



AI at Your Side

The Student's Guide to Smarter Learning

Forthcoming from Oxford University Press

Sebastian Galiani
Raul A. Sosa

Tulane University



Book website

AI Adoption

Three surveys of students and knowledge workers, 2024 to 2026

95%

of UK university
students use AI
HEPI, 2026

85%

of US college students
used it for coursework
Inside Higher Ed, 2025

75%

of knowledge workers
use AI at work
Microsoft & LinkedIn, 2024



AI can make you *sound* smarter without making you **think** better, and from the outside the two look the same.



The book shows how to use AI so that your thinking improves too, and how to tell whether it has.

The Book's Four Principles



Augmentation, not replacement

Build on what you
can do without
skipping the
learning that
matters most.



Appropriate automation

Decide, task by
task, when to rely
on it and when to
do the work yourself.



Ethical AI literacy

Know its limits,
biases and potential
harms, and use it
responsibly.

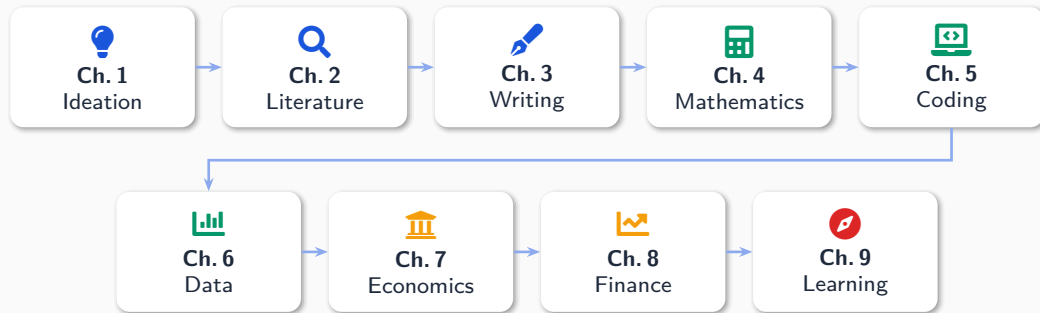


Intellectual ownership

Say why you asked,
why you trust the
answer and which
parts are yours.

The Book

The Book at a Glance



+ Appendix: *How Modern AI Systems Work*

Research and Writing

Ideation

From a topic to a project: a precise question, a testable hypothesis, a data plan and its limits. The chapter also shows how to widen the options before narrowing them.

Literature

Map the field before searching, pull the same four items from every paper (claim, method, finding, link to your question) and keep track of where authors disagree.

Writing

The CREO map: claim, reasons, evidence, objection. The chapter also sets an integrity standard for what you can delegate, what you must disclose and what is out of bounds.



AI can cite **a paper that does not exist**, or cite a real one **for something it does not say**. Its syntheses also **manufacture consensus** where the literature is divided, and that disagreement is often where your own contribution starts.

The Quantitative Toolkit



Mathematics

How AI can find the exact step where your reasoning broke, help you plan the approach before computing and check a finished proof for the gap a grader would flag.



Coding

A ladder of growing responsibility: a **tutor** that explains, an **assistant** that reviews, an **agent** that runs whole workflows, and finally you, as the **builder**, writing code that calls the model.



Data

The hard part is deciding what to look at and whether a pattern is real. Choosing which comparison a chart shows is part of your argument.



The *protégé* effect: instead of asking the model to explain, you *explain* the method to it, and it marks what is right, what is wrong and what is missing.



In code and data work, the most dangerous mistake does not crash the program. The code runs and the output looks fine, but it answers a question you did not ask.

Economics, Finance and Learning

Economics

You work through a formal model with AI, one step at a time, until you can say what each condition means. The chapter does this for public debt, Becker and Spence.

Finance and strategy

There is no right answer written down anywhere, because the decision depends on assumptions you choose and have to defend. The model's job is to question them and argue the other side.

Learning

Each time you use AI, you choose between getting the assignment done and building the skill it was meant to teach. The chapter also names four things the model cannot supply: domain depth, the will to do the hard thing, judgment and metacognition.

Chapter Structure

Chapters 1 to 8, each adapted to its subject



1. The domain's *harness*

A prompt with rules that you save and reuse. Most chapters build their own, because what counts as a good answer differs between, say, a literature review and a valuation.



