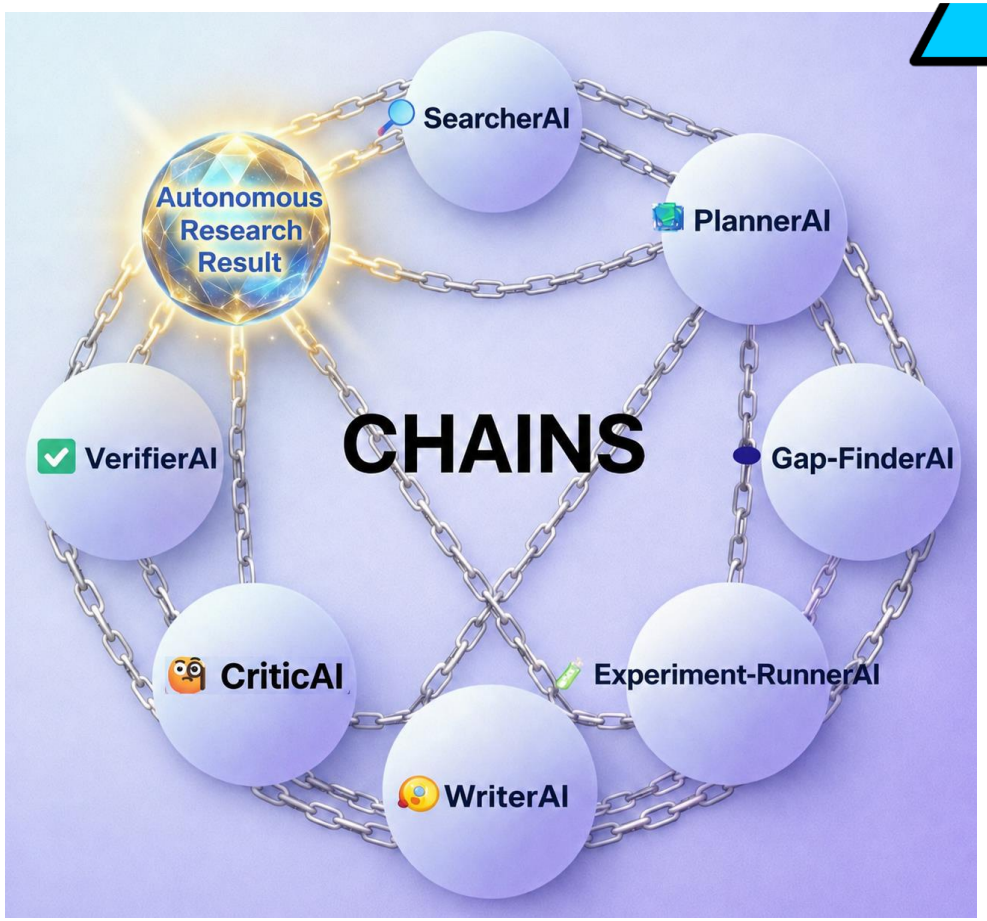
Collaborative **H**uman-"**AI** age**N**t **S**warm" (**CHAINS**) for Conducting Scientific Research



Authors:

Wilbert Villalobos, Yulia Kumar, J. Jenny Li, Dov Kruger, and Jose Marchena

Presenters:

Wilbert Villalobos:

- B.S in Computer Science
  - Undergraduate LLM and AI Researcher
- Full time employee as an Associate Systems Engineer at Northrop Grumman Corporation Mission Systems
- Co-author of 1 publication under Springer Nature
- Presented research at multiple conferences:
  - Great Minds in STEM
  - Richard Tapia Conference
  - ICICT Internation Conference
- Email: villalow@kean.edu

Jose Marchena:

- B.S. in Computer Science
- Pursuing a Master's degree in Computer Science at Kean University
- Graduate Assistant and AI Researcher
- Research experience in AI, Machine Learning, and RAG systems
- Co-author of 5 research publications in international conferences including Springer chapter book
- Presented research at national conferences including GMiS 2025
- Email: marchenj@kean.edu

Table of Contents

Introduction & Motivation

- Current autonomous AI systems for science often act as "black boxes" with obscured intermediate decisions
- Key Issues:
  - Generation of unsupported claims.
  - Lack of auditability and reproducibility in loosely orchestrated workflows.
  - High-risk actions performed without explicit oversight.

