Argonne NATIONAL LABORATORY

AGENT-ASSISTED DATA ANALYSIS

HENRY KLEST

AI Town Hall – EICUG/ePIC
July 16, 2026

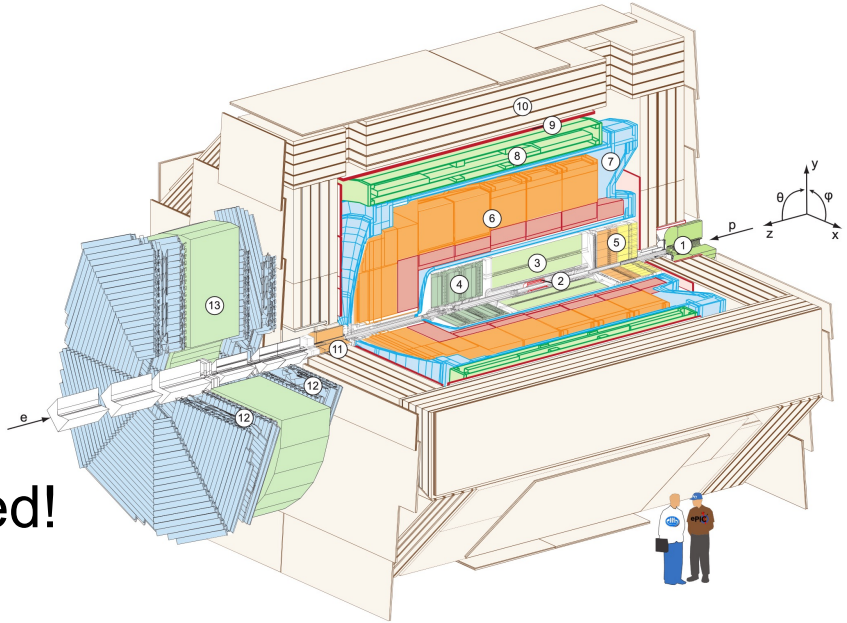
U.S. DEPARTMENT of ENERGY

Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

HERA & THE H1 EXPERIMENT

- **Lots of great science left on the table from HERA!**

- H1 was a leader in the Data Preservation for High-Energy Physics (DPHEP) collaboration
  - —H1 data are immaculately preserved!



- **Can agents do a whole H1 analysis?**

WHY AGENTS?

- **Agents have enormous benefits over humans for analysis tasks**
  - —Can work **autonomously** and **in parallel** on tasks
  - —Can digest huge bases of code in FORTRAN, C++, Java, Python, etc. **without documentation or input from experts**
  - —Can **maximize HPC throughput** & efficiency, less “wasted” jobs
  - —Can easily perform **“high-risk, high-reward” studies on the data**, that nominally a human wouldn’t spend the time to do
    - I.e., “can we do things in a radically different way and achieve a more robust result?” Or “Is there new physics hiding over there?”
  - —Gracefully handles **high-dimensional data** that confuses humans easily
- Primary disadvantage is **loss of context**

AGENT-ASSISTED ANALYSIS

Examples:

- "Could you produce cross section predictions using the Sherpa, Rapgap, and Pythia8 event generators for the binning of our analysis?"
  - Agents dug in CVMFS for 30 minutes and came back with predictions!

- "Could you run those generated events through the H1 GEANT3 detector simulation package and write output data in the same format as our real data?"
  - Ancient codebase, written in early 90s, troublesome even for the only remaining experts
  - After ~ 1 hour of back and forth, agents managed to do it!

H1 ANALYSIS

- In **2 months** working with agents, accomplished essentially **all analysis tasks** for a 4D SIDIS measurement
  - —Optimization of binning
  - —Unfolding
  - —Generating MC predictions
  - —Radiative corrections
  - —Backgrounds
  - —Uncertainties
  - —Writing the analysis note

- Biggest hurdles surround poorly documented analysis components
  - —And finding internal referees… (lesson for ePIC publication committee)

