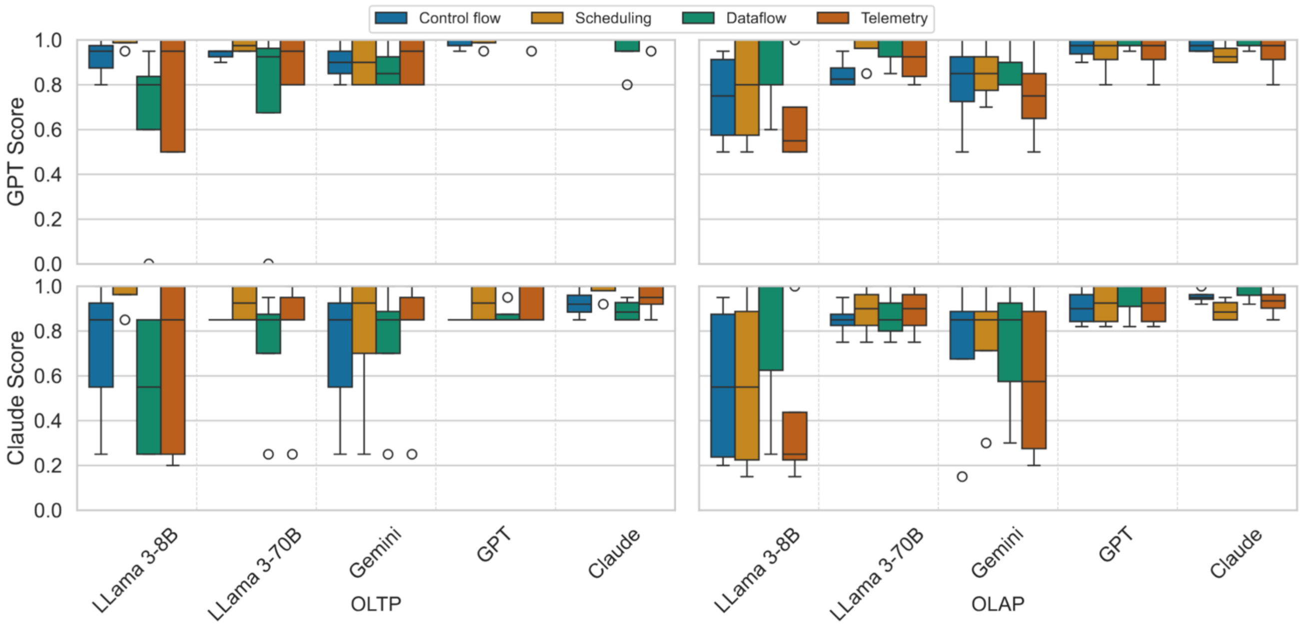
- Graph queries hardest in all models

- No LLM is one-size-fits-all

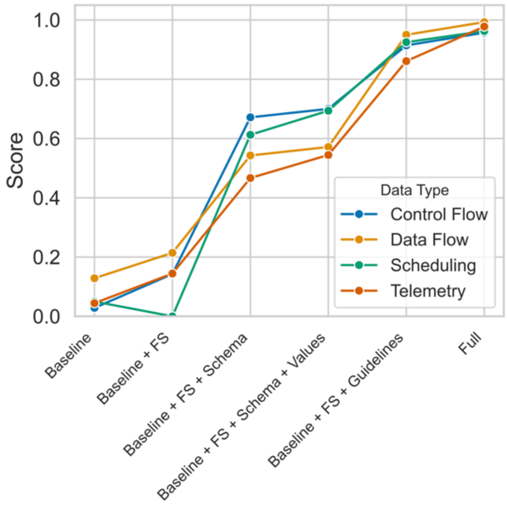
- Guidelines, schema, and few-shot examples: large boosts with low token overhead

- Response times stay within interactive bounds, about a couple of seconds.

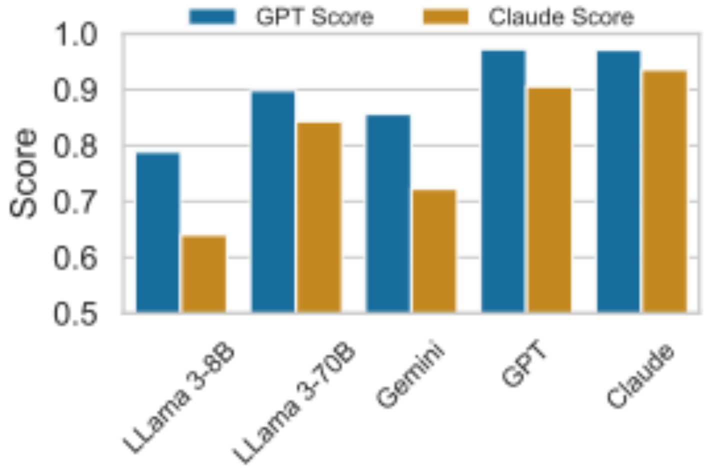
- Good generalization when running on real workflows



