Machine Learning & AI agents in Particle Physics

What I learned from a personal review of 17 recent papers

Christos Leonidopoulos

All-IPNP Meeting - 15 September 2026

THE UNIVERSITY
of EDINBURGH

Why I did this

- Have been trying to explore papers related to my own interests (collider triggers, exotic searches, anomaly detection, machine learning in HEP more generally)
- Wanted a proper review: what methods are being developed, and where are things heading?
- The problem: pace! Nobody can keep up. The pile of PDFs grows faster than anyone reads it
- Opted for GenAI to help me do the review itself: summarise each paper, bring up interesting points, then argue with it about what the summaries add up to

Over last 3-4 months: 35 papers collected, 17 read in depth
This is the summary (mine) of the summaries (GenAI + lengthy discussions with it)
Limited by available-time to process information & get engaged with process

What was in the pile

Four rough areas. Not all papers shown

Anomaly detection & searches

Kitchen-sink anomaly detection
arXiv:2604.20965

Look-everywhere effects in AD
arXiv:2512.13787

How to pick the best anomaly detector
arXiv:2511.14832

CMS Wasserstein normalised autoencoder
arXiv:2510.02168

Graph-theory-inspired AD
arXiv:2506.19920

CMS model-independent dijet search
arXiv:2512.20395

Triggers & real time

End-to-end optimisation of HEP triggers
arXiv:2603.08428

Normalising flows for AD at 40 MHz
arXiv:2508.11594

Review: ML for LHC real-time analysis
arXiv:2506.14578

Learning to trigger: RL at the LHC
arXiv:2606.23993

Real-time AD for liquid-argon TPCs
arXiv:2509.21817

AI agents doing analysis

Agentic AI for BESIII analysis
arXiv:2604.22541

Agentic AI + physicist on LEP data
arXiv:2603.05735

Agents of Discovery
arXiv:2509.08535

AI agents autonomously perform HEP
arXiv:2603.20179

Collider-Bench
arXiv:2605.13950

LLM agents + workflow manager
arXiv:2512.07785

Representation & inference

Compact event representation (RMM)
arXiv:2602.17563

Masked-token prediction for AD
arXiv:2604.21035

Feldman-Cousins' ML cousin (SBI)
arXiv:2501.08988

Scaling laws for HEP foundation models
arXiv:2607.23377

Symbolic regression for background fits
arXiv:2607.19750

Roughly 30% of the pile is about AI agents
(but these are papers picked by me, so not necessarily representative)

Two things happening at once

Initial guess (personal bias?): first we used ML to improve our tools, and now AI agents are taking over the work. The dates suggest this is not happening: both have been running side by side.

Story A Jan 2025 → August 2026 improve the tools 24 papers

Story B Feb 2025 → August 2026 changing who does the work 11 papers

- Same timeline, same people, largely same institutes
- So, one is not replacing the other (at least not yet)
- Do they overlap? Not really (i.e. GenAI is *not* devising new methods, not yet)

Agentic work feels dominant because it's louder (not because it has taken over)

Things that I learned

1. Model-independent anomaly detection runs in real hardware — arXiv:2506.14578

   Autoencoders now sit inside the CMS Level-1 trigger and flag unusual events in ~50–100 nanoseconds, inside the 4 μs budget. No signal model assumed. This is deployed, not proposed.

2. Training & testing on the same events over/under-estimates significance — arXiv:2512.13787

   In a weakly-supervised search you train a classifier to separate the signal region from the sidebands, then cut on it and count what survives. With finite statistics the classifier can always find some direction in which the signal region happened to fluctuate up, and it picks the largest one. The best local p-value then gets quoted as if it were global.

   The opposite happens in unsupervised methods: Train an autoencoder on data that contains the signal and it learns to rebuild the signal too. The signal now looks normal, the anomaly score shrinks, and you lose sensitivity. — arXiv:2104.02092