MANAGING AGENTS

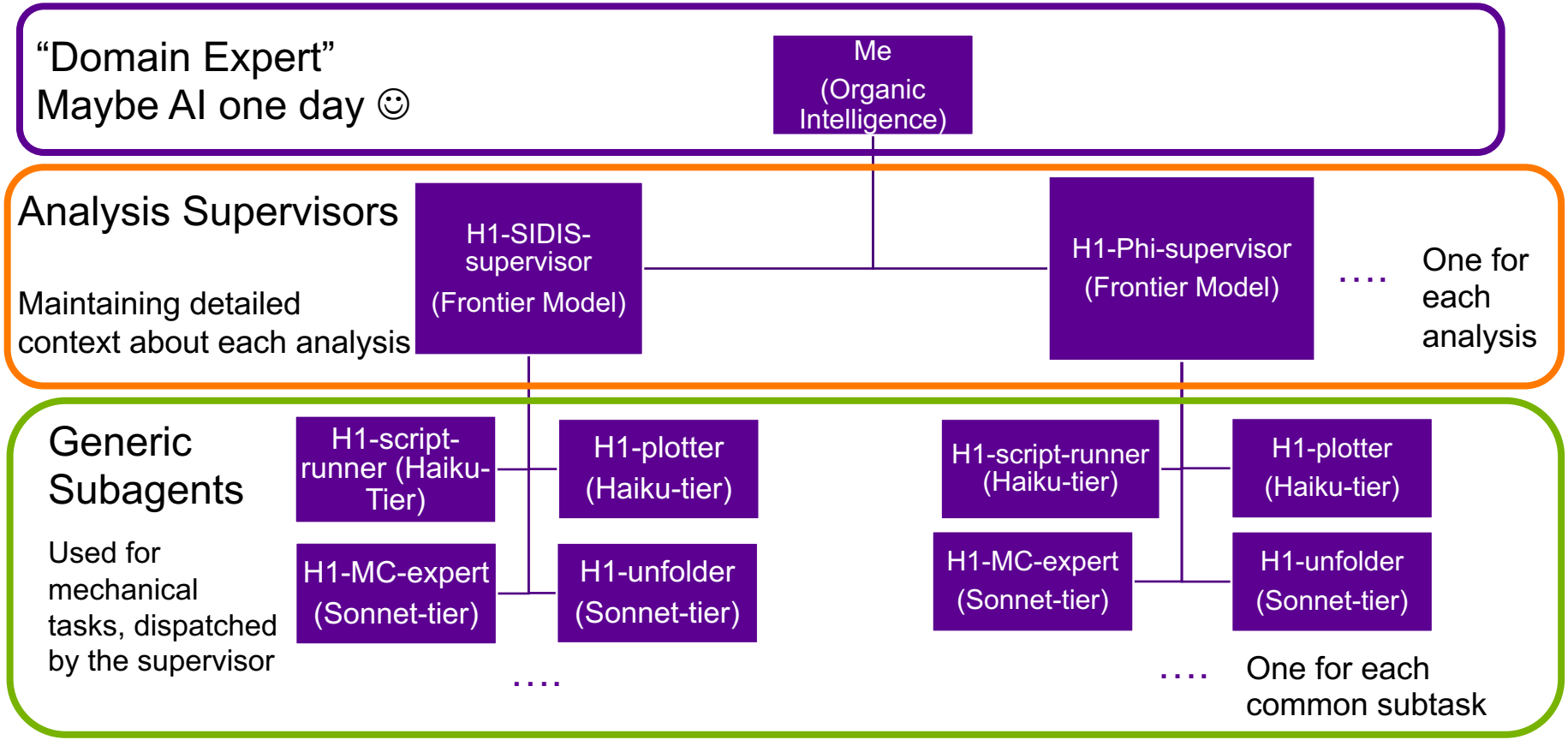
- **Loss of context is the primary challenge**
  - —The “skill” in doing lengthy analysis tasks with agents is **managing their context** to make sure they can handle their task
  - —If you feed **too little information**, agent won’t know what to do
  - —If you feed **too much information**, agent will run out of context (and cost lots of tokens!)