Query Classes Experiment



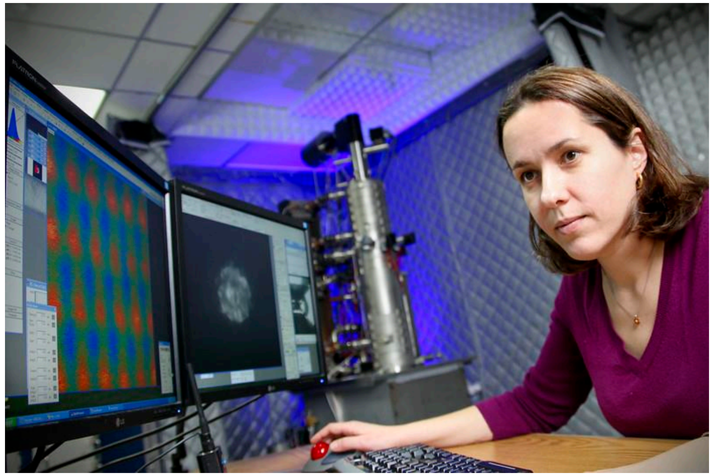
Prompt Parts

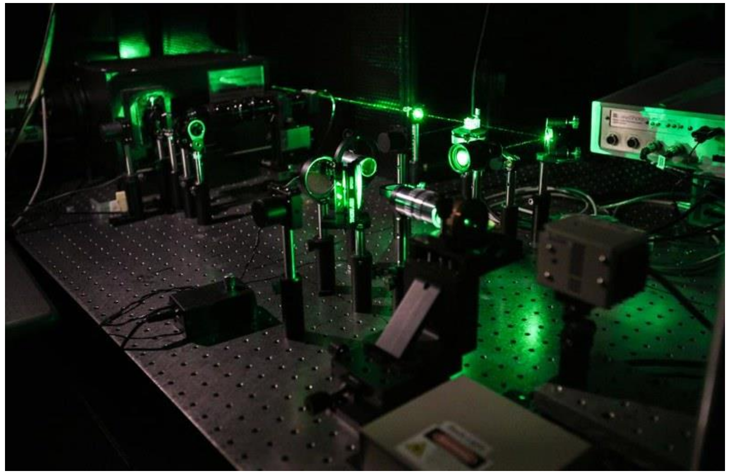


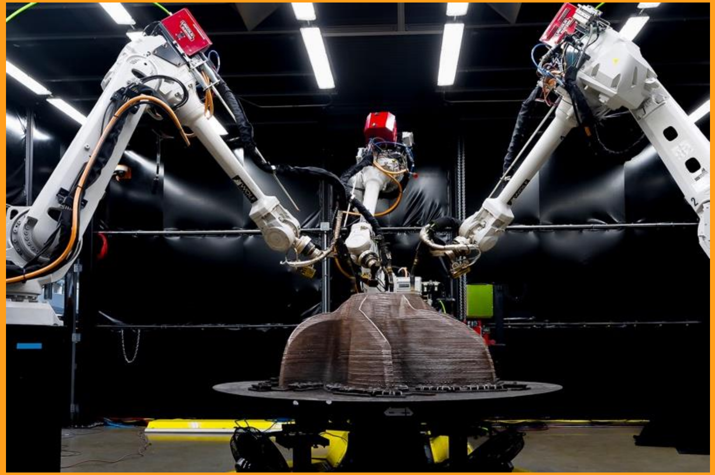
LLM as a Judge Trends

Use Case: Agentic Workflow for Adaptive Control of Metal 3D Printing

The DOE national labs have among the best scientific tools and facilities. AI agents will unlock new ways to use them