- The CHAINS Solution:

Decomposes the research process into **task-specialized agents** with a structured run manifest and explicit **Human-In-The-Loop (HITL) checkpoints**

Background & Definitions

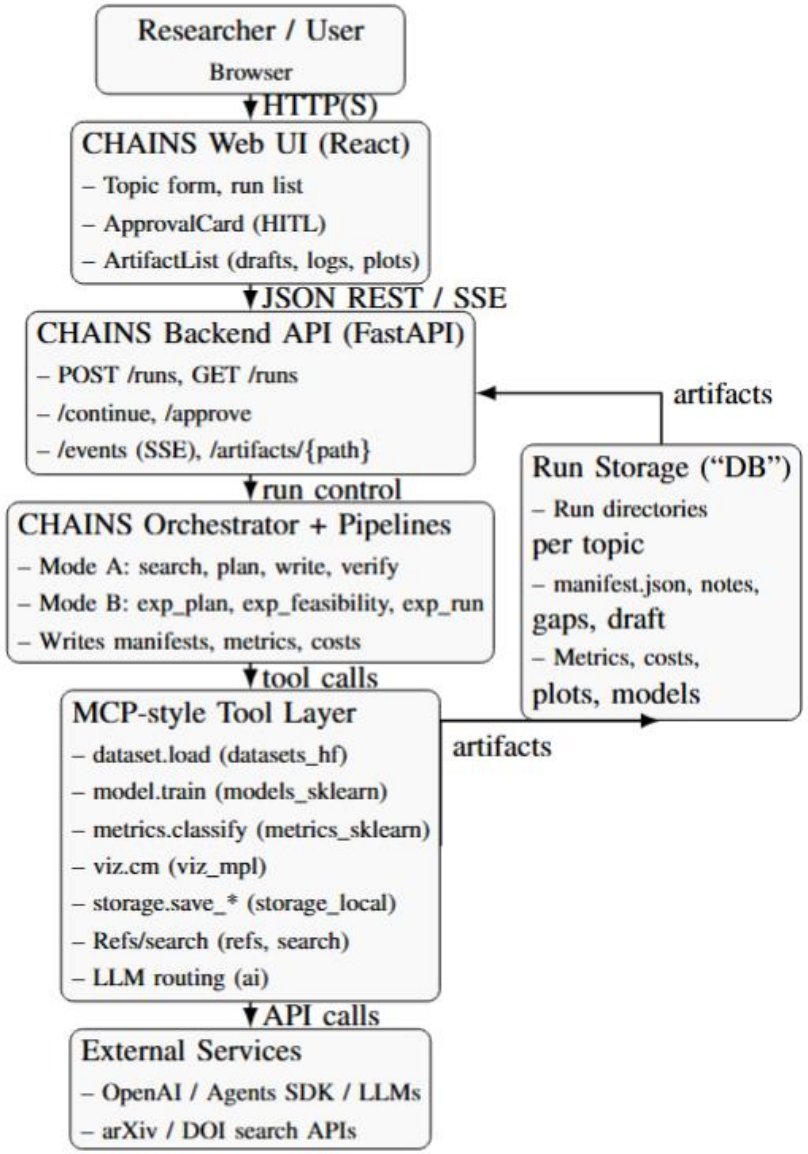
- **Agentic AI:**
  - Specialized AI agents collaborating to perform complex tasks.
- **Artifact-Driven Workflow**
  - Systems where every intermediate output (notes, plans, drafts) is saved as an inspectable artifact (e.g., JSON or Markdown)
- **Human-In-The-Loop (HITL):**
  - A governance model where humans retain the authority to approve, decline, or continue agent runs at critical checkpoints.
- **Deterministic Verification:**
  - Using rule-based (non-LLM) checks to ensure structural and citation integrity

Research Questions

RQ1 (Operationalization):
How efficiently can CHAINS turn knowledge gaps into **feasible micro-experiments?**

RQ2 (Transparency):
Does CHAINS's artifact-driven workflow improve transparency and reproducibility compared to black-box LLM systems?

RQ3 (Trust):
How do **Human-In-The-Loop (HITL) checkpoints** influence safety, failure prevention, and user trust?