- **Both cases translate to a waste of time and money!**

THE SUPERVISOR

- The primary way to combat loss of context is to have a very high-level **supervisor** which delegates tasks to subagent
  - —The supervisor has: context about the status of the analysis & its dispatchable team of subagents
  - —The supervisor does **not** have: **the permissions to do anything by itself**
    - Otherwise, it will regress to doing tasks itself and eat up all of its context
- I essentially **only interact with the supervisor** when working on my analysis

MY AGENT SETUP “ORG CHART”



MY AGENT SETUP “ORG CHART”

“Domain Expert”
Maybe AI one day

Analysis Supervisors

Maintaining detailed context about each analys

Generic Subagents

Used for mechanical tasks, dispatched by the supervisor

H1-s runner T

H1-MC (Sonr

The new skillset needed for this is **how to be a good manager!**

For example:

Don’t want workers repeating work done by another worker;
Need to allocate tasks based on the skillset of the workers;
Have well-defined goals and milestones for the workers;
Need to be able to judge the quality of the worker’s work

common subtask

ASIDE: WORKING ON REMOTE FARMS

- Working on remote filesystems (e.g. the DESY computers) is a bit clumsy, **agents can't do 2FA!**
- Agents figured out that they can interact with the remote system via ssh when ssh is configured with a persisting control socket

```
# Lists manifest directories in .

$ ssh naf-hone.desy.de 'ls -d /data/dust/group/h1/klesthen/
BaryonAsym/H1ChargedSidis_HighQ2/studies/
```

Sometimes will even send a 1000 line python script over ssh!

N.B. Proper safety guidelines are important for this, if the ssh tunnel drops, agents will spam 100 ssh commands and get your IP banned…

```
# Inspects release bin mapping details

$ ssh naf-hone.desy.de "python3 - <<'PY'
import json, collections
r='/data/dust/group/h1/klesthen/BaryonAsym/H1ChargedSidis_HighQ2/output_v6/central/AziAsym/
ibu_trento_phi_S3xmerged_zlow010_pT4_phi7_alpha0p1_v6syst'
rows=json.load(open(r+'/systematics_releases/zero/ibu/cross_sections.json'))['rows']
print('nrows',len(rows))
charge='pos'
keys=[]
for row in rows:
 if row['charge']==charge and row['phi_idx']==0:
  keys.append((row['bin_index'], row['cell_label'], row['q2_min'],row['q2_max'],row['x_min'],row['x_max'],row['z_idx'],row['z_min'],
row['z_max'],row['pT_idx'],row['pT_min'],row['pT_max']))
print('first phi0 rows', keys[:20])
print('cells', sorted(set((r['cell_label'],r['q2_min'],r['q2_max'],r['x_min'],r['x_max']) for r in rows if r['charge']==charge)))
print('z', sorted(set((r['z_idx'],r['z_min'],r['z_max']) for r in rows)))
print('pt', sorted(set((r['pT_idx'],r['pT_min'],r['pT_max']) for r in rows)))
coeff=json.load(open(r+'/systematics_fits/zero/trento_phi_coefficients.json'))['coefficients']
for e in coeff[:10]:
 print('coeff key', e['charge'],e['cell_label'],e.get('z_idx'),e.get('pT_idx'),e.get('z_min'),e.get('z_max'),e.get('pT_min'),e.get(
'pT_max'),e['B_over_A'])
PY"
```

THE ELEPHANT IN THE ROOM

- The obvious question that will be asked is **“how do you know it’s right”**
  - We’re doing science, after all… **It needs to be right!**

- Treat the agents as a student with great coding skills but less physics intuition
  - Professors are rarely reading the code of their students!
  - **How do they know the results are correct?**
  - Need to ask them to reproduce existing results, make QA plots, etc., same for agents!

- One bonus of agents is that you can send waves of **critic agents** to critique the code implementation
  - Useful to give them different “personalities”, i.e. high temperature, low temperature
  - If 3 agents all say the implementation is correct, you have 1/sqrt(3) confidence that it’s right ☺

CONCLUSION

- Can agents do a whole H1 analysis?
  - **Kind of**, currently needs constant supervision
  - With the rate of improvement, could imagine in a couple years that a frontier model could do an entire analysis by itself ***(if proper documentation is available)***

- Agent-assisted analysis is the future
  - Currently like riding a bucking bronco vs. walking, but we'll have cars someday ☺
  - Allows us to just do the fun part of science! (i.e. **the science**), also **improves accessibility** to our field!
  - **Saves a huge amount of time and life force** that is generally spent on mechanical tasks and not actual data analysis
    - Running jobs on HPC clusters, writing plotting scripts, debugging segfaults
  - Enables performing checks that otherwise may go undone and actually helps **make the analysis more rigorous**

Thanks! Questions?

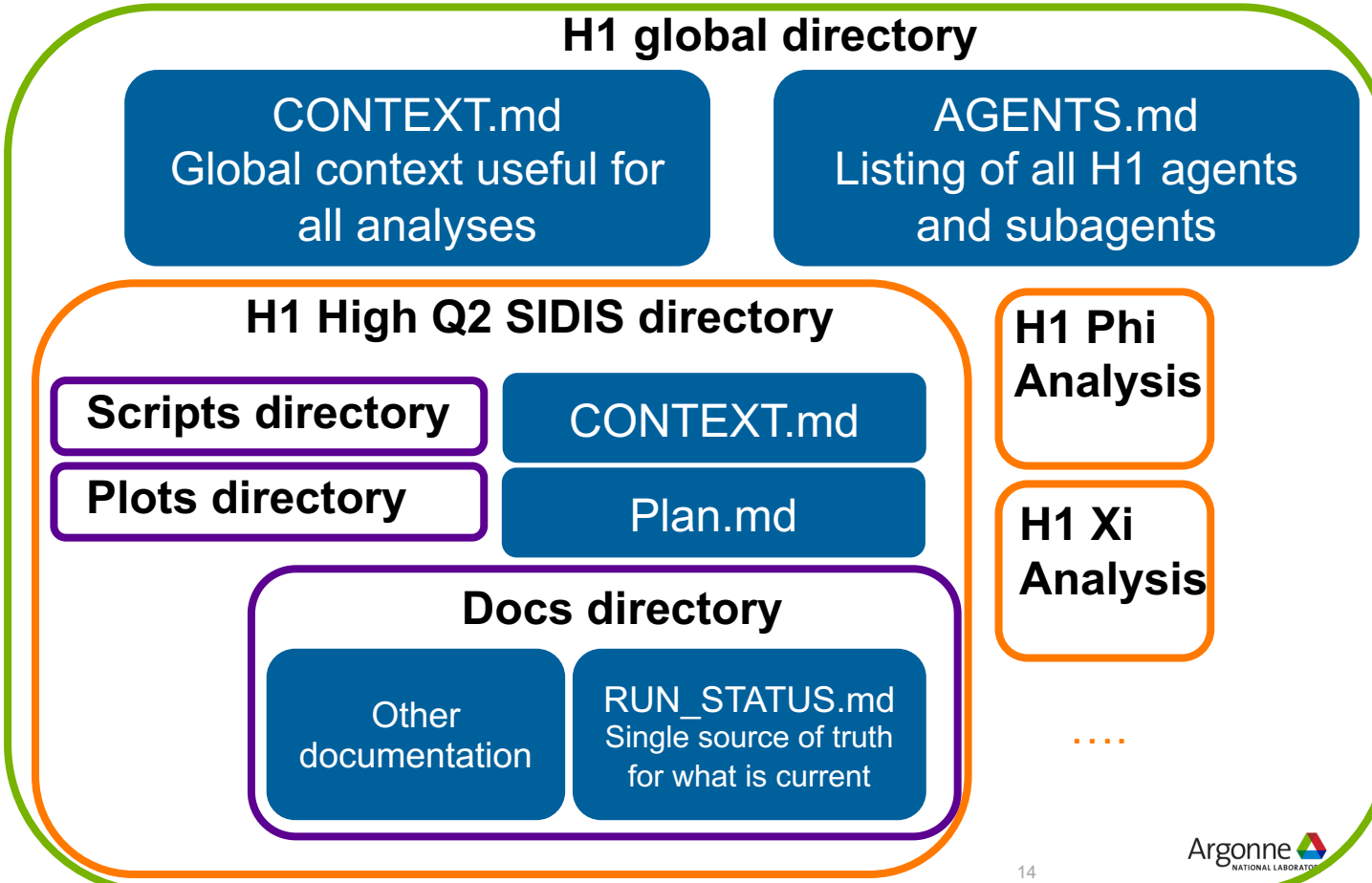
BACKUP

CONTEXT MANAGEMENT

Maintaining good documentation is absolutely critical, since we need to work across many sessions

Every directory should have a short manifest that tells agents what's there

Agents need to know what scripts to use, what files, etc.