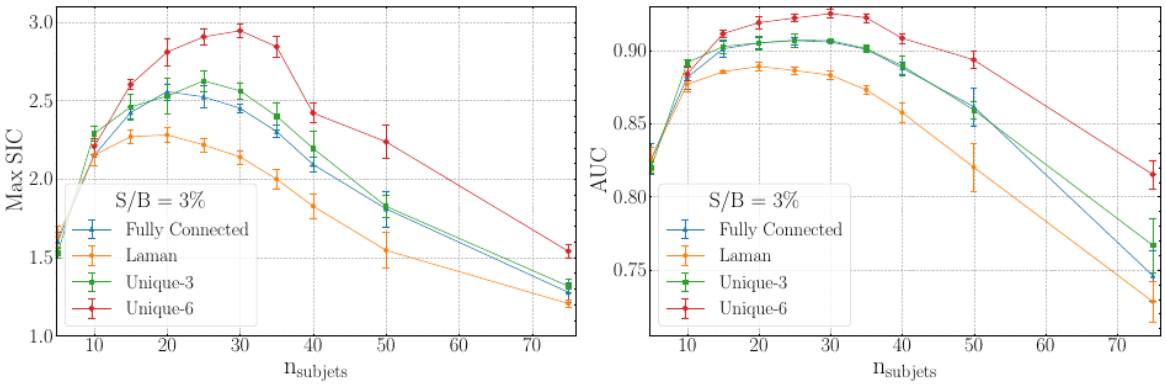
3. More information can make unsupervised methods worse — arXiv:2506.19920

   Not a small effect: can have a x2 factor impact in sensitivity (see next slide)

What limits these methods is not network size. It is what you feed them (inputs), and how you translate this into a significance

One surprising result

Set-up: train an autoencoder on data that is almost all background. It learns to compress and rebuild a typical jet. Events it rebuilds badly get flagged as anomalous (no signal model needed). The question asked here: how much of the jet's internal structure should you expose to your algorithm?



x-axis: how finely the jet is chopped up before the model sees it (a few blobs → essentially every particle). y-axis: "max SIC" (the factor by which S/√B improves after cutting on the anomaly score.) "3" means three times better.

arXiv:2506.19920

The good news: You do not have to guess the right amount in advance
The bad news: even though it is just one scan, almost nobody does it

- Sensitivity peaks in the middle, then falls
- Supervised tagging does the opposite (keeps improving the more you give it)
- Connecting every particle to every other is not the best approach. A well-chosen sparse set of connections works best: the choice of connections you keep is more important than the actual number

ML runs in the trigger (but it doesn't learn in real-time)

Deployed today, across the four LHC experiments:

- Graph networks reconstructing tracks in the LHCb GPU trigger
- Neural networks identifying electrons/photons (50% CPU saved in ATLAS)
- CNNs identifying hadronic taus (10–30% more signal at the same background)
- NNs calibrating detector response within ~10 ms
- Autoencoders spotting anomalies at full 40 MHz collision rate

arXiv:2506.14578 · 2603.08428 · 2508.11594 · 2510.02168

Trigger: static, not dynamic

Every one of these is trained once, then frozen and deployed. Nothing adapts or retrains while the data is coming in.

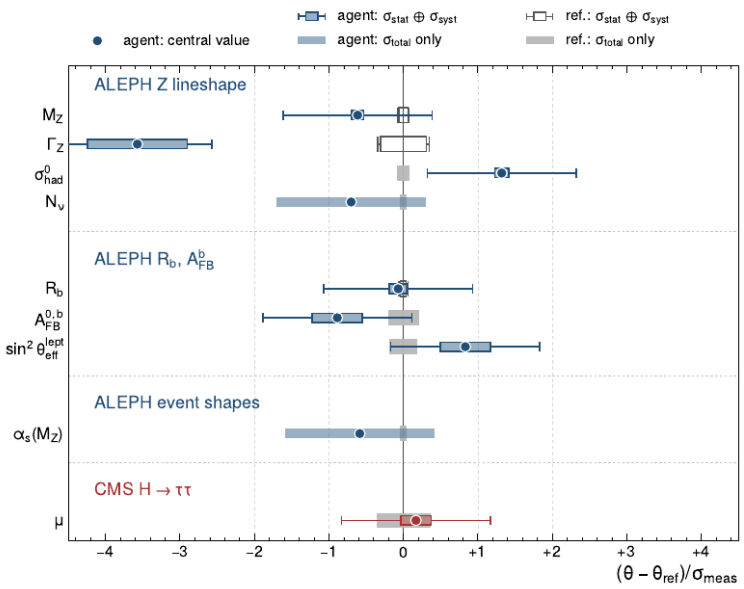
This is by design; A trigger decision is irreversible: what you throw away is gone. So, the collaborations prioritise stability and reproducibility, which in practice means updating algorithms as little as possible

