A field guide

Discovery with Machines

How to research, test and publish when your collaborator is an AI

INCOPREA RESEARCH

Contents

1. Preface
2. What a machine can and cannot discover
3. The economy of ideas, and why volume fails
4. Before you generate: search first
5. The shape of a testable claim
6. Designing experiments that can fail
7. The counting rule: identifiability before precision
8. Seven defects that survive review
9. Reading your own results honestly
10. The corrections register
11. Writing it up so it can be checked
12. Publishing, licensing and being found
13. What to do when the answer is no
14. Working with a machine collaborator
15. Credit, disclosure and the ethics of assistance
16. Three worked cases
17. Objections, answered
18. Techniques: the arithmetic of not fooling yourself
19. Techniques: designing a measurement that can decide
20. Techniques: public datasets, honestly used

21. Building your own error corpus

22. A year of practice

23. Where this came from, where it is, where it is going

24. Doing research without a laboratory

25. Money, time and the shape of a small research practice

26. Appendix A — one-week protocol

27. Appendix B — claim checklist

28. Appendix C — fifty-eight tools

29. Appendix D — glossary

Preface

This is a short book about a long problem: how to tell the difference between a machine that has read everything and a machine that knows anything.

It is written for one person — someone using AI to do research, alone, without a laboratory's worth of colleagues to check their work. That person gets enormous leverage from these tools and almost no error correction, and the failure modes are quiet. Nothing crashes. The numbers look right. The prose is fluent. The errors surface months later, in public, or not at all.

Everything here was learned the expensive way, in a single programme that produced six thousand generated hypotheses, fifty-eight tools, two real datasets and sixteen published corrections. Eight of the strongest claims were checked against the literature at the end. All four were already known.

> *This book is what I would have wanted on day one: the order of operations, the shape a claim needs, the tests that cannot fail, and the discipline of publishing the corrections.*

Each chapter is short and stands alone. If you read only three, make them three, six and nine — search before generating, count before measuring, and publish what you got wrong. Those three account for

most of the difference between work that survives and work that merely impresses.

1. What a machine can and cannot discover

A language model is a very fast, very widely read colleague who is never tired, never embarrassed, and never checks whether the thing you are both excited about is already known. That combination is genuinely useful and genuinely dangerous, and almost every failure in this book came from mistaking one half of it for the other.

What machines do well: breadth. Ask for four hundred ways a measurement can mislead, and you will get four hundred. Ask for the failure modes of an instrument you have never used, and you will get a plausible, organised list in seconds. Ask for the three fields in which a particular statistical trap appears under different names, and the answer will usually be right, because that is a retrieval problem and retrieval is what the machine is for.

What machines do badly: knowing what is already true in the world. A generated claim is not a finding. It has the grammar of a finding and none of the warrant. The first law of this book follows directly.

> *The machine proposes; the world disposes. Any sentence that could be false must be made checkable before it is believed — and most sentences can be false.*

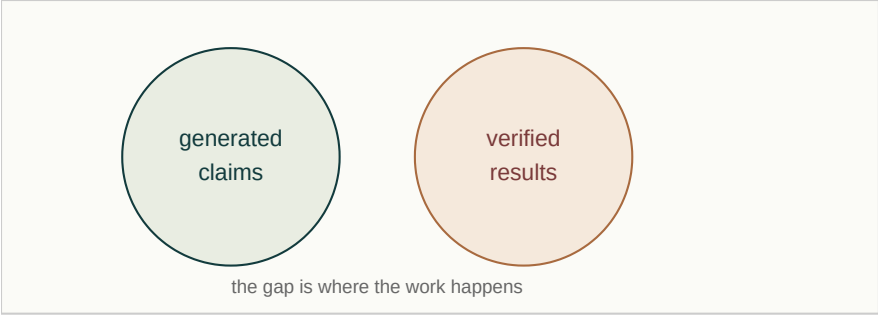


Figure 1. Generation and verification are different activities with different failure modes. Most of the effort belongs in the second circle.

2. The economy of ideas, and why volume fails

Ideas are cheap. That sentence is easy to say and almost impossible to feel, because producing ideas feels like work. It is not the same kind of work as testing them.

In the project this book came from, six thousand hypotheses were written, each with a mechanism and a test. They were organised across sixty domains, each entry carrying a distinct logical form. It was, by any measure of craft, a good body of work — and by the only measure that mattered it was worth very little. Then four of the strongest claims were checked against the literature, and all eight turned out to be known — one of them published under almost exactly the same title, decades earlier.

The arithmetic is brutal and worth internalising: if the strongest claims are old, the weaker ones are older. Volume does not compound. It dilutes.

Activity	Cost	Value per unit	Scales?
Generating hypotheses	seconds	very low	no — saturates immediately
Searching prior art	minutes	high	yes, up to a point
Designing a test	hours	high	poorly
Running the test	days	very high	no
Publishing the correction	hours	very high	no

Notice the shape: everything that scales is cheap, and everything valuable does not. That asymmetry is the whole economy of research, and a machine shifts it in the wrong direction — it makes the cheap column nearly free.

3. Before you generate: search first

The most expensive mistake described in this book is the order of operations. Search before you generate, not after.

Four queries would have redirected a week of work. Instead they came at the end, and their answer reframed everything: the field was crowded exactly where the ideas felt freshest, because familiar ground is where confident-sounding ideas come from.

A practical protocol, in the order that saves the most time:

Step 1 — Name the claim as a sentence

Write it in the form "X is caused by Y, testable by Z." If you cannot, it is not a claim yet.

Step 2 — Search for the sentence, not the topic

Topic searches return your own field. Sentence searches return prior art. Search the mechanism, the measurement, and the word "review".

Step 3 — Search against yourself

Deliberately search for the result you are most afraid of: that it is known, that it is false, that the effect is an artefact. The first page of results is usually enough to tell you which.

What the checks found, and what they do not prove

Eight claims from the strongest set were searched. All eight returned prior art:

Claim	Prior art found
Three-axis machine has 21 independent geometric error terms	published table of exactly that
Nuclear fuel ridging pitch tracks pellet length	1974 study on pellet length and cladding interaction
Radial diffusion coefficients from phase-space gradients are non-unique	active literature on that extraction
No interlaboratory comparison detects a bias common to all	proficiency-test consensus literature
Dim light damages more per unit exposure	microfading study of the reciprocity principle
Electro-osmotic flow reverses by excluded volume in narrow pores	reversal in charged nanopores, multivalent electrolyte
Standard detection limits understate the power requirement	detection-limit literature; the correction is standard
Peak width biases decomposition when the instrument is ignored	routine in spectroscopy and photoemission

This is a purposive sample and must be described as one. The eight were chosen because they were the strongest, and strong claims are the ones most likely to correspond to something real — and therefore the ones most likely to be already known. **Eight of eight tells you the prior-art risk for confident claims is high. It does not tell you the base rate across all claims, and any reviewer who points that out is right.**

> **How to measure it properly.** Draw a random sample — not the strongest — of, say, one hundred claims from the population you care about. Search each with a fixed protocol. Report the proportion with a confidence interval: eight of eight is consistent with anything from about 60% to 100%; eighty of a hundred is a measurement.

> Two days of work, and it converts a strong prior into a number you can defend.

Until that exists, treat the finding as what it is: a reason to search before generating, not a statistic to quote.

Step 4 — Record the null

Write down what you did not find and when. An absence recorded is a defensible statement; an absence remembered is a guess.

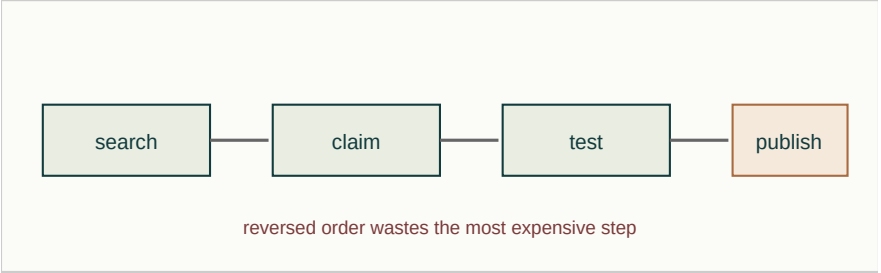


Figure 2. The order of operations. Search is cheap and disqualifying; generation is also cheap but feels productive.

4. The shape of a testable claim

Most claims fail not because they are wrong but because they are unkillable. An unkillable claim cannot accumulate evidence, so it never becomes knowledge. The fix is structural: give every claim a shape that admits a single decisive observation.

Six shapes carry most of the weight. Learn to recognise them, because a claim in one of these shapes is worth a hundred in none.

Bound

"No method can do better than X." Strong because a single counterexample destroys it, and because bounds transfer between fields.

Count

"There are at most N independent quantities here." The cheapest shape to test: count the parameters, count the measurements, compare. Most identifiability failures are count failures wearing a lab coat.

Identity

"A and B are the same thing described twice." Falsified by finding a case where the two predictions diverge — and when true, it collapses a literature into a paragraph.

Threshold

"The behaviour changes at X." Testable by bracketing. Watch for the version that is unfalsifiable because X is fitted after the fact.

Reversal

"The effect changes sign above Y." These are the most valuable claims in practice, because standard practice usually assumes monotonicity.

Artefact

"The effect is produced by the instrument, not the sample." Falsified by swapping the instrument while holding the sample fixed.

Shape	Cheapest test	Common failure
Bound	construct a counterexample	the bound is true only in one regime
Count	count parameters vs measurements	counting the wrong quantity
Identity	find a divergent prediction	the identity is a definition
Threshold	bracket the transition	threshold fitted post hoc
Reversal	measure both sides	the reversal is a saturation artefact
Artefact	swap the instrument	two instruments share the artefact

5. Designing experiments that can fail

A test that cannot fail is not a test. This sounds obvious and is violated constantly, usually by accident, because synthetic constructions tend to make the expected answer exact.

Two of five checks in one evaluation could not possibly have detected the effect they were aimed at: one used a numerical gradient dominated by noise, the other fitted an exponential that was exact by construction. Both returned confident-looking numbers. Both were worthless.

The negative control is the whole design

Before running the positive case, run a case where the effect must be absent. If the test fires there, it is measuring something else. In this book's own work, a derived-column detector passed its positive control and its negative control — because the negative control was a monotone inverse pair, and the detector was sensitive to correlation rather than to derivation. That single check would have found the defect a month earlier.

Pre-register the decision rule

State, in advance, what result would make you abandon the claim. If every possible outcome is consistent with the claim, you are not testing.

Two implementations, one quantity

The most productive defect-finding habit in this book: compute the same number two ways. A 28% error in a headline statistic was found only because two independently written scripts disagreed. Neither script was wrong about anything visible; the disagreement was the signal.

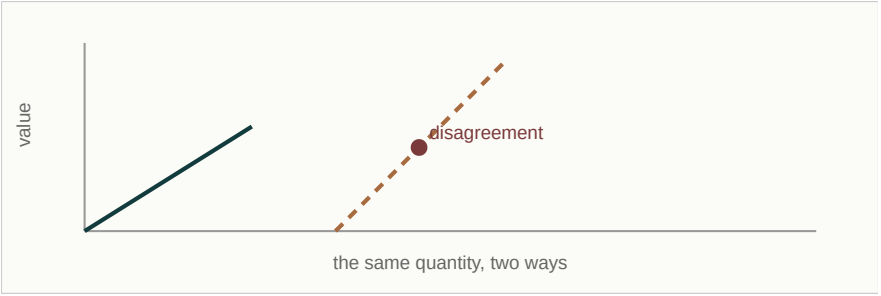


Figure 3. Two independent computations of one quantity. The systematic offset is invisible inside either script and obvious between them.

