

CLARIPHY AI COLLABORATION MEETING · MADISON · SEP 14–16, 2026

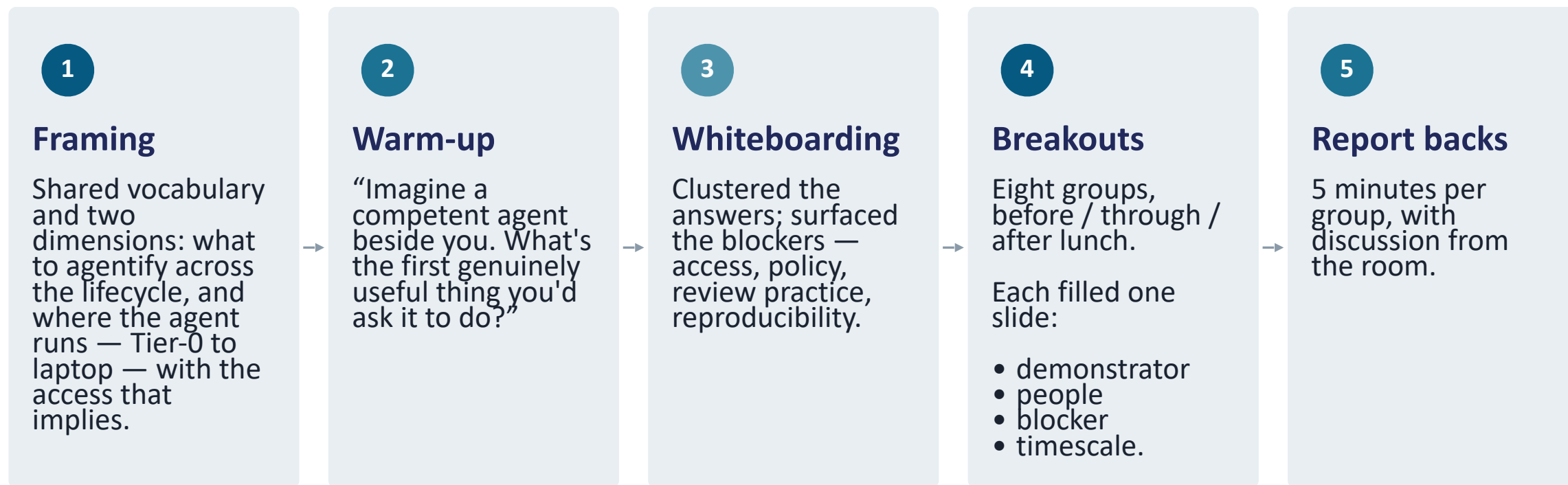
AI Agentic Analysis Systems

Summary of the Sep 15 parallel session <https://indico.global/event/18639/timetable/>

Session organizers

Giordon Stark (UChicago) · Haichen Wang (Berkeley / LBNL)
Andrzej Novak (MIT) · Petra Maksimovic (JHU)

What we did



~120 minutes across two blocks. One group (training / best practices) had no takers and folded into the workforce development session.

Intro slides: https://docs.google.com/presentation/d/1yy25lob9hJB6Sgkp8xVJifSMTWmsjR24dBdP3jZskS4/edit?slide=id.g3fb4c14070c_0_1#slide=id.g3fb4c14070c_0_1

Breakout slides: https://docs.google.com/presentation/d/1OW3PKbxyvfvTBUIUo41IXKxslFIYXqgKhH6Z05Fz6MI/edit?slide=id.g3fb669b0322_50_0#slide=id.g3fb669b0322_50_0

Live notes: <https://docs.google.com/document/d/1SD1PJSub1LT23nauTMZpi4ov2byZMO03GROVWdmYIV4/edit?tab=t.0#heading=h.crq7hdx4w7s>

The first thing people would hand to an agent

Repetitive work nobody enjoys

- Find the datasets and current recommendations, then use with them
- Triage thousands of failed batch jobs; then fix the trivial ones
- Resurrect last publication's code against the current release
- Check that the analysis note is self-consistent

Things we can't hold in our heads

- Enumerate every background process for a final state
- Look up trigger and detector metadata on demand
- Literature review: prior results, limits, reinterpretations
- Overlap detection: where do existing analyses already cover this phase space?
- Critique the plan: cuts, statistics, alternatives from other analyses

Almost none of these are physics problems. They are access, metadata, provenance and retrieval problems.

Four key themes from warm-up discussions

1 Access is technical and political

Reaching calibrations, recommendations, code, and internal docs from outside a collaboration is technical; making things truly public is policy.

2 The analysis workflow is not the whole system

Facilities aren't featureless or limitless. How teams and shared infrastructure negotiate for resources is unsolved and we need a steering committee.

3 Several blockers predate AI

Review processes, note-writing, parallel groups unknowingly duplicating each other. The momentum is a chance to fix things that could have been fixed decades ago.

4 Acceptable use, credit and the PhD

What level of AI use is acceptable in a thesis? What qualifies as a HEP PhD is set by the field, not the university. **This is moved to workforce development.**

Eight groups, eight demonstrators

AREA	DEMONSTRATOR	MAIN BLOCKER	HORIZON
Revisiting how analysis is done	Tech notes that are easy to review — AI edits, student still writes the first draft; test with a real review committee	Social — collaboration & pub-board buy-in	~6 mo if buy-in
Noodling (idea → plot)	Authenticated web interface for fast questions on a small experiment's data; returns the plot and the source code	Access to files, metadata, CLI; auth	~6 mo, 1 FTE
Discoverability	Public skill marketplace for the field + internal versions per experiment; MCP servers to reach existing infrastructure	Sociological; no standards yet	0–3 mo tech, 6+ social
Benchmarking	A standard set of measurements across experiments, scored against published values with a scalar	No clean metric; open data scarce off-collider	~6 mo collider
Review	Split an open-data analysis into analysis & reviewer teams; write the paper about how the review was done	Objective benchmark; expertise; acceptance	Start now
Debugging / reproducibility	Checkpoints that validate individual elements along a workflow providing debugging and reproducibility at once	No common thrust across experiments	~6 mo template
Data standardization	Agentified fit exploring treatments of correlated systematics; recommendations incl. for tokenization	Scale; person-power, coordination, tokens	~6 mo demo
AI use / policy	Supervisor LLM over task-specific small language models, looping until success or a capped return to a human	Sub-agents that are themselves LLMs	now – 6 mo

Five patterns across all eight groups

- 1 **The blocker is almost never the model.** It's access, metadata, standards or people. Sociological and organizational far more often than technical.
- 2 **Everything is six months.** Every group landed on a 3–6 month demonstrator. The field is ready to build. It just needs coordination, not more discussion.
- 3 **Benchmarking, review and reproducibility are one problem.** All three reduce to evaluation: what can an agent be trusted to do, and how do we know.
- 4 **Discoverability is the meta-problem.** Several groups independently proposed building things that partly exist. Nobody knows what everybody else is doing.
- 5 **A collider-only picture is a risk.** Open data, benchmarks and demonstrators skew to ATLAS/CMS; small and mid-scale experiments must be designed in from the start.

Concrete commitments

- **Biweekly “AI Agentic Analysis Systems” (AAAS) meetings**

A standing cross-experiment forum so the six-month demonstrators are tracked, not forgotten. Open to all.

- **IRIS-HEP topical meeting with CLARIPHY participants**

Before the next CLARIPHY meeting in 3–4 months, to connect the agent builders across experiments and theory.

🔗 <https://clariphyworkspace.slack.com/archives/C0BB1JV7FME>

- **Small & mid-scale experiments in scope by design**

Demonstrators and benchmarks include at least one non-collider case from the start.

- **An awesome-hep-agentic-analysis repository (AHAA)**

A curated public index: skills, marketplaces, MCP servers, plugins, agent personas, tutorials.

🔗 <https://github.com/awesome-hep-agentic-analysis>

- **Cross-cutting first**

CLARIPHY AI has no money to allocate; its leverage is convening and identifying common collaborations.

- **Design around agent strengths**

Agents are use prescriptive, checklist-style verification over open-ended creativity; return the source code with every result so a human can check it.

Follow-up continues on the CLARIPHY Slack channel.

Existing roadmaps this connects to

IRIS-HEP agentic-analysis milestones, 2026–27

	Facilities	Analysis
Oct '26	Elwood agent @ UChicago; AGC-notebook scripts	OpenData invariant-mass; AGC statistics
Nov '26	MCP-platform MVP @ Nebraska	AGC with Dask (coffea-casa / Jupyter)
Dec '26	Multi-user, O(10) users / 1 agent each	AMI + Rucio discovery, metadata, download
Jan '27	PanDA + ATLAS Glance @ UChicago	TopCPToolKit skills + C++
Feb '27	Benchmark harness for repeatable runs	Grid jobs, file handling, Rucio rules
Mar '27	Per-user custom skills	Custom library skill + pixi skills
Apr '27	—	LLM physics tasks; ServiceX + AGC
May '27	—	Integration challenge: histserv
Jun '27	—	Integration challenge: Triton / ML inference
Jul '27	—	Build analyses from scratch, checkpoints
Aug '27	—	Full analysis from Snakemake
Sep '27	—	Documentation: Twikis, internal notes

Synergy possible!
benchmark harness → benchmarking group
checkpoints → debugging
skills → discoverability
docs → review-of-notes.

Community whitepaper

“Building an AI-native Research Ecosystem for Experimental Particle Physics”
arXiv:2602.17582 · Feb 2026 · endorsed by DPF CPSC

- Organizes the field's AI vision around four Grand Challenges. This session is Grand Challenge 4: From Data to Discovery — agent-orchestrated analysis aiming to cut design-and-evaluation time 100–1000×.
- Explicitly names small and mid-scale (ASTAE-class) experiments as agile testbeds whose methods transfer to flagship facilities.
- Argues for a national-scale collaboration over institution-by-institution effort — exactly the coordination gap this session found.

Still open

OPEN QUESTIONS

- What should deliberately not be agentic? Where is it cheaper to just write the software than spend tokens?
- Who owns the system-level problem: many agents, many teams, shared facilities negotiating for resources?
- Authorship: if the physicist designs the analysis, steers the agents, verifies the numbers and takes responsibility... is that authorship?
- Should agents parse tracebacks, or should something translate failures into “the dataset has no files” first?

Action items for all

- Do you build skills or agents or MCP tooling, or use AI on analysis notes? Reach out to us!
- Bring a non-collider or small-experiment use case to the first AAAS meeting.
- Join the CLARIPHY Slack channel for follow-up.

 <https://clariphyworkspace.slack.com/archives/C0BB1JV7FME>