TOKEN EFFICIENCY

- Token costs for heavy agent usage mitigated somewhat because ~90% of tokens are cached input tokens (cache hits)

- I switched to GPT5 models because Opencode doesn't show me any cached tokens for Anthropic models
  - Don't want to unnecessarily cost the lab 50x more!

Prices per 1M tokens.

	Short context			Long context		
Model	Input	Cached input	Output	Input	Cached input	Output
gpt-5.5	$5.00	$0.50	$30.00	$10.00	$1.00	$45.00
gpt-5.5-pro	$30.00	-	$180.00	$60.00	-	$270.00
gpt-5.4	$2.50	$0.25	$15.00	$5.00	$0.50	$22.50
gpt-5.4-mini	$0.75	$0.075	$4.50	-	-	-
gpt-5.4-nano	$0.20	$0.02	$1.25	-	-	-

Model	Base Input Tokens	5m Cache Writes	1h Cache Writes	Cache Hits & Refreshes	Output Tokens
Claude Fable 5	$10 / MTok	$12.50 / MTok	$20 / MTok	$1 / MTok	$50 / MTok
Claude Mythos 5 (limited availability)	$10 / MTok	$12.50 / MTok	$20 / MTok	$1 / MTok	$50 / MTok
Claude Opus 4.8	$5 / MTok	$6.25 / MTok	$10 / MTok	$0.50 / MTok	$25 / MTok
Claude Opus 4.7	$5 / MTok	$6.25 / MTok	$10 / MTok	$0.50 / MTok	$25 / MTok
Claude Opus 4.6	$5 / MTok	$6.25 / MTok	$10 / MTok	$0.50 / MTok	$25 / MTok
Claude Opus 4.5	$5 / MTok	$6.25 / MTok	$10 / MTok	$0.50 / MTok	$25 / MTok
Claude Opus 4.1 (deprecated)	$15 / MTok	$18.75 / MTok	$30 / MTok	$1.50 / MTok	$75 / MTok
Claude Opus 4 (retired, except on Vertex AI)	$15 / MTok	$18.75 / MTok	$30 / MTok	$1.50 / MTok	$75 / MTok
Claude Sonnet 4.6	$3 / MTok	$3.75 / MTok	$6 / MTok	$0.30 / MTok	$15 / MTok
Claude Sonnet 4.5	$3 / MTok	$3.75 / MTok	$6 / MTok	$0.30 / MTok	$15 / MTok

AGENT-ASSISTED ANALYSIS IN H1

- One analysis typically takes **3-4 years** from beginning to publication
  - —Usually, would pay one full Ph.D student FTE, part of a lab staff/professor during that period
- Since beginning of May, have been performing **5 unique experimental analyses in parallel by myself** using agents
  - —3 of 5 are now under collaboration review for public release
  - —Agents self-document, update an analysis note/paper draft as they go that can be shared with collaboration referees
  - —Usually takes ~2 years of analysis effort to get to this point! **With agents, ~ 1 month**