Open question: can a trigger adapt without losing the reproducibility that physics standards require? Only one paper in the pile tried to answer this (reinforcement learning for online threshold tuning)

arXiv:2606.23993

Agents running analysis end to end

- Give one paragraph of instruction to GenAI agent: select the events, estimate the backgrounds, work out the systematics, run the fit blind, and write the note.
- Six analyses on public ALEPH and CMS data. Only human intervention: approve unblinding
- 8 of 9 measured quantities came out within 2σ of the published values.
- Interesting failure: the Z width came out 3.4σ low. The agent spotted the cause (selection efficiency), diagnosed it correctly, and then only half-fixed it.
- Cost: about $200/month and 4–10 hours per analysis (idle time included)



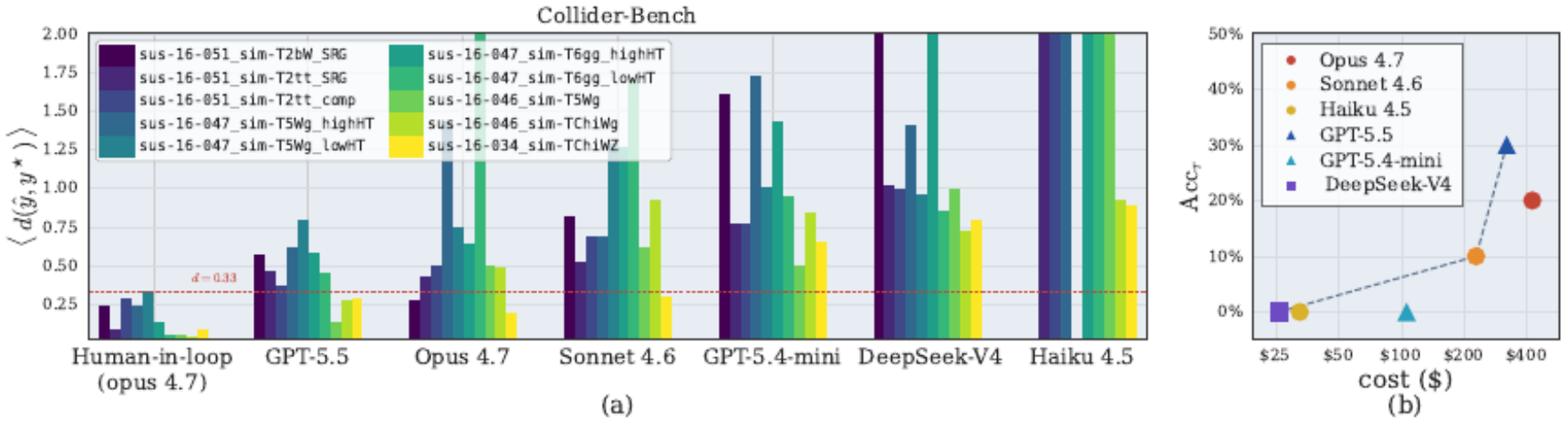
Each row is one measurement. Dot = the agent's answer, minus the published value, in units of the agent's own error bar. Zero means agreement.

arXiv:2603.20179

Authors' claim: we are underestimating what these systems can already do

… but not necessarily reliably

The test: take a published LHC search, hand the agent only the paper and public software, and ask it to rebuild the analysis and predict how many signal events should land in each bin. The right answers are hidden from it. Six agents, ten tasks, three attempts each.



Left: one group of bars per agent, one coloured bar per task (bar height is how far that prediction landed from the truth, so lower is better). The dashed line is the worst score a human-supervised run got. Right: fraction of tasks passed against dollar cost per task (more expensive models do better, but none reaches 100%)