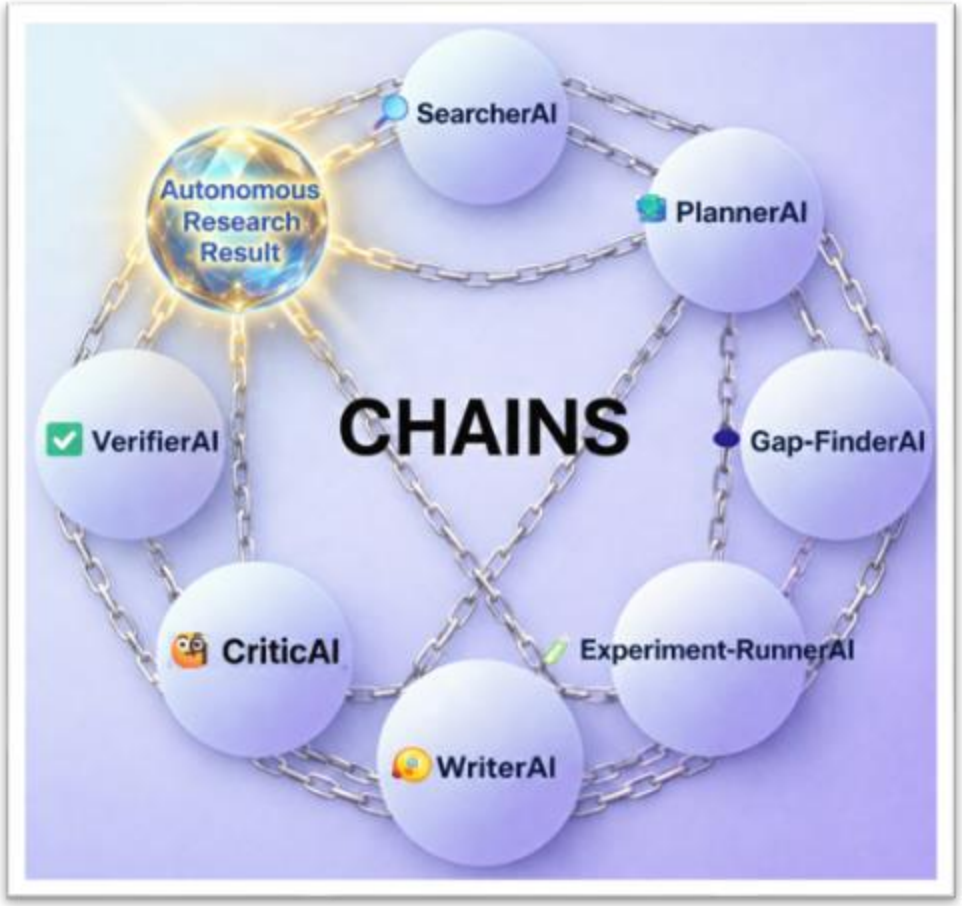
System Architecture:

- **Frontend**
  - React + Tailwind dashboard
- **Backend**
  - FastAPI
  - Uvicorn
- Core
  - OrchestratorAI
- **External Services**
  - OpenAI APIs
  - arXiv
  - MCP-style tools
- **Storage**
  - run manifests
  - artifacts
  - metrics

