AI Agents at MBI - first Experiences

A manuscript revision with Codex

Addressing referee comments and drafting the response

Daohan Wang

Marietta Blau Institute for Particle Physics
Austrian Academy of Sciences

October 23, 2026

The case: a GraphNet manuscript revision

Submission to SciPost Physics

- The referee requested clarifications and additional evidence.
- Some comments required LaTeX edits.
- Others required a new graph neural network (GNN) baseline and further analysis.

3
questions

2
requested changes

The practical question

How can an agent help complete the revision and write a response that accurately describes the work?

The workspace is part of the prompt

I collected the relevant files in `2nd_reply/`.

Material	Why the agent needed it
Referee report	The exact questions and requested changes
Manuscript source and PDF	The current argument and where to edit it
Analysis and batch scripts	Existing model and evaluation conventions
`references.bib`	The bibliography already used by the paper
`reply.txt`	The format for the response

Initial instruction

"Let us address them one by one."

A sequence of bounded tasks

Stage	My decision	Agent contribution
Manuscript	Meaning and insertion point	Edit the LaTeX source
New baseline	Architecture and training scope	Adapt the existing code
Evaluation	Data and procedures to preserve	Run the requested scripts
Comparison	Models and figure layout	Adapt plotting code
Response	Scientific interpretation	Draft each reply

Each stage produced an artifact I could review.

I checked the reported result before defining the next task.

Precise instructions for manuscript edits

Example 1: missing references

Add the CMS Combine and ATLAS HistFactory citations.

Context: both entries already existed in `references.bib`.
Location: immediately after a specified sentence on page 21.

Example 2: the inference method

Clarify the unbinned likelihood fit in the leading FAIR-HUC solution and our use of binned classifier outputs for final inference.

My interpretation: avoiding binning may improve statistical precision.

I supplied the scientific meaning and the exact place to edit.

An additional single-GNN baseline

Condensed instruction to the agent

Create a single GNN using all **35 features**, trained only on **nominal samples**. Adapt the existing code in `newGNN/`.

Remove	Training perturbations and the uncertainty-aware branch, including its attention, gating, and branch fusion.
Preserve	The deterministic branch's message-passing layers, hidden dimensions, and activation functions.
Adapt	Reorganize all 35 features into one graph with node, edge, and global inputs. Update graph construction and input encoders. Adapt the classifier, training loop, and model saving to a single branch.

The new baseline required adapting the full model pipeline.

Training and evaluation on CLIP

The agent adapted our existing scripts to the **new single-GNN model**.

Task	**Agent contribution**
Training	Use the new model with the existing training batch script.
Perturbed predictions	**Successfully submitted 11849 jobs on CLIP** to calculate and save **probability arrays**.
AUC evaluation	Adapt model loading. Keep the test data and metric code fixed and calculate the AUC.
Wasserstein distance	Change the input directory. Reuse the existing calculation.

The agent handled both script adaptation and job submission.

A comparison figure with an explicit scope

Model	Training setup
Single GNN	Nominal samples, all 35 features
Nominal dual GNN	Nominal samples
Varied dual GNN	Training with variations

- Compare **four processes** on the same underlying nominal dataset.
- Show relative variation from nominal versus output probability.
- Preserve the existing plot fonts and font sizes.
- Keep Fig. 6 and add a separate comparison figure.

The agent plotted the comparison using the results it obtained during training and validation (Slide 7).

Drafting the reviewer response

1. **I updated the manuscript with the new results and discussion**.
2. **I marked the changes in red** and supplied the latest `.tex` file.
3. The agent used that version with the referee report and `reply.txt`.

Condensed response-drafting prompt

"Using the referee report and this revised manuscript, draft a response to every referee comment. The file `reply.txt` provides the response template. Please edit `reply.txt` directly."

The response should describe changes already present in the paper.

A reply tied to the manuscript

Referee request, paraphrased

Clarify the difference between the leading FAIR-HUC solution's unbinned likelihood fit and our binned final inference.

Illustrative response

Thank you for the suggestion. We have clarified in Section 5.3 that the leading FAIR-HUC solution employed an unbinned likelihood fit, whereas our method uses binned classifier outputs for the final inference, and we now explicitly note that the unbinned treatment may provide additional gains in statistical precision.

Illustrative wording based on the documented edit.

Review points: direct answer, actual change, correct location, cautious interpretation.

Scientific judgment stayed with me

What does the added baseline establish?

The single GNN and the varied dual GNN differ in both architecture and training setup. Their comparison alone cannot isolate the effect of perturbations during training.

Before finalizing the response, I needed to check:

- Does every reply address the referee's actual question?
- Does it describe the analysis we performed?
- Are numerical claims and manuscript references correct?
- Does the wording stay within what the comparison supports?

First experiences and limits

What helped in this case

- Direct edits across manuscript and analysis files
- Reuse of existing scripts for repetitive tasks
- A response draft based on the revised manuscript

What I would document better

- Code versions, configurations, and seeds
- Tests, job IDs, and completion logs
- Output counts and final metric values

Scope of this experience

The benefit was qualitative. I did not measure a time saving, and the notes do not provide a complete reproducibility record.

A prompt structure I would reuse

Element	Example from this revision
Context	The referee requests an additional single-GNN baseline.
Requested change	Adapt and train the existing model as a single GNN using all 35 features.
Constraints	Use nominal samples only. Preserve the deterministic branch's message-passing layers and activation functions.
Deliverables	A working single-GNN implementation, an adapted training script, and trained model weights.

For the revision as a whole:

- Work through one reviewable task at a time.
- Inspect the actual artifact or job output before proceeding.
- Draft the reply from the latest manuscript, then review every claim.



Thank You!