6. The counting rule: identifiability before precision

Before asking how accurately something can be measured, ask whether it can be measured at all. The answer is usually a counting question, and it is usually answered in minutes.

If you have K independent measurements and N unknown parameters, and K < N, no instrument quality will help you. The system is rank deficient and the parameters are not separately identifiable. What surprises people is the next step: at exactly K = N the system becomes identifiable in the algebraic sense and simultaneously useless numerically, with condition numbers in the billions.

Measurements K	Unknowns N	Condition number	Practical meaning
2–8	10	rank deficient	cannot separate at all
10	10	6.8×10^{9}	identifiable, unusable
12	10	3.2×10^{8}	still needs regularisation

Two consequences follow, and both are counter-intuitive.

Contrast beats repetition. Two measurements that weight the unknowns differently are worth more than ten that agree. A close pair of constraints can condition an order of magnitude worse than a separated pair.

The penalty belongs to the parameter count, not the measurement count. Adding a third measurement as a redundant observation improves conditioning. Adding it as a new free parameter, nearly collinear with the others, can degrade conditioning by two orders of magnitude. In one test the same third observation moved the condition number from 26.9 to 2,001.8 — purely because it carried its own offset.

> *Identifiability is a counting problem before it is a precision problem. Count first; measure second.*

7. Seven defects that survive review

These are the failures that pass peer review, pass code review, and pass the author's own re-reading — because none of them produces an error message.

1. The threshold comparison

A selection written `m >= mc`, where `mc` was computed by addition, silently excludes the values exactly at the threshold, because `4.35 + 0.05` is not 4.4. In a magnitude catalogue the excluded bin is the modal one, and the effect was a 28% shift in the headline statistic.

2. The derived column

A variable computed from another is not an independent measurement. Correlating against it produces inflated coefficients that vanish under scrutiny. Detect by regressing the log-ratio and checking the residual spread; flag below about 0.10 decades.

3. The unstated estimator

Two standard estimators of the same exponent can disagree by 10% to an order of magnitude, and the disagreement does not shrink with more data. Any single reported value hides a choice.

4. The mixture

A pooled statistic can be perfectly stable while its components move by a factor of three. Stability is not evidence of homogeneity.

5. The default convention

Library defaults decide results: whether columns are normalised, whether centring happens, what counts as a threshold. If the convention is unnamed, the result is unnamed too.

6. The clean test

A synthetic construction that makes the expected answer exact cannot detect the effect. Always include a case where the effect must be absent.

7. The unread literature

Not a computational defect, and the most expensive. Eight of eight checked claims were prior art.

8. Reading your own results honestly

The hardest skill in research is not generating results. It is noticing which of them you are protecting.

Three tests, applied in order, catch most self-deception.

The swap test

Would you accept this result if a competitor produced it with the same evidence? If not, you are holding a standard you would not accept from others.

The refutation test

Name the observation that would change your mind. If the answer is "none that I can imagine", the claim is not empirical.

The mechanism test

Separate the effect from the explanation. It is entirely possible to be right about the number and wrong about the mechanism — twice in the same project, in this book's case, for the same claim.

The honest posture is not constant doubt, which is paralysing. It is a standing commitment to publish the corrections, which is the subject of the next chapter.

9. The corrections register

Every research programme has a private list of things it got wrong. Publishing that list is the single most credible act available to a researcher, and almost nobody does it.

Sixteen corrections were recorded in the project behind this book: claims withdrawn, narrowed, restated, or found to be model-dependent. The register has a fixed format — the original claim, what the test showed, the outcome — and its value is that it is checkable against the rest of the work.

What the sixteen had in common

The effect was real and the mechanism was wrong. The derived-column detector genuinely fails at low noise; the explanation of why was wrong twice. Reporting the number and inventing