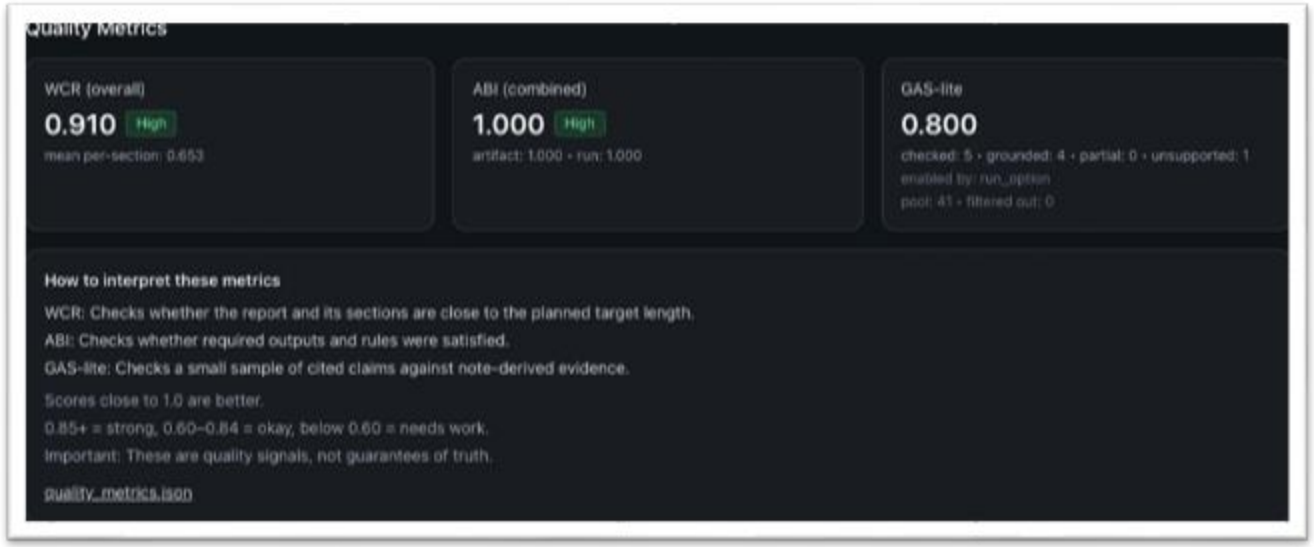
Multi-Agent Workflow

Agent Roles:

- **PlannerAI:** Designs high-level research and writing plans.
- **SearcherAI:** Queries arXiv and converts results into structured notes.
- **GapFinderAI:** Analyzes notes to identify 1–3 specific knowledge gaps.
- **ExperimentRunnerAI:** Executes micro-experiments and saves metrics (e.g., confusion matrices).
- **WriterAI:** Drafts the research paper based on the collected evidence.
- **CriticAI & VerifierAI: Review the draft for unsupported claims and structural compliance.**



Quality Metrics



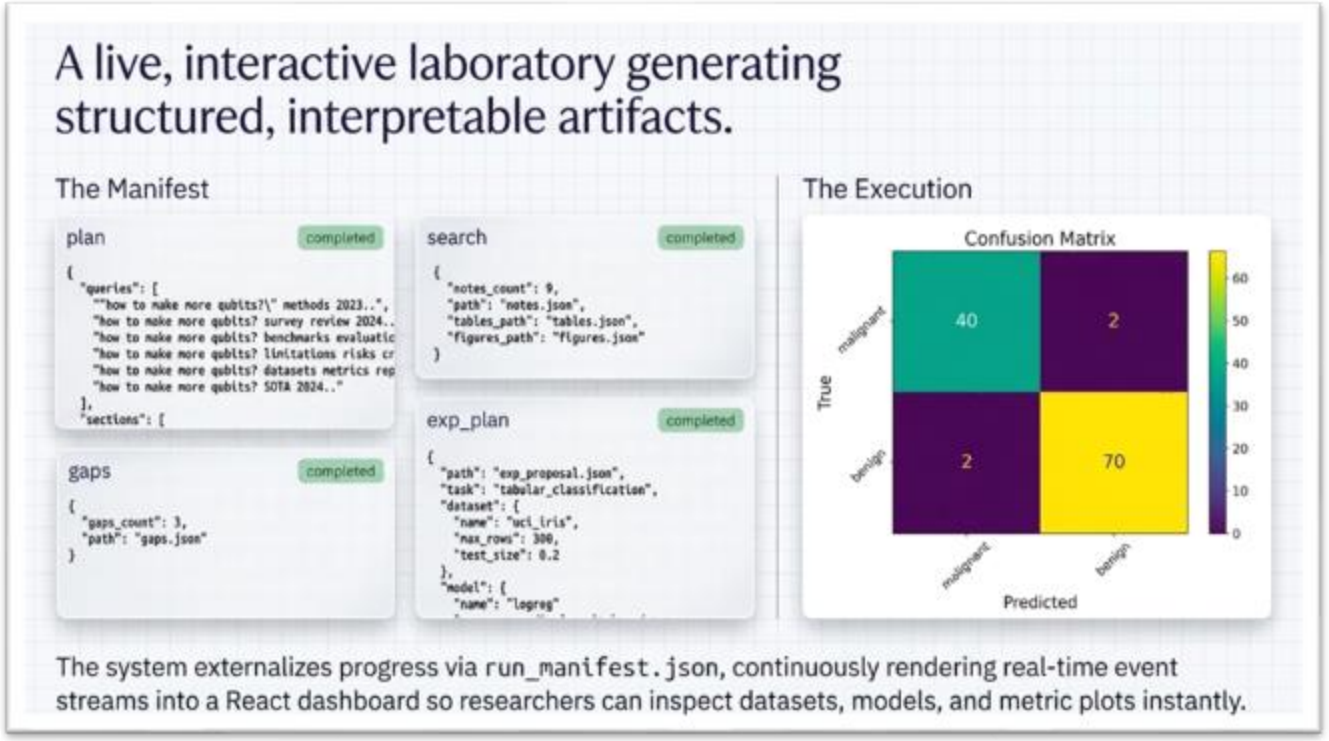
WCR (Word Count Ratio): Measures length conformity against the target plan.

