Agents of Discovery - Referee comments

Summary of paper

This manuscript demonstrates the first use of agentic AI for a realistic LHC analysis. The authors use a “team of agents” based on large language models (LLMs) to perform anomaly detection with the LHC Olympics dataset. Several models by OpenAI are investigated for this task (GPT-4o, o4-mini, GPT-4.1, and GPT-5), and their stability, reliability and physics performance are compared. The paper demonstrates that the “team of agents” are able to solve this problem, and employ the same techniques (weak supervision, bump-hunt etc) as used by humans. GPT-5, as the most advanced LLM release, is shown to be the most capable model. However, the improved performance comes at the price of increased runtime and execution costs. The authors go on to study how different prompting and the addition of a feedback loop can affect the technical and physics performance, showing a reasonably large variation in some metrics. These studies were carried out with the GPT-4.1 model, as it offered the best trade-off in terms of performance and cost.

Overall assessment

This paper is of high interest as it is the first thorough investigation into the use of agentic AI for LHC analyses. The methodology seems technically sound in all aspects and is presented in a clear way throughout. The use of language is excellent, and I found little-to-no spelling/grammar mistakes. Therefore, I highly recommend this paper for publication in SciPost, subject to the minor revisions/clarifications listed below. Thank you for the interesting work.

Minor revisions/questions

General

- It should be noted that anomaly detection in a limited mass range is a very simple analysis, compared to the breadth of analyses that we perform at the LHC. More complex analyses require calibration, control regions, statistical modeling, extraction of confidence intervals etc (to name just a few). While you have to start somewhere and anomaly detection is a good baseline, I think it’s important to mention in the paper that this is a simple problem, and we are some way off using agentic AI across the board. This is not to say that you are overstating the significance of the results, but it would be good to remind the general reader of this. Perhaps in Section 2.1 and in the conclusions.
- The results demonstrate a large spread in performance (both technical and physics) with different models, and subsequently with subtle changes to the configuration/prompting. How should one explore this extremely large space to obtain the best “team of agents”? This becomes especially difficult when the performance can vary substantially upon repetitions with exactly the same configuration. Do you have any guidance on this? If so, I would consider adding a paragraph or two to help those who read the paper and want to implement for their use-case.
- Did you investigate the Type-2 error rate i.e. the possibility of the “team of agents” claiming discovery with no signal in the data set? If the study would not take too much

time/resources, then I would recommend adding it to the paper. It is often a worry (also in day-to-day use-cases) that LLMs can be over-confident in their responses.

2. Dataset & Task

- Add reference of n-subjettiness ratio.
- Consider showing the 1D distributions of the input features for data, background and signal at the end of Section 2.2. This will help inform the reader on the information that the agents have access to.

4. Results

- Figure 3 caption: change order of metrics in the caption to match the ordering of the plots in the Figure (Input tokens <-> Output tokens).
- Section 4.2.1:
    - There is a fixed amount of signal in the data (0.6%) yet the p-value can change substantially. This is most notable for GPT-4.1, where many of the runs suggest that there is no signal in the data (p-value = 1). What should one conclude from this? My immediate reasoning would be that you can't trust what the agents are telling us about the probability of the background-only hypothesis.
    - Add a comment about the outlier that observes 100% signal percentage for GPT-4.1. What have the agents really concluded in this run?
    - There is no discussion of the "max SIC" score in the text, yet this is an important metric to show how well the agents have classified the anomalous signal. Can you add a paragraph with this discussion.
    - Can you explain what is happening for the runs which obtain a Max SIC of ~15? These outliers are much higher than the bulk of points, and are pulling up the average significantly. Without these, there seems to be little difference in the performance between GPT-4.1 and GPT-5. Are they real cases where the methods employed are much more sensitive to the signal? If so, which methods were used.
- Section 4.2.2:
    - Can you provide a quantitative comparison of the physics performance between humans and the "team of agents"? While it is impressive to see similar methods used, we don't know if they have been used to their full capacity. Without the quantitative comparison it questions comments in the conclusions section like: "To summarise, the most advanced LLM that we tested, GPT-5, achieved AD results that are comparable to those produced by humans."
- Section 4.2.3:
    - Sentence containing "with at most 3 failed… apart from the Ideas+ML prompt which failed 6 runs", sounds odd given there are only four options tested. Please change the wording.
    - I recommend adding "GPT-4.1" into the top-right hand corner of each plot shown in this section, to remind the reader that all points in the plots are for GPT-4.1.

    - As above, please comment on the outliers of max SIC in Figure 6. With the “ML” prompt, have the agents been able to use a method that achieves a much higher Max SIC?
- Section 4.3:
    - The discussion at the end of this section feels like cherry-picking. Is it not the case that with enough runs you will eventually obtain a team of agents which predict mass and signal percentage values which are close to the truth?
    - Figure 11: Add a comment in the caption that the standard deviation band is missing for the resonance mass for “ML+FBL^+” because only one run provides a mass value.

5. Conclusions

- Regarding the two points mentioned above:
    - Add a sentence or two re-stating this is a simplistic analysis (anomaly detection) and used as a starting point to explore agentic AI. Most analyses at the LHC require more complex methodology.
    - Unless you are able to add a quantitative comparison to human results, I would change the phrasing “To summarise, the most advanced LLM that we tested, GPT-5, achieved AD results that are comparable to those produced by humans.”, to reflect that the methods chosen are comparable (rather than results).