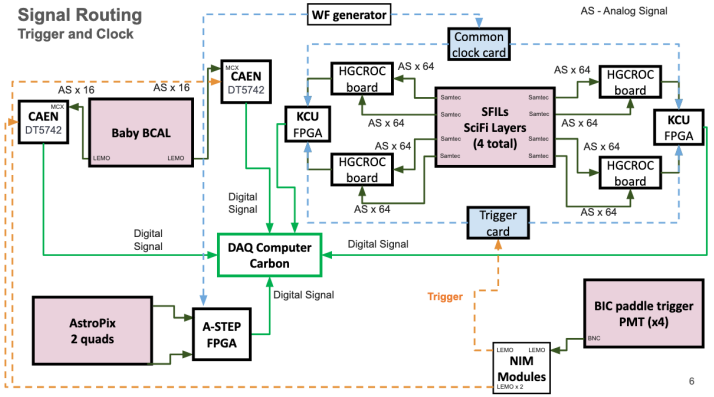
AGENT-ASSISTED ANALYSIS IN H1

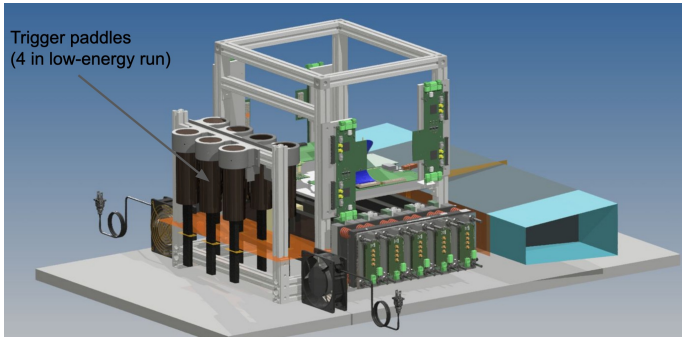
- Argonne's current API approach gives us an enormous **competitive advantage over other labs & universities**

- **Other labs will pay for a student or postdoc FTE to take three years to complete an analysis that $20k in tokens can complete in a couple months**
  - —This is clearly the future of physics data analysis, and we have a head start

TEST BEAM AT JEFFERSON LAB

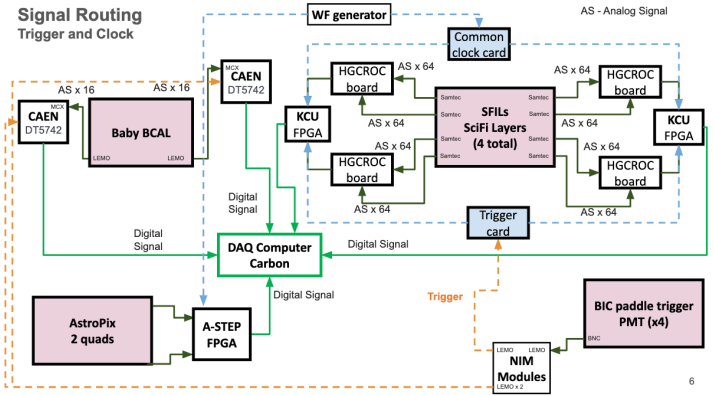
- In March & April, had a prototype of the Barrel Imaging Calorimeter that our group is building for the EIC exposed to beam parasitically at Jefferson Lab
  - Very complicated readout setup employing 3 heterogenous readout systems, 2 different ASICs, 5 FPGAs
  - Senior data acquisition expert from BNL who wrote our DAQ software was oversubscribed and mostly not available to help during running
  - Had no direct access to setup, agents were instrumental in keeping all systems functioning remotely

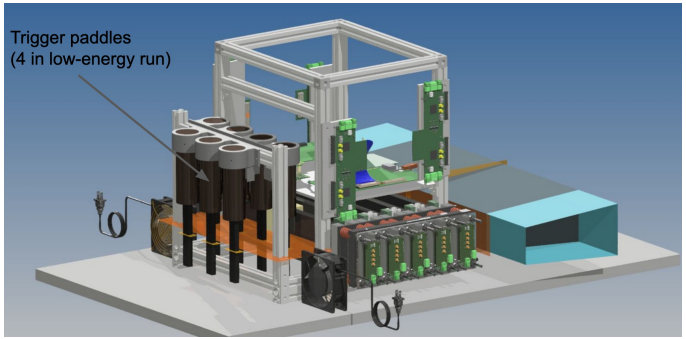




TEST BEAM AT JEFFERSON LAB

- Claude was able to digest the entire data acquisition software chain & replace the expert.
  - —Suggested and implemented improvements on the DAQ software that allowed us to run at a 3x higher data rate
- Agents also instrumental in analysis of the very large dataset, finding O(10-100x) speed improvements in decoding, data processing, etc.
  - —Saving huge amount of CPU hours





PHENOMENOLOGY

- Currently 3 agent-assisted papers in the works on possible physics opportunities at the EIC and JLab

- Less mechanical than data analysis, requires strong reasoning capabilities + cross check
  - —Generally less token-heavy overall since the task set is smaller and less heterogeneous