ABI (Adherence to Binary Instructions): Fraction of satisfied structural requirements (e.g., presence of required sections).

GAS-lite (Grounding Accuracy Score-lite): A sampled audit checking if cited claims match the evidence in literature notes.

A decentralized swarm where distinct agents handle specialized scientific task

Results & Findings



Performance: CHAINS demonstrated more consistent completion and structured output than LLM-only baselines.

The "Paper-Shaped Text" Problem: LLM-only systems often produce non-executable text that *looks* like a paper but lacks structure; CHAINS ensures **structured execution**.

Improved Auditability: Run manifests and typed artifacts allow for easier failure localization and replaying of runs.

Safety: HITL gates prevent wasteful or unsafe executions

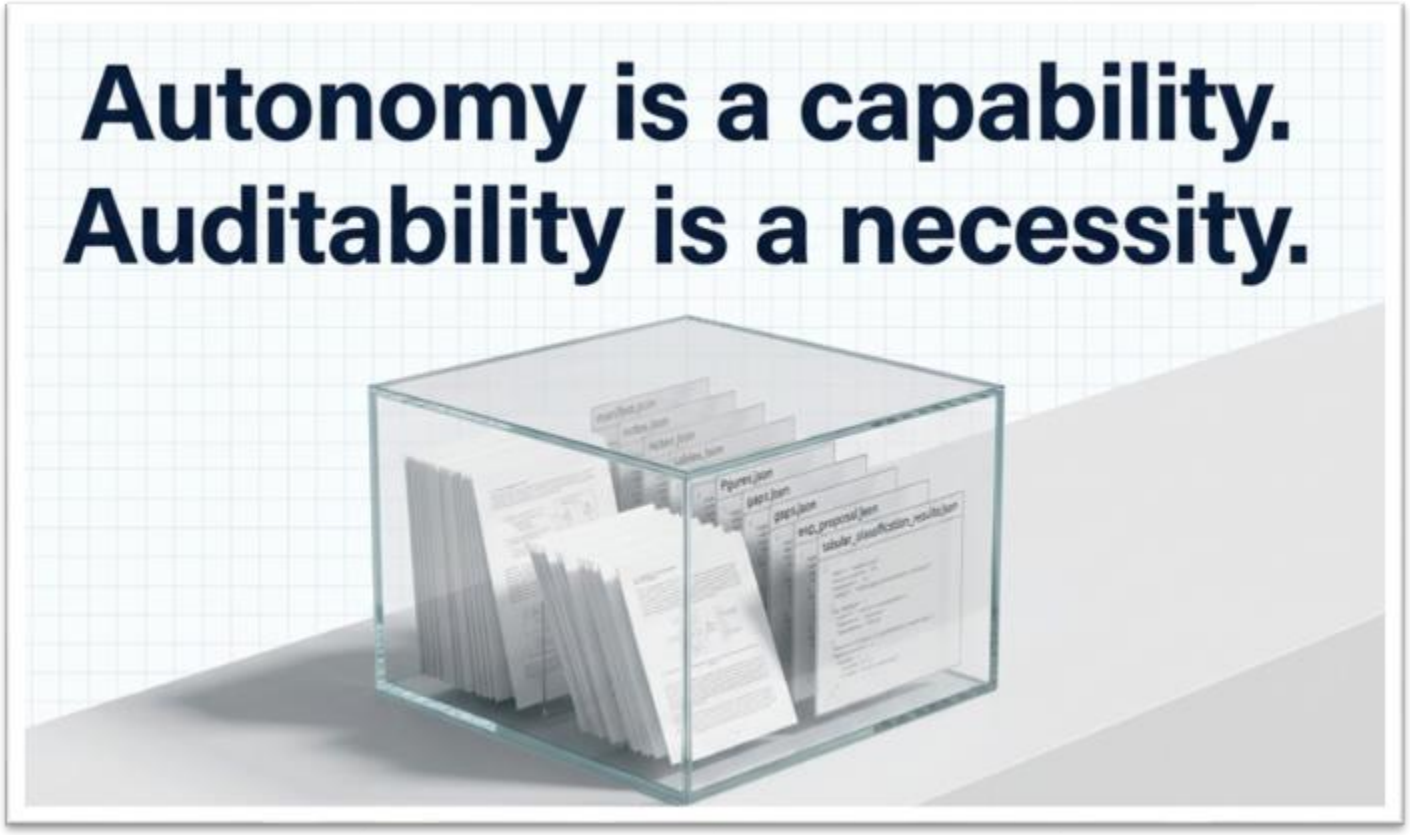
Limitation & Future Work

Limitations:

