# (Mis)Adventures of GenAI in the Scientific Workflow

 Joshua Bloom Follow 5 min read · Dec 29, 2025



Astronomers are very comfortable adopting and coöpting tools built elsewhere to further our science (c.f. the telescope, CCDs, random forest, Bayesian inference, Python, etc.). So it was natural for me to try out generative AI tooling to assist in the development of my academic work. In this post, I highlight some of the trials and tribulations in coding and paper writing for academic work.

**Vibe Coding in the Early Days (Spring 2023):** Stepping gently into the fray, and already using ChatGPT to answer stackoverflow-like questions about code, in March 2023 I thought it would be useful to have a way to save ChatGPT chats in a way that could be properly referenced in academic work. *Maybe something interesting happened therein and we'd want to save it for posterity and cite it*. So I thought I'd ask ChatGPT for help making a Chrome Extension to generate a DOI from a saved chat. To make the chat persist, I decided to upload it to Figshare as part of the workflow. The whole coding

exercise was a cut-and-paste tussled affair—with me copying error messages into the chat window, trying it out, pushing to GitHub, etc.. Knowing little about browser extensions I had to rely almost entirely on ChatGPT to help me out of rabbit holes. The back and forth was painful but I got the basic functionality of **DOIit** working to the point I used it to capture my interactions and the responses. Here's a meta example where I'm asking for help writing a README to install DOIit (DOI: 10.6084/m9.figshare.22350943; "A conversation using OpenAI chat: 03/28/2023 14:54:06"). The whole exercise validated some of the initial promise of GenAI, but took longer than I thought it would to get something functional. I also didn't create a mock test suite so all my testing basically polluted my ORCID listings forever!

**Code Assistants (Fall 2024 — Spring 2025):** I learned about Cursor on hackernews and my first experiences were amazing. Adding rich and NumPy-style docstrings to Python code was the first power up. Next, it was code-snippet completion and smart variable renaming. Then, in December, I had a bold but wacky astronomy idea and went all in on Cursor. This time, I asked for also help rewriting and modernizing some C code from other repos and building Python bindings with Cython. The parallelism approaches suggested in Cursor worked at the function level but was definitely not optimized for batch processing over the entire codebase. In the end, I found a big old null result after confronting data from the whole sky (with jobs running for weeks on AWS). But by Spring 2025, I was convinced that Cursor was at least 5x-ing my coding productivity.

**Academic Paper Writing with GenAI (Fall 2025):** In learning about some of the progress that frontier models were making in axiomatic fields, like

math and theoretical physics, I wanted to experiment with my own work. I *briefly* went down the "let's solve a Millennium problem!" spiral (Collatz conjecture) before coming up for air, realizing it would be better to stick to a problem I could understand how to evaluate the correctness of. Rather than start from scratch, I decided to resurrect an unfinished paper of mine that had languished in the hopper of "good ideas, couldn't push it out" (this hopper is large, unfortunately 😰). About 15 years ago I had envisioned a new method for measuring distances in the nearby universe but required specialized know-how to manage a complex and careful hierarchical Bayes setup (ie., stats beyond my comfort zone). An astrostatiscian colleague and I got an initial start on the problem about 12 years ago but ran out of steam.

To make progress on the paper, instead of coding, I decided I would use only use natural language—my first **vibe science** experience. Feeding the draft to ChatGPT, I asked for help for the next steps and to fill in some of the TBDs we had left as breadcrumbs more than a decade ago. The response was heartening ("*Great project.*") and the updates to the LaTeX seemed spot on. At the end of each update I was prompted to do more ("*If you want me to tweak tolerances, add Student-t into EM by default, or add ARI-based label-invariant checks in tests, say the word and I'll roll those in too.*"). But it became clear we were just layering in more and more complexity without homing in on the simplest, correct analysis path.

So I brought in Claude Code, running it within conductor.build (now easily my favorite tool of 2025), and started to create toy datasets to try out the algorithms, build a test suite, and added GitHub actions to avoid regressions merging onto my branch. I started Oct 11 and within a dozen PRs, I was feeling pretty excited about the progress. I started hitting my daily limits on

Claude (and upgraded to the Max version)! I brought in AI code review to help; enabling CoPilot on GitHub was easy.



But then, after a week of vibe grinding, I started to get this nagging sense I was being slowly led into a state of stuporous acquiescence, that the whole package was working even if I couldn't understand all of it. CoPilot was finding a few bugs but we were going off the rails. So I started asking questions: "*Is the code doing what the paper is saying it's doing? Why is that fit to that line with simulated data so perfect?"* It turns out no, the paper and code had diverged. The line fits were cheating, making use of the "true" data rather than the noisified ones. I needed to use the full suite of professor/mentor tools I had honed over three decades to coax us back, calling foul on figures that looked just a little too good and generating new toy datasets that challenged baked in assumptions. "*What about this reference? Didn't say they the opposite?*" "*Oh you're right…*"

A month and a half later, 62 PRs in, I got to a paper draft and codebase that I thought my human co-authors could finally review. It remains to be seen whether this vibe science experiment work will survive their scrutiny and if we ever submit the paper for peer review.

**Conclusions**: At first blush, it might seem that I'd be turned off after all these (mis)adventures of GenAI in the scientific workflow. To be sure, I am

a little prickly: I was careening down a steep hill with no training wheels and no established norms. For vibe science to work, we need built-in guardrails like:

- constant critical reviews from different LLMs than those producing the work, perhaps fine-tuned on the sub-domain at play,
- unbreakable contracts that do not allow the algorithms described in papers to diverge from those in the accompanying code,
- approaches (e.g., specialized agents) that keep papers tethered to the correct and best citations for that unit of work.

I feel like these are all solvable. Looking back, the derivative of what's possible have been so positive it's hard not to feel optimistic about the future.

Vibe Coding | Astronomy | Academic Research | Claude Code | ChatGPT


**Written by Joshua Bloom**

721 followers · 193 following

Astrophysics Prof @UCBerkeley; wise.io co-founder (->GE); Dad