Edge-Cloud-HPC Infrastructure for Additive Manufacturing at ORNL

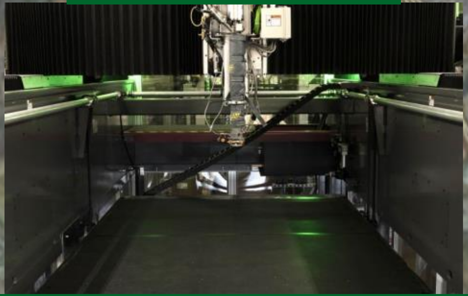
HPC Simulations

Cloud Services

CADES
Compute & Data Environment for Science

ORNL Main Campus

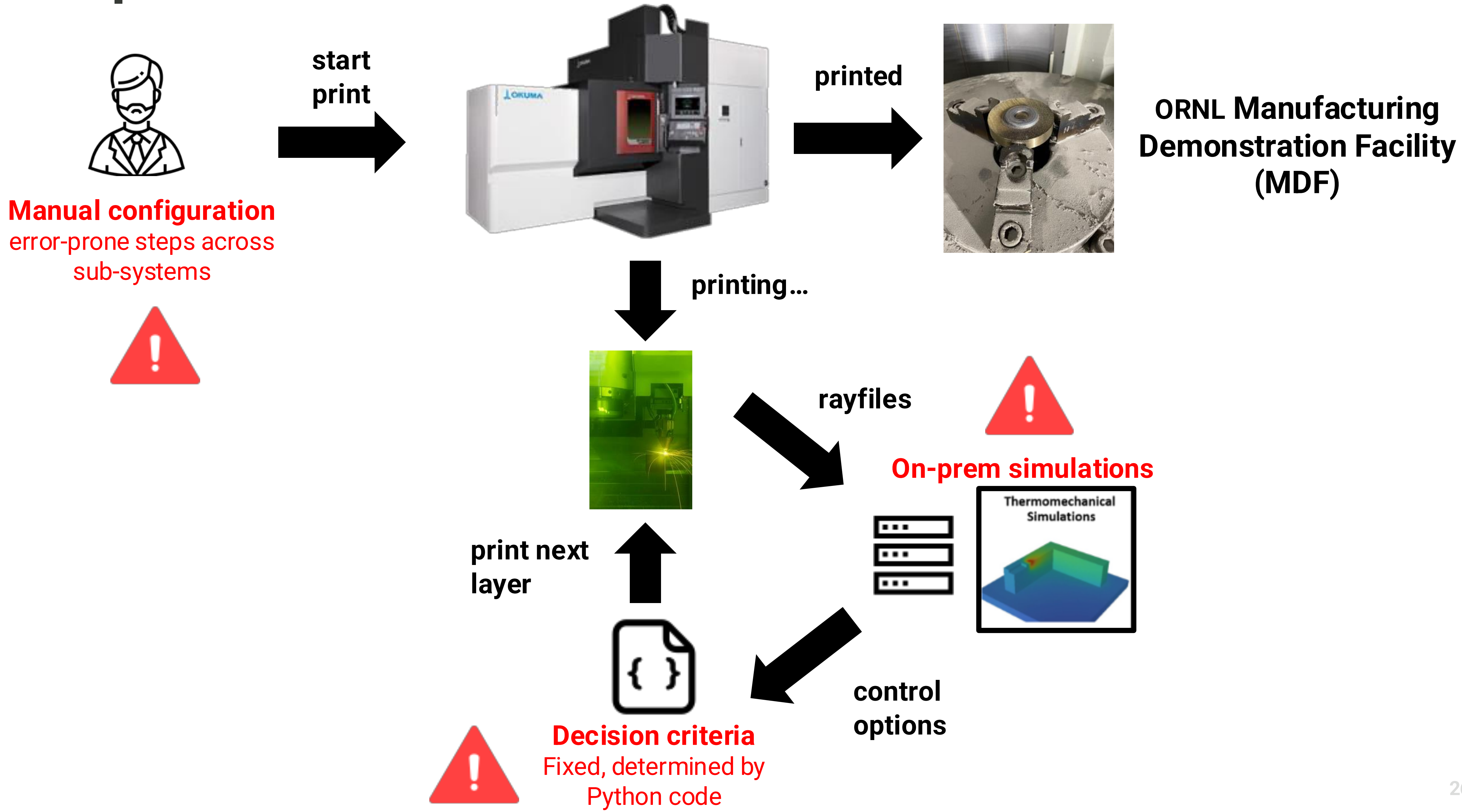
Edge: 3D Printing



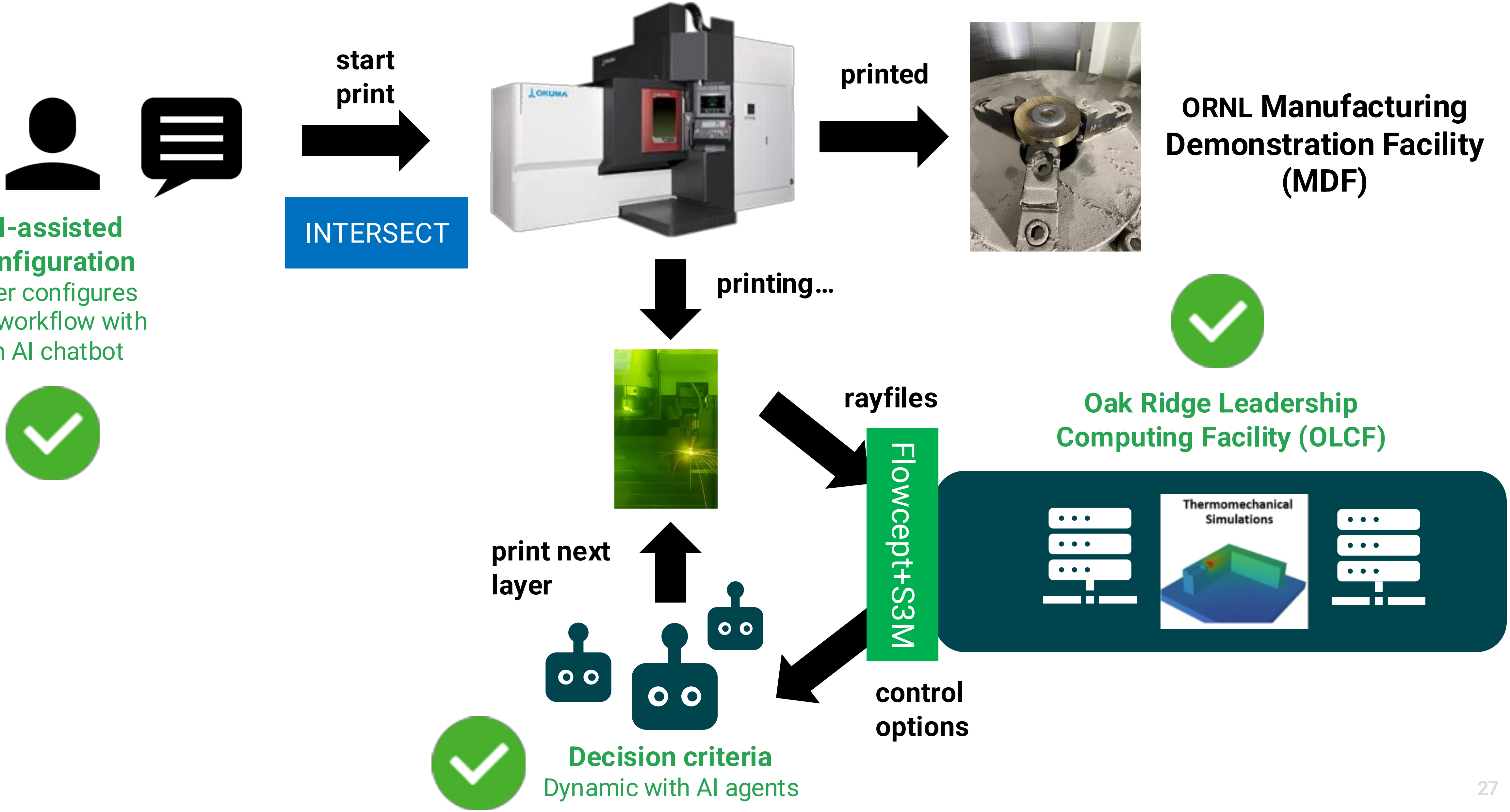
ORNL's Manufacturing Facility

Oak Ridge, Tennessee

Current experiment workflow requires manual configuration with on-prem simulations and fixed decision criteria