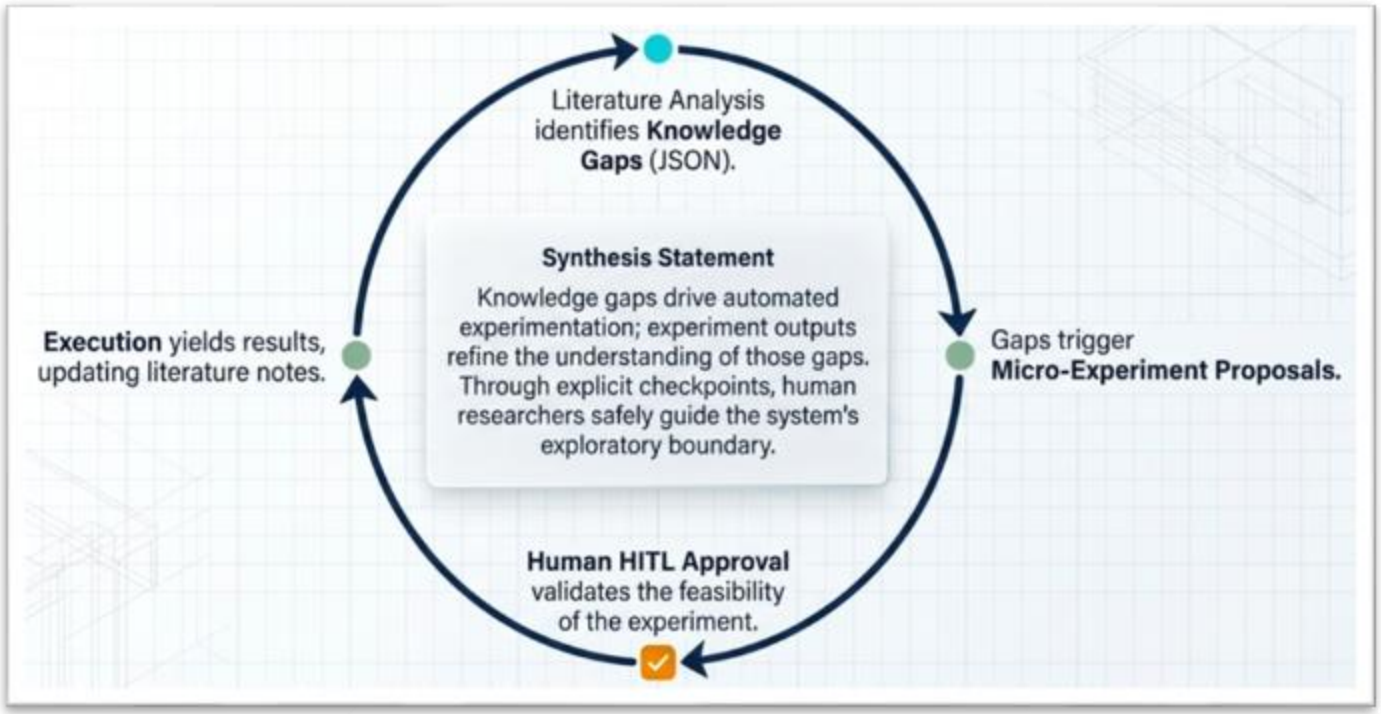
- **Preliminary Evaluation:** Based on a limited number of runs.
- **Semantic Truthfulness:** VerifierAI and GAS-lite are strong proxies but don't fully guarantee semantic truth.
- **Scale:** Lacks large-scale benchmarking across diverse tasks.

Future Work:

- Stronger semantic-grounding verification.
- Support for richer, real-world datasets and specialized tools.
- Comparative benchmarking against other orchestration frameworks



CHAINS proves that autonomous agents must be treated as coordinated components within an artifact-driven workflow, not opaque end-to-end replacements. Transparency, traceability, and human governance are the prerequisites for AI-assisted scientific discovery.



Conclusion

CHAINS frames AI agents not as opaque replacements, but as **coordinated components** in a human-governed workflow. The core value of CHAINS is transparency, traceability, and controlled execution in AI-assisted scientific research.

DEMO

Thank you

KEAN UNIVERSITY

RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY