Lecture VIII — Your Research Project

AI-Assisted Programming for PhD Researchers

Dr. Tobias Vlćek
Helmut Schmidt University — September 2026

Research Code Is Different

The Artifact Is the Claim

In research, code correctness *is* result validity.

. . .

- A bug here is not a crash but a wrong Table 1
- Remember Lab 3: one silent `fillna` shifted every summary statistic
- Nobody sees a stack trace; they see a number that looks fine and is not
- The code and the result stand or fall together

Reproducibility Checklist

Four things make a result someone else can rerun:

- **Pinned environment** — `uv lock`, `renv`, or `Project.toml`
- **Fixed seeds** — every random draw is deterministic
- **Data versioned or archived** — findable and persistent [1]
- **One command** from raw data to every figure

Notebooks: Handle with Care

- Hidden state and out-of-order execution quietly wreck reproducibility
- Of ~1.16M public notebooks Pimentel et al. mined, only **~4%** reproduced their own saved output [2]
- **36%** ran cells out of order — a contributing cause, not the failure rate itself [2]
- Fine for exploration; the *record* should be a pipeline — like the one you built

Agents and Reproducibility

Agents are good at exactly this cleanup work:

- “Make this run end-to-end from scratch in a clean environment”
- “Pin my dependencies and generate a lock file”
- “Find the hidden state in this notebook and flag out-of-order cells”

AI Beyond Code

Literature Research with Agents

The deep-research pattern: search → read → synthesize with citations.

. . .

- Point an agent at a question; it fans out, reads, and summarizes with sources
- Genuinely useful for mapping an unfamiliar field fast
- A strong *starting* map — not a finished bibliography

The Catch: Invented Citations

Models produce references that are confident, well-formatted, and fake:

- **55%** of GPT-3.5's citations were fabricated; GPT-4 still **18%** [3]
- Even the *real* ones carried errors — **43%** (GPT-3.5) / **24%** (GPT-4) [3]
- Newer is better, not safe: GPT-4o still ~**20%** fabricated, worst on niche topics [4]

Spot the Fake

Same format, same confidence — one exists, one does not:

```
Fabricated: Smith, J. (2019). "Deep learning for
  epidemic forecasting." Nature Methods, 16(4), 331–339.

Real:       Pimentel et al. (2019). "A large-scale study
  about quality and reproducibility of Jupyter notebooks."
  MSR 2019, 507–517.
```

You cannot tell from the text — so every citation gets verified.

The Verification Discipline

For every AI-suggested source, three checks:

- **DOI resolves** — the link actually goes somewhere
- **Paper exists** — real venue, real authors, real year
- **It says what was claimed** — read the abstract, minimum

. . .

AI does not remove the work; it moves it — from finding to verifying [5]. Yesterday's Lab 5 rule, now for your thesis.

Drafting Text with Skills

A clunky sentence through `conciseness` + `academic-grammar`:

```
Before: It should be noted that, in order to obtain the
  results, we made use of a large number of samples.

After:  We used a large sample.
```

The skills suggest; you decide what stays.

Cross-Check Everything

AI drafts are drafts: verify claims sentence by sentence, numbers against sources, citations against papers.

. . .

> ⚠ Warning
>
> The deeper risk is producing more while understanding less — "illusions of understanding" [6]. You sign it, you own it.

Institutions Are Responding

- AI agents in research are now a policy topic — journals, funders, and institutions are writing the rules [7]
- The disclosure and verification habits from Lecture VI keep you ahead of them
- Do the honest thing now; the requirements are only catching up

Your Capstone

The Assignment

Tomorrow you build. Today you scope.

. . .

- Deliverable: a working demo **plus** a presentation, Friday afternoon
- Judged on the four points: demo works, process reflection, honest challenges, transferable insights
- The same four you have seen all week — no surprises

What Makes a Good Capstone

- Comes from your own research, a real problem you have
- Core is finishable in ~5 hours, with stretch goals beyond it
- Has a verifiable output — a figure, a table, a passing test suite
- Uses the full workflow: spec → plan → implement → verify

No Project? Option A: Literature Manager

A BibTeX/DOI tool built from the Lab 5 pieces:

- Parse a messy `.bib` or a list of DOIs
- Fetch metadata via the MCP server
- Deduplicate, then emit a clean reading list
- Full brief on the capstone page

No Project? Option B: Parameter Study

A small simulation you sweep and plot:

- Pick a model — a queue, an epidemic, your choice
- Sweep two parameters over a grid
- Produce result plots and a findings table
- Full brief on the capstone page

Scope It Now

- Open the capstone page and write your spec with the interview pattern
- I circulate while you work
- By 16:30: spec + plan reviewed by a fresh agent session or a neighbor

The Capstone Project

Continue Your Journey

Next Up

- You now know why research code carries a higher bar, and how to keep AI honest in it
- Tomorrow, **09:00**: build day — bring coffee
- The Capstone Project — scope yours this afternoon
- Lecture IX — Judgment and Presenting — tomorrow morning
- Course literature and references

Bibliography

[1] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.

[2] J. F. Pimentel, L. Murta, V. Braganholo, and J. Freire, "A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, Montreal, QC, Canada: IEEE, May 2019, pp. 507–517. doi: 10.1109/MSR.2019.00077.

[3] W. H. Walters and E. I. Wilder, "Fabrication and errors in the bibliographic citations generated by ChatGPT," *Scientific Reports*, vol. 13, no. 1, p. 14045, Sept. 2023, doi: 10.1038/s41598-023-41032-5.

[4] J. Linardon, H. K. Jarman, Z. McClure, C. Anderson, C. Liu, and M. Messer, "Influence of Topic Familiarity and Prompt Specificity on Citation Fabrication in Mental Health Research Using Large Language Models: Experimental Study," *JMIR Mental Health*, vol. 12, p. e80371, Nov. 2025, doi: 10.2196/80371.

[5] H.-P. (. Lee *et al.*, "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, Apr. 2025, pp. 1–22. doi: 10.1145/3706598.3713778.

[6] L. Messeri and M. J. Crockett, "Artificial intelligence and illusions of understanding in scientific research," *Nature*, vol. 627, no. 8002, pp. 49–58, Mar. 2024, doi: 10.1038/s41586-024-07146-0.

[7] “AI scientists are changing research — institutions, funders and publishers must respond,” *Nature*, vol. 651, no. 8107, pp. 853–854, Mar. 2026, doi: 10.1038/d41586-026-00934-w.