2. Worked examples

Prompts, the model's answers and what to take from them. Most chapters open by setting a weak prompt against a strong one.



3. A case study

A real problem, with the false starts left in. Most cases do not work out the first time, and the chapter shows why.



4. Where AI fails in the domain

Where its errors are hardest to spot, because the wrong answer reads well and has the right form.

Examples from the Book

Public debt

From a broad topic to a plan with sources, an empirical strategy and robustness checks.

Spence's signal

You write the separating-equilibrium conditions and the AI checks them. Then comes the puzzle of why education that produces nothing still pays.

A DCF under test

Ferrari valued by discounted cash flow, with a counter-thesis, a sensitivity analysis and a pre-mortem on what would make it wrong.

Becker and time

The Lagrangian solved, and then the harder step of saying what the first-order condition means.

A 10-K in two hours

Reading Costco's annual report with the finance *harness*, without handing the conclusion to the model.

Charts that fail

The cluttered chart that answers nothing, the misleading dual axis and Simpson's paradox.

Each one shows the prompts and what the model returned, and most say where not to trust it.

AI in the Classroom

Galiani's 2026 courses in development and health economics

 **The classroom rule**, adapted from Tyler Cowen: about two thirds of the course is the subject itself, and one third is AI used inside that subject, to learn it and to do research in it.

Course rules

- ✓ Prompts follow a set structure
- ✓ Process counts as much as product
- ✓ Hallucination checks are required
- ✓ An AI Process Log, a record of intellectual ownership, replaces “show your work”
- ✓ Exams cover the problem sets, so delegating them does not pay

In the problem sets

- ➔ Give the model the case and paste its full answer
- ➔ Check one fact against another source and mark what the model simplified
- ➔ Attach the exact prompt and model version
- ➔ Mark where the model takes a side

Capabilities and Limits

Language models and the evidence on productivity

Language Models

Next-word prediction, repeated



The model picks the next word given everything before it, adds it to the text and repeats. Its reasoning, its code and its citations all come out of that one loop.

✓ Its strength

Trained on more text than you will ever read, it can connect ideas across fields you would not have put together.

⚖ Its limit

The ranking comes from patterns in the text it was trained on. Nothing in it checks whether a word is true.

Confident Errors



1. It was trained to assist

A second round of training, graded by people and by automatic tests, turns the text predictor into an assistant. As a side effect, the grading tends to reward agreeing with you and to prefer a confident guess to “I don’t know.”



2. It writes without checking

Where the text it learned from is dense, the most probable word is usually the true one. Where the text runs thin, it often writes something just as fluent that is false. That is a hallucination.

Where the errors cluster

citations

numbers

names

dates

anything recent

anything obscure



It often sounds just as sure when it is wrong, so asking it afterward whether it is sure is not a check. If you want it to push back, ask for it. Web search helps by putting real sources in front of the model, but the model can still misread them.

Productivity Gains

Three studies comparing work with and without an AI assistant

BCG consultants

+30%

higher-quality work, 25% faster

Dell'Acqua et al., 2026

Customer support

+15%

more cases resolved per hour

Brynjolfsson et al., 2025

Software developers

+26%

more tasks completed

Cui et al., 2026

→ Only the first study included a task chosen to fall just outside what the AI could do well.

The Jagged Frontier

Evidence from BCG consultants and experienced developers

 **The jagged frontier** is the edge of what AI does well. Tasks that look equally hard can fall on either side, and past the edge people can do worse with AI than without it.

BCG, the task just outside the frontier

19 points

lower share of correct answers with AI than without

Dell'Acqua et al., 2026

Developers who knew their code well

19% longer

with early-2025 AI tools, in a randomized trial

Becker et al., 2025

 They could not tell which side they were on. At BCG, the wrong answers still read as convincing. The developers misjudged their own speed:

Expected: **24% faster**

→ Took: **19% longer**

→ Believed afterward: **20% faster**

Execution and Judgment

When one input gets cheaper, its complements become more valuable

↓ Cheaper with AI

- Writing code
- Searching and summarizing
- Algebra and derivations
- Running simulations
- Drafting

↑ Worth more with AI

- Choosing the question
- Choosing the assumptions
- Choosing the model
- Interpreting the result
- Knowing whether it is right

 **Cognitive debt** (Kosmyrna et al., 2025, preprint) builds up when work carries your name but the tool did the thinking. It comes due when the tool is gone, or when it is wrong and you cannot tell.

A Prompting Demo

One research question about New York rents, asked two ways

The One-Line Prompt

The first of two new chats

Claude Sonnet 5 · clean session · Oct 9, 2026 condensed

Give me some research ideas on why rent in New York is so expensive.

 Claude

Research Ideas: Why Is Rent in NYC So Expensive? [...]

1. Supply-Side Economics [...] **2. Demand-Side Factors [...]** **3. Policy & Regulation [...]** **4. Market Structure [...]** **5. Comparative/Historical [...]** **6. Data Sources to Use [...]**

Want me to help you **design a specific research question**, draft a **literature review outline**, or pull together **data analysis code** (e.g., scraping/analyzing public rent datasets) for one of these angles?

A clean session is a new chat with no tools (so no web search) and no memory of other chats.

⚠ It gave a good menu, eighteen angles in five families plus data sources, and nothing in it looks wrong. But it cannot choose for you. It does not know your field, deadline or format, and it did not ask.

The Same Prompt in Two Chats

The first chat (top) and a second new chat (bottom)

Claude Sonnet 5 · clean session · Oct 9, 2026

condensed

 Claude

[...] **Construction costs:** Compare NYC construction costs per unit vs. other major cities (labor, materials, union requirements, environmental review (ULURP) timelines). [...]

✗ **Chat 1 calls ULURP environmental review.** ULURP is the city's public review of land use. Environmental review is a separate process, CEQR.

Claude Sonnet 5 · clean session · Oct 9, 2026

condensed

 Claude

[...] Quantify how ULURP (Uniform Land Use Review Procedure) delays compare to other cities and their cost impact. [...]

✓ **Chat 2: right.** ULURP is the Uniform Land Use Review Procedure: community board, borough president, planning commission and city council.

The second chat gave nineteen numbered angles in five families, plus three methods, then an offer to keep going.

⚠ Both chats name the same procedure, and one is wrong about what it is. Neither flagged any doubt, and the request did not ask for one. If you ask only once, nothing in the answer tells you whether you got the right one.

The Five Fields of a Prompt

The details the one-line prompt left the model to guess



Context

Who you are, what you can do and what you already suspect.



Task

The one thing you want done.



Constraints

Time, data and what is out of scope.



Output

What the answer must contain, item by item.



Verification

What it must flag so you can check it, and what you will check yourself.



Most people type a prompt like a search query. These five fields turn it into an **instruction set**.

→ Save that instruction set for a task you repeat and you have what the book calls a *harness*.

The Five-Field Prompt

The rent question from slide 19, rewritten



Claude Sonnet 5 · clean session · Oct 9, 2026

reply on the next slide

Context: I'm a second-year economics major in New York. I've taken intermediate micro and intro econometrics, so I can run OLS and a simple difference-in-differences in R. My hunch is that the 2023 crackdown on short-term rentals took thousands of Airbnb listings off the market, and rents still haven't come down.

Task: propose research designs that could explain why rent in New York is so high.

Constraints: one semester, about four hours a week. Only free data I can download myself. No surveys, no scraping, nothing beyond OLS and difference-in-differences.

Output: two designs. For each one: the question in one sentence, the dataset (name and link), the regression, and the main threat to it.

Verification: for every dataset, tell me the first year and the smallest geography it covers for New York. If you are not sure of something, write UNSURE instead of guessing. I will open the data myself.

We highlighted the three phrases that do the most work.

The Reply to the Five-Field Prompt

Design 1 from the second five-field chat, condensed

● ● ● Claude Sonnet 5 · clean session · Oct 9, 2026 condensed

 Claude

Design 1: Within-NYC exposure design [...] **Question:** Did New York City ZIP codes with more Airbnb listings before Local Law 18 (effective September 2023) see rents rise faster afterward than ZIP codes with fewer listings?

Datasets: Inside Airbnb [...] Zillow Observed Rent Index (ZORI), ZIP code level [...]

Main threat: Parallel trends. High-Airbnb ZIPs (Manhattan core, parts of Brooklyn) were likely already on a different rent trajectory before the law [...]

Verification [...] I believe the public ZORI series begins January 2015, but I am not fully certain of [...] whether every NYC ZIP has data back that far [...] **UNSURE, verify by opening the file.**

We sent the prompt from slide 22 to two new chats. Both proposed this design, and both marked the start of Zillow's series UNSURE.



The fields

The **context** brought in the hunch, and design 1 studies the crackdown.

The **output** field is why each design came with its data and its main threat.

The **constraints** kept both designs to OLS and difference-in-differences.

The **verification** field made it say where it was unsure.

Checking the Zillow Data

The coverage it marked UNSURE, checked against the file

Zillow's rent index, New York

ZIP codes in the file, and how many have data from before the law (September 2023).

	ZIPs	before the law
Manhattan	45	43
Brooklyn	37	33
Queens	38	22
Bronx	18	10
Staten Island	11	5
All	149	113

 Our check: Zillow ZORI by ZIP, through August 2026.

The doubt was in the right place

The series does start in January 2015, but only 55 of the 149 ZIPs have data that early.

36 ZIPs have no “before”

Their data start after the law, so they cannot enter the design. Thirty of them are in Queens, the Bronx and Staten Island.

 The five fields **made the answer checkable**, and the model flagged its doubt about the data, as the request asked.

Neither chat caught that Inside Airbnb's files from before the law must be requested and may cost money.

Agents

Chatbots, Agents and the API



Chatbot

You write,
it answers.

ChatGPT, Claude, Gemini



Agent

It plans and runs the
work. You check it.

Claude Code, Cursor, Codex



API

You write code that calls
the model, at scale.

OpenAI, Anthropic, Google

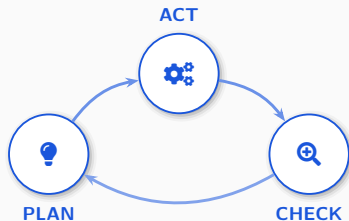
>_ Agent example: “Inspect this dataset, run a fixed-effects regression, produce a chart.” The agent runs the whole sequence.



API example: a Python script sends 200 abstracts through the same fixed prompt and writes each one's hypothesis, method and findings to a CSV.

Plan, Act, Check

An agent is a model that works toward a **goal** with little supervision: it plans the steps, uses tools (files, the web, code) and checks what they return.



Chatbot and agent, side by side

	Chatbot	Agent
Unit of work	a reply	an assignment
Works from	what you give it	what it goes and finds
Ends in	text	files, analyses, reports

! An agent can check that its code runs and that the file exists. Whether the answer is right is still yours to check.

Agents in Research

From building the data to replicating a paper



Build the dataset

Assemble it from primary sources, with every coding decision documented and every value tied to its source.



Stress-test your own result

Rerun it with other reasonable controls, samples and definitions, and see whether it holds.



Review the draft

Flag inconsistencies and errors in the text before your own final read.



Replicate a published paper

Download the replication package, rerun the code behind each table and mark which ones reproduce.

All four are structured cognitive work, with clear rules and a result that a person can check against a source.

Robots



Decades of hardware

The first industrial robot went into a General Motors plant in **1961**. Arms, motors and sensors have improved for more than sixty years, but robots almost always repeat a fixed sequence in a controlled setting.



Moravec's paradox

Games like checkers proved easier for machines than a one-year-old's **perception and movement**, which stayed the bottleneck for decades. The same kind of model that writes reports, trained further on action data, is starting to ease it.



Already in use

Warehouses and logistics, self-driving vehicles, camera-guided arms on factory floors. On farms, driverless tractors and sprayers that hit only the weeds. Humanoids are in pilots and their first small paid deployments.



Putting a robot on a new task has meant redoing the engineering around it, and that engineering is most of the installed cost. If one model can handle new tasks without that work, the **fixed cost** collapses, and so does the minimum scale at which a robot pays off.

AI and Work

Two ideas from the economics of technological change

01 Non-routine tasks are exposed now

A job is a bundle of tasks. The classic task framework treated non-routine cognitive tasks as safe from automation, and AI can now do some of them.

02 The gains come with redesign

Electric motors reached factories in the 1890s, but productivity jumped only in the 1920s, when plants were rebuilt around one motor per machine.

When AI takes over a task, value moves to the tasks around it (slide 17), and knowing whether the output is right becomes the bottleneck.

The Golden Rule

For delegating research work to agents

✓ Where to start

- ✓ Document-based tasks with clear rules
- ✓ Work you repeat across many cases
- ✓ Anything you can check against a source

⚖ The golden rule

People **design the instrument**, a prompt with rules for one concrete task, and **audit the result**. The agent does the steps in between.

🔒 Never delegated

- 🔒 The goal and the standard of quality
- 🔒 The final judgment on what gets published
- 🔒 Responsibility for the result

🎓 **You can only judge work you know how to do.** That makes the golden rule an expert's rule. A student first does the work, often enough to judge it, and only then hands it over.

Agents at Scale

Datasets built by agents, one unit at a time

Deep Research on a Loop (DRIL)

Afonso, Galiani, Gálvez & Sosa (2026)

 Building datasets from primary sources is one of the costliest tasks in empirical economics. **DRIL** writes the coding rules into a frozen **instrument** and runs it on each case separately.

Two stages

Design. The researcher sets the objective and brings any reference material. An agent drafts a codebook from them (the rules for what to measure and how) and flags the ambiguities. The researcher resolves them and freezes the design.

Implementation. Agents run it on every case and log sources, gaps and uncertainty.

One real run

A 2025 partial update of the Global Tax Expenditures Database: **86 jurisdictions**, about 40% of the 218 in the database, for about **\$22.50** at API prices. It produced quantitative estimates for 66 and recorded a gap, rather than a guess, for the other 20.

 Before anyone uses the data, a separate verification stage has to check it.

A Loop vs. a Single Agent



One agent, one answer

Ask an agent about one country and you get a tailored answer. Ask about the next and the format and sources shift, so the coding is no longer comparable.



The same instrument, in a loop

The agent applies one codebook to every unit, so every value is coded by the same rules and the output is a single **dataset**.

Like a pre-analysis plan, freezing the instrument before the run puts every design choice on the record, so others can audit the coding and reproduce it.



The loop suits whole populations whose information is public, even when scattered across sources and languages, and whose variables fit in a *codebook*. Coding them by hand is costly, and that cost limits which questions get studied.

A Sample Country Brief

A mock-up of a report the same loop could produce

Country X

Fiscal and debt brief, illustrative draft

Debt/GDP: 62.4% [IMF, *Fiscal Monitor* 2025]

Fiscal balance: -3.1% of GDP [Ministry of Finance, 2025 budget]

External debt, share of total: 48% [World Bank IDS]

Tax expenditures: 34 measures, $\approx$ 2.1% of GDP [Global Tax Expenditures DB]

⚠ Gap flagged: subnational debt is not broken out in public sources. The brief leaves the figure blank.

Analyst note: check the Q4 balance against the revised budget before release.

The numbers are invented and do not come from the paper's run. Each figure cites a source, as the loop's output would.

AI and Productivity Gaps

Evidence from a randomized experiment in Argentina

The Experiment

Cruces, Fernández Meijide, Galiani, Gálvez & Lombardi (2026)

Most studies look at AI *inside* firms, where hiring has already narrowed who is in the room. We ran a **randomized online experiment outside firms**, on adults in Argentina, to test whether AI narrows productivity gaps by **education**.

The design

- **1,174 Argentine adults**, aged 25 to 45
- An incentivized, workplace-style **problem-solving task**
- Randomized: a generative-AI assistant and a short tutorial, or **neither**

The follow-up, without AI

After the main task, everyone completes a second module **without AI**. It shows whether any of the gain carries over, or whether it was **delegation** that disappears once the tool is gone.

Main Results

Treatment effects by education level

75%

of the education gap in task performance closed, with AI access and a short tutorial

- ✓ AI improves performance for **everyone**, with much larger gains for those with **less education**.

⚠ **Without AI, a sizable gap re-emerges.** Within the AI group, those with more education used the tool somewhat more effectively, and follow-up scores were high only where people also spent real time on the task. Both patterns are associations, since only the AI treatment was randomized.

What you get from the model depends on what you bring to it.

In the New York demo, a hunch and a few limits changed the answer, and checking it against the source was still our job.

AI at Your Side

Sebastian Galiani · Raul A. Sosa
Forthcoming from Oxford University Press

Thank you!



Book website