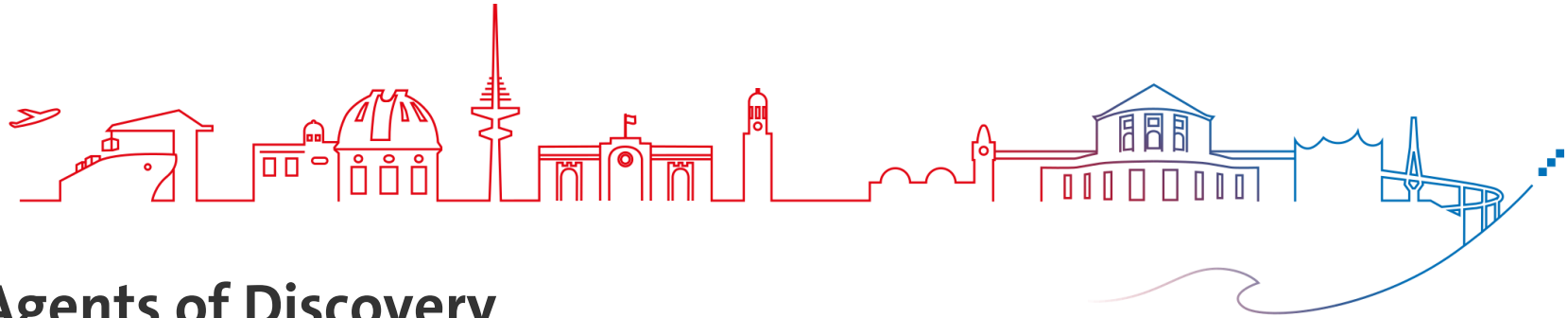
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

PUNCH 4NFDI



Agents of Discovery

8th Round table on Deep learning @ DESY
Hamburg, Nov 14 2025

arxiv 2509.08535
Sascha Diefenbacher, **Anna Hallin**, Gregor Kasieczka, Michael Krämer, Anne Lauscher, Tim Lukas

anna.hallin@uni-hamburg.de

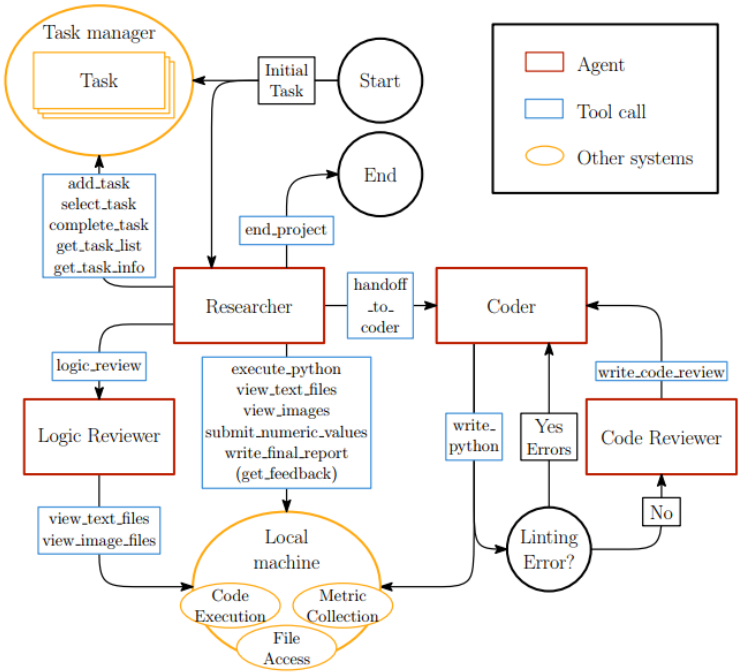
Physics analyses

- Some tasks are more interesting than others
- Some tasks are repetitive
- Can we use AI to take care of those repetitive tasks, so we can spend our times on things that are more interesting?
- AI agents could potentially be inserted at different points in the analysis chain, with access to necessary resources

This work

- Investigate capabilities of commercial LLMs
  - Task: anomaly detection
  - Come up with ideas
  - Produce and run code
  - Analyze the output and report results
- Dataset: LHCO anomaly detection challenge
  - Simulated dataset
  - Find the small amount of signal in a large amount of background

Agent orchestration



- The user supplies the initial **task**, the system then runs without further user input
- **Researcher**: main agent, has acess via tools to task manager
- The Researcher can hand off coding tasks to the **Coder**
- The Coder writes python code, that is checked by a **Code Reviewer**
- The Researcher can use tools to run the code on the local machine, view output files, get feedback (if activated), submit a file of anomaly scores and write the final report.
- The **Logic reviewer** checks the reasoning of the Researcher when requested
- When the score file and final report is submitted, the Researcher can end the project.

Testing different setups

- Tested different LLMs (GPT-4o, o4 mini, GPT-4.1, GPT-5)
- Tested different prompts
  - Different hints
  - Single or multiple task reporting (LHCO challenge questions)
  - Different paraphrasings
- Tested stability over time
- Optional feedback loop tool

Results

- GPT-5 outperforms other models
    - Reaches performance levels comparable to state of the art weak supervision models (given a pure background sample), and even nails the origin of the anomaly
- GPT-4.1 provides good balance between cost and performance
- Hint to use ML seems to be necessary, feedback loop helps
- Telling the LLM that it's the best AI in the world and that humanity and/or the future of the LHC depends on it, gives better results than a more concise prompt
- Variation is quite large – can be "lucky"

- Paper: arxiv 2509.08535; Code: github.com/uhh-pd-ml/AgentsOfDiscovery