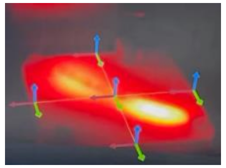


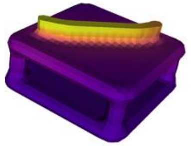
This work enables cross-facility experiments and explores AI Agents for autonomous decision-making



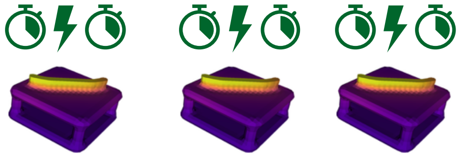
The agents make informed decisions based on human guidance, sensors, and simulations

Human guidance

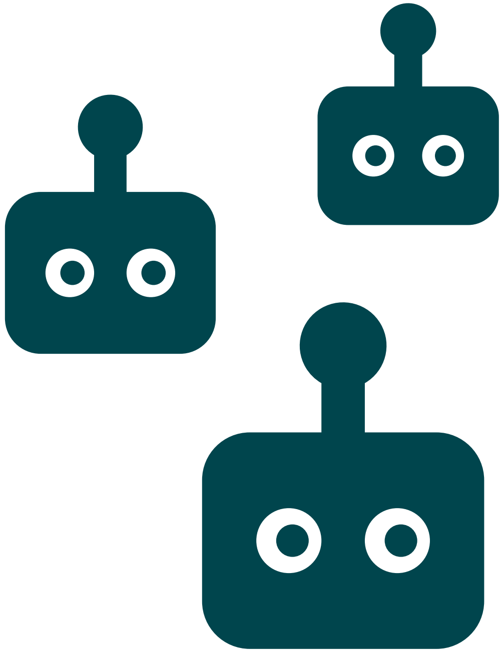




State estimate



Control options and simulated futures



Control decision

These agents can help you if you have ever been in a situation where…

…you base your plans on simulations, only to find that they miss crucial physics

…you know there's some simple logic to add, but there's no time to change code

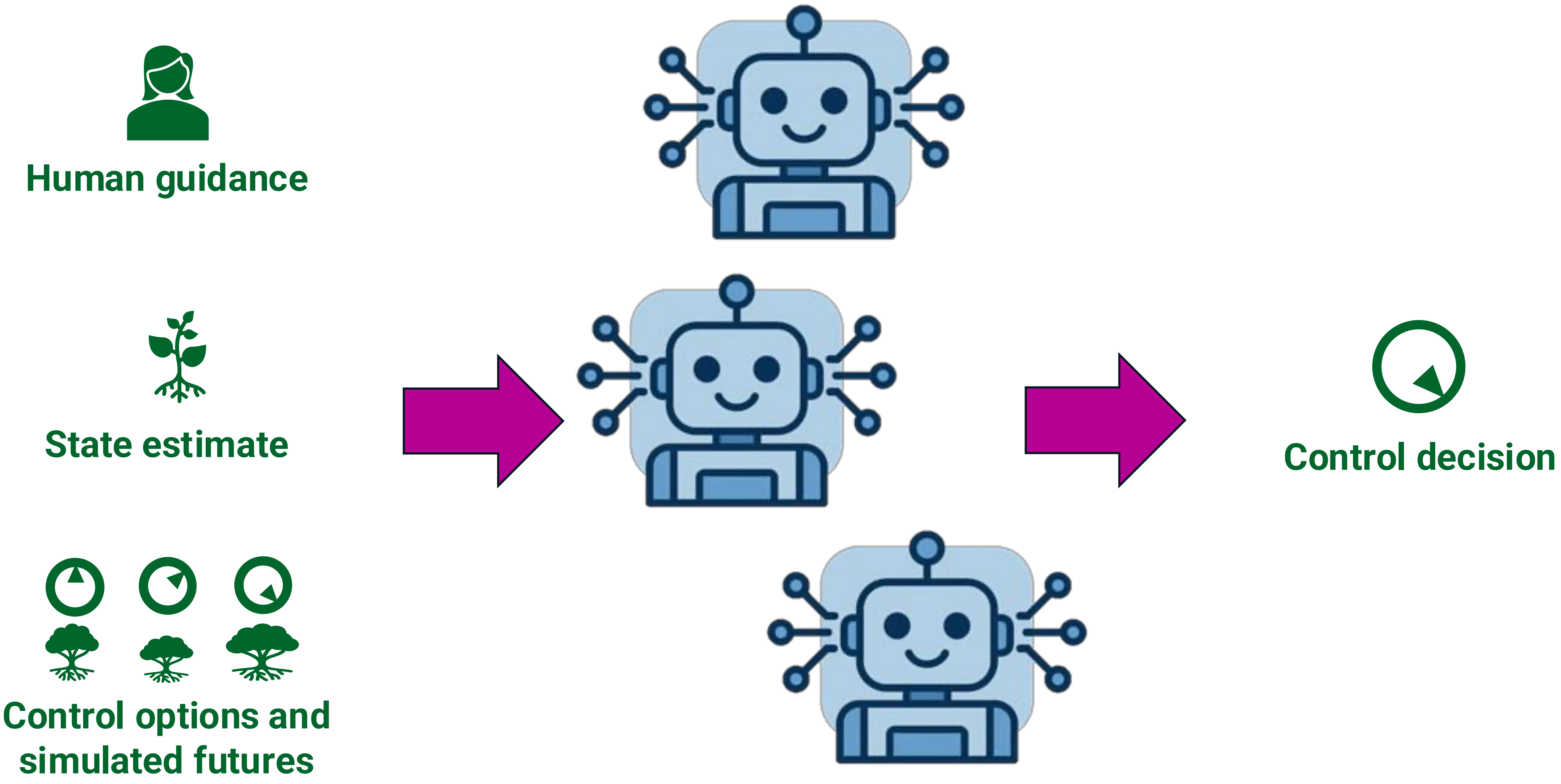
…something is going wrong, but you don't know what or why

"The simulation didn't tell us that printing in the same direction causes the part to droop"

"Actually, this paper says another transition can happen at 550C"

"Why didn't the robot move? Did the message go through?

Next you will see AI agents controlling an experiment, making real-time decisions with the help of simulations, **while Flowcept tracks provenance using the PROV-AGENT model**



MDF Demo Video

Video available at: https://zenodo.org/records/16801502

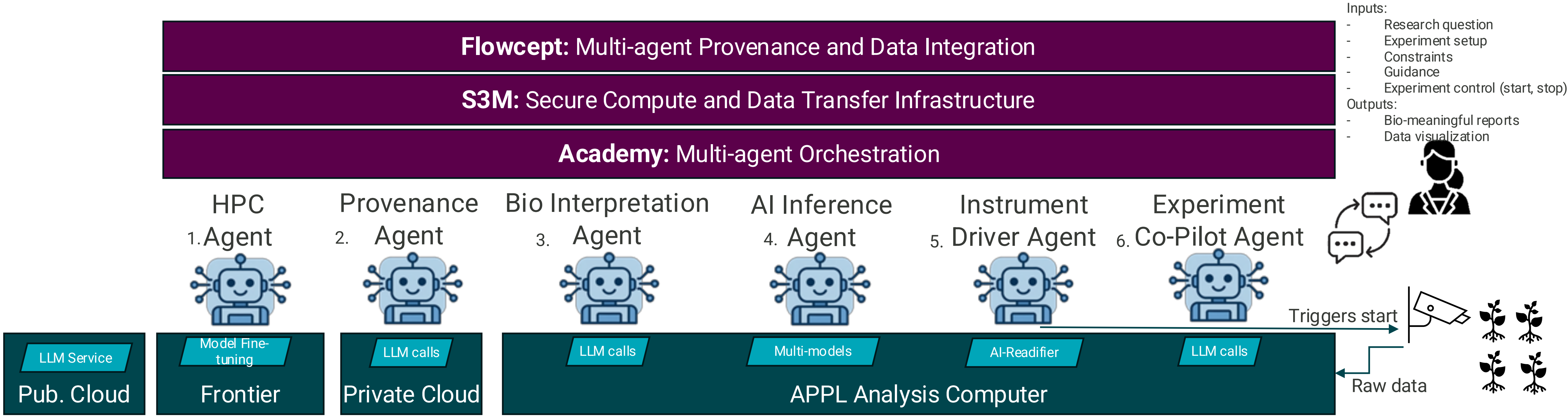
DEMO!

pip install flowcept[llm_agent]
flowcept --start-agent

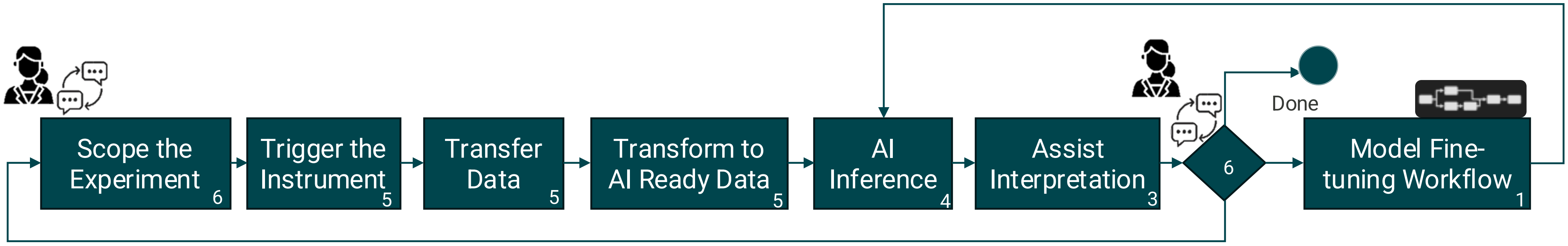
https://github.com/brianetz/CompChem/blob/main/src/bde_workflow.py

Next Steps: Integrating across various DOE initiatives

American Science Cloud (AmSc) and Orchestrated Platform for Autonomous Laboratories (OPAL)



Long-term Vision Workflow



Begin next iteration

Match the numbers with the agents above

Final Remarks

Provenance *Of* Agents

Ensuring Trust:

Capturing agent decisions, actions, and interactions as provenance enables accountability, transparency, and reproducibility in agentic workflows.

Provenance *With* Agents

Democratizing Access

Leveraging agentic AI as an interface transforms complex provenance databases into natural, interactive tools for scientists.

Real-world scenarios

Demos

Across additive manufacturing and chemistry workflows using ORNL/OLCF supercomputers, provenance acts as both safeguard and enabler, making AI-driven discovery more reliable and usable.

- The approach fine-tuned on a synthetic, simple workflow generalized well to real, complex workflows that run in ECH environments in other scientific domains
- The Prov Agent architecture, prompting techniques, and evaluation methodology are generic and could be reused by any other system that manages provenance data
- LLMs add to the scientific analysis toolkit, but will not replace existing methods

workflows community initiative

OAK RIDGE National Laboratory

Argonne NATIONAL LABORATORY

Dec 10, 2025 – Workflows Community Initiative

O B R I G A DO
For Agents and with Agents: The Role of Provenance Data in Agentic Workflows

Presenter: Renan Souza, PhD

Research Scientist
National Center of Computational Sciences
Oak Ridge National Laboratory

U.S. DEPARTMENT of ENERGY

ORNL IS MANAGED BY UT-BATTELLE LLC
FOR THE US DEPARTMENT OF ENERGY

ACKNOWLEDGEMENTS


Amal Gueroudji

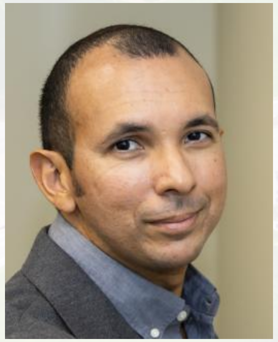

Daniel Rosendo


Brian Etz


Woong Shin


Stephen DeWitt


Fred Suter


Rafael F. da Silva

BACKUP

Main Contributions

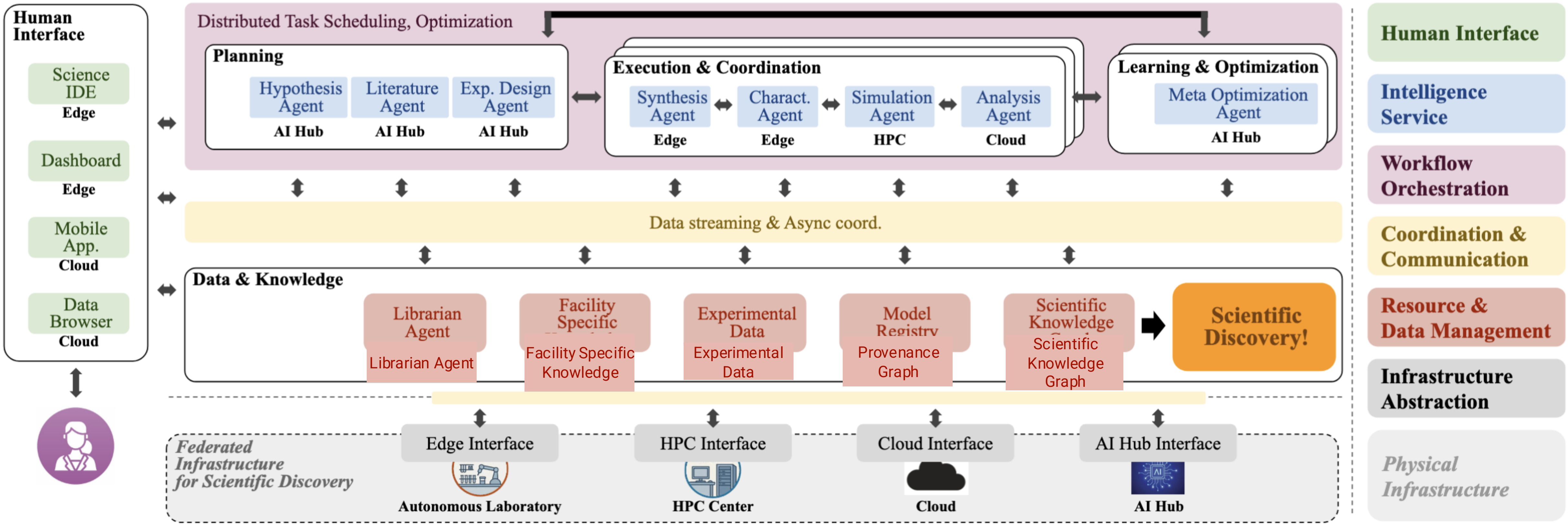
- Evaluation methodology for LLM-based provenance interaction
- Reference architecture for a provenance AI agent
- Open-source implementation on Flowcept
- Experiments with multiple LLMs + Exp on Frontier + Live demo

The Anatomy of an LLM Prompt to the Prov. AI Agent

Role & Expertise	You are an expert in HPC workflow provenance data analysis…
Job	You are going to generate a structured query to answer a user query in natural language
Static Fields	The prov. data have the following static fields: *task_id*, *workflow_id*, …
Dynamic Fields	The prov. data also contain inputs (in the used field) and outputs (in the generated field). These are the INPUT fields: {INPUT_FIELDS} These are the OUTPUT fields: {OUTPUT_FIELDS}
Example Values	For each of the dynamic fields above, the agent keeps track of a few example values in its context
Query Guidelines	Utilize the field *started_at* to sort by the timing of the tasks
Few Shot Examples	A few pairs of (Natural Language Query, Expected perfect structured Query)
Custom User Guidance	App-specific guidance that will be especially considered when generating the query
Output Formatting	Return only a valid structured query, DO NOT say anything else.
Normalized User Query	The actual *normalized* user query in natural language

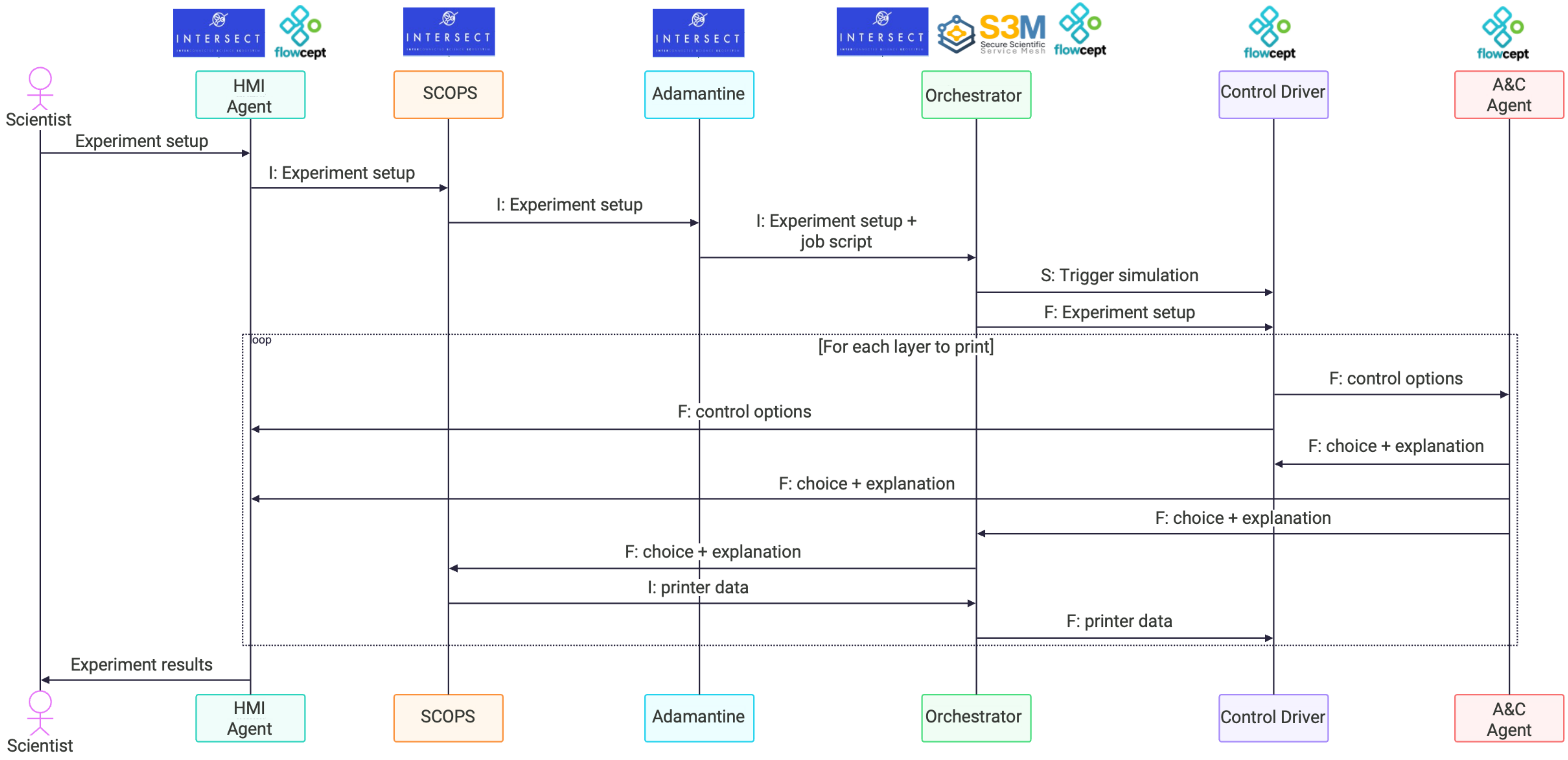
Raw prompt: https://github.com/flowcept/FlowceptAgent-WORKS25/blob/main/raw_prompt_example.txt

Full end-to-end Architectural Vision for Agentic Scientific Workflows



W. Shin, R. Souza, D. Rosendo, et al. The (R)evolution of Scientific Workflows in the Agentic AI Era: Towards Autonomous Science. 2025. Best paper on WORKS'25@SC.

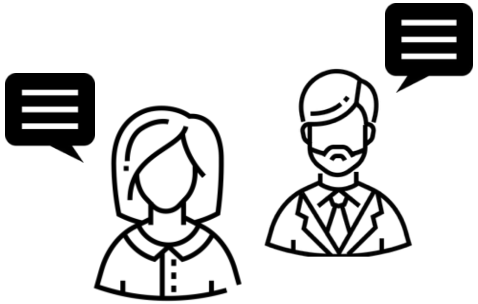
Simplified sequence diagram shows interactions during the demo



Current experiment workflow requires manual configuration with on-prem simulations and fixed decision criteria

*"Complex parts can take **months to print** and the **cost of a mistake** in the middle of a print is high."*

*"True, and we **learn a lot during the print** that should influence our **control decisions.**"*

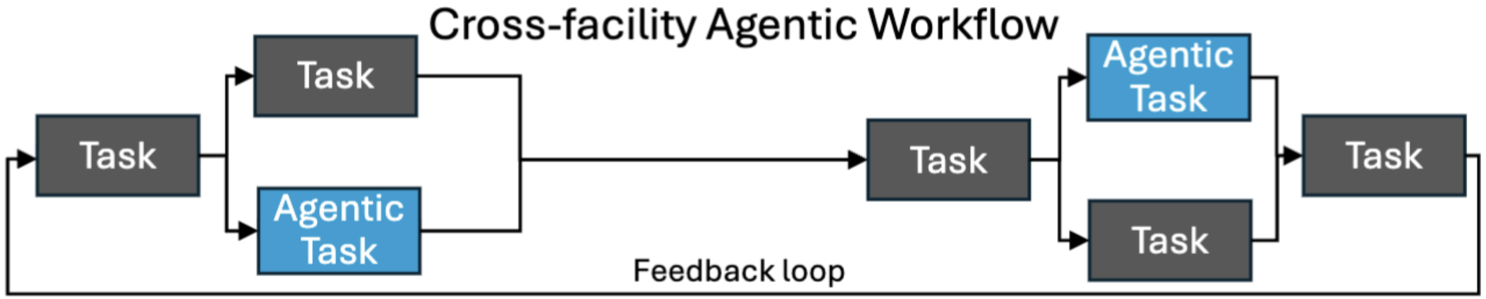


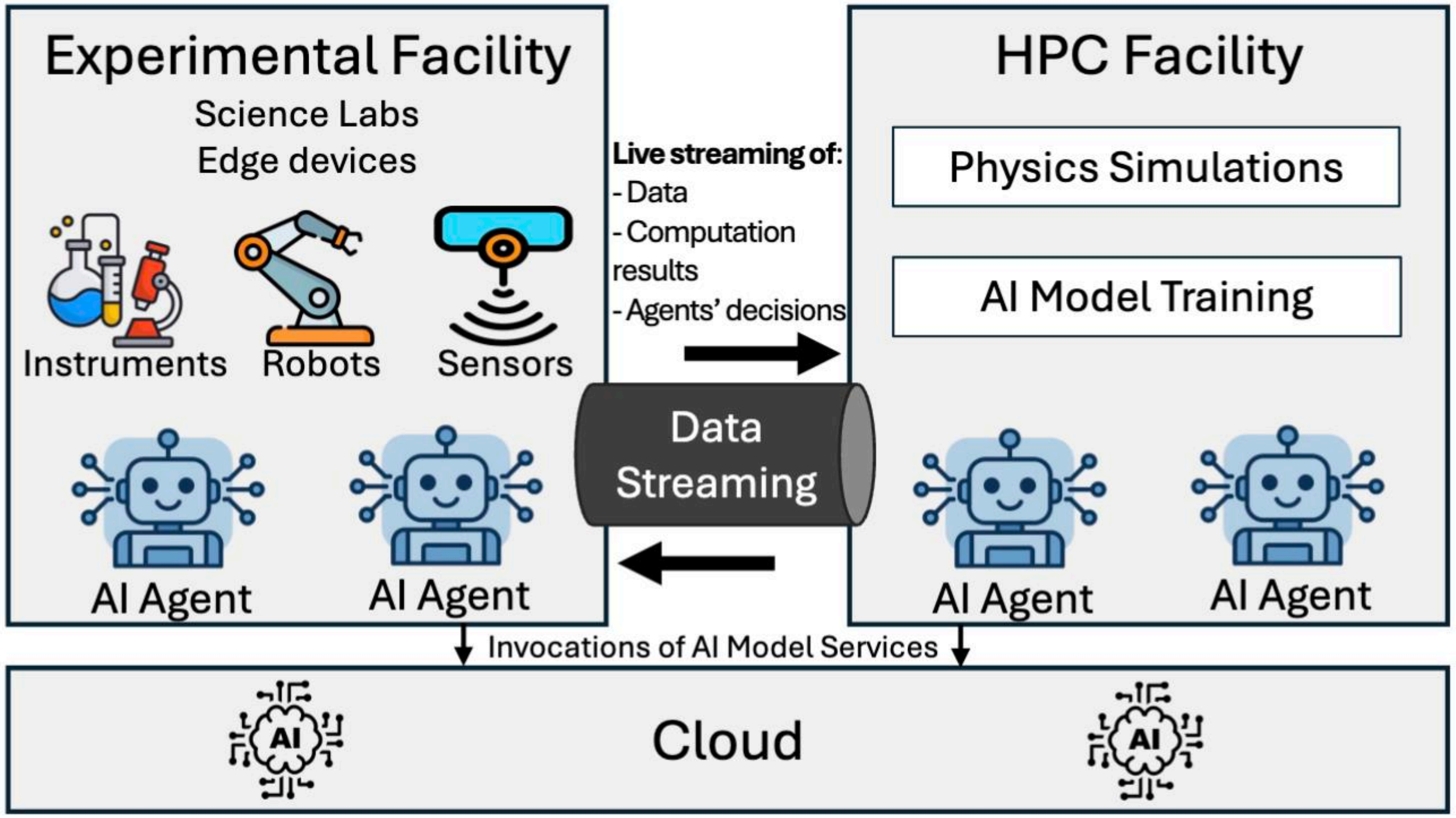
*"We definitely need a **dynamic** approach, where we could **adapt** control decisions **during the print**."*

Agentic Tasks in Workflows

Definition: workflow tasks that employ one or more GenAI services to make a decision or perform an action

They are chained into other agentic or non-agentic tasks in the workflow





Oak Ridge National Laboratory