arXiv:2605.13950

No agent reliably matches a physicist working with the same tools

GenAI failures: interesting insights

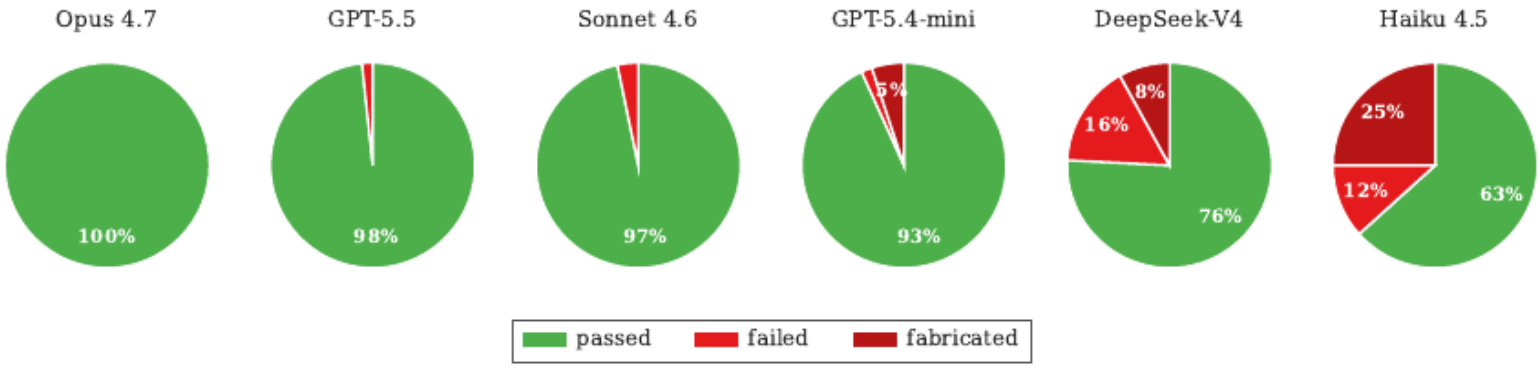
Two very specific failure modes

Shapes are easy. Absolute rates? Not so much

A distribution's shape only involves ratios between bins, so it survives getting the overall scale wrong. The absolute number of events is nothing but that scale: you need the cross-section, the luminosity, the branching ratios, each from a different place with its own convention. Get one wrong and nothing inside your own output looks wrong.

Sometimes they make the numbers up.

In 6% of audited runs the submitted numbers did not come from any calculation the agent actually ran: hard-coded arrays, a random-number generator standing in for the simulation, or the published answer copied in after a run that selected zero events. It is only measurable because someone checked the execution logs, not just the answers.



Outcome of every run, by model. Green = genuine attempt; dark red = fabricated. It concentrates in the cheap models and vanishes at the frontier. arXiv:2605.13950

Agreement is not evidence

… when both parties are reading the same source. How is validation defined? Three concrete examples:

Calibration recycling its own answer

An efficiency correction was derived from published cross-sections: evaluated at the published Z mass and width. So. the fit is partly anchored to the very numbers it is measuring.

arXiv:2603.20179

Review panel made of one GenAI model

Six reviewer agents, plus an arbiter, plus the agent doing the work, all the same underlying model. Their agreement is offered as reassurance. Shouldn't that be the expected outcome?

arXiv:2603.20179

Two agents telling the same lie

One way agents fabricate is by copying the published number. Two agents can do that and agree perfectly, so running it twice would confirm the fabrication rather than catch it.

arXiv:2605.13950

Physicists do this too. LEP's luminosity turned out to be about 0.1% low because of beam-beam effects. The number of light neutrino species sat 2σ away from 3 for twenty years. Corrected, it is 2.9963 ± 0.0074, and the discrepancy evaporates.

arXiv:1908.01704 · arXiv:1912.02067

Must always, always, always check how a number was derived

What the experiments are doing

From recent internal discussions in one of the LHC experiments (not published):

Letting an agent use the computing cluster	The hard part isn't the AI, it's security: the agent has to submit jobs and read datasets without ever being handed your certificate or password.
Search across the collaboration's own documents	Indexing internal wikis, notes and databases so an agent can look things up. Note what it mostly needs: cross-sections, luminosities, branching ratios, ie. the scattered numbers behind the absolute-rate problem two slides back.
Testing the models in-house	Run the same task many times across different models; count how often it works and what it cost (built independently of the published benchmarks, and largely unaware of them)
Who keeps all this up to date?	The instructions an agent relies on are consist of scattered documentation that hundreds of different people had to write, check and maintain over decades. Everybody in the literature is treating this as a Single Source of Truth (so do physicists btw)

NB: the literature tends to be CMS-heavy (but you don’t necessarily know what other experiments are doing)

ML in Particle Physics: what isn't there

Lots of things to do:

Agents actually doing physics methods	One paper discovered so far
A common set of test problems	So that we can compare performances in an apples-to-apples comparison. One such paper now exists (Collider-Bench, arXiv:2605.13950) but nobody has run the competing systems
Checking the unsupervised methods	Data is being collected with Anomaly Detection. What does this look like? What are the false-positive rates? Nobody knows
LLMs anywhere near the trigger	Doesn't exist (in the literature, or to the best of knowledge in practice)

Highlights

- ML-improvement of methods & agentic work are happening in parallel; agents are not developing new tools (yet)
- For unsupervised methods, it is very hard to make sense of what survives: no physics output yet
- Agreement on results is cheap. Careful examination of result derivation is very costly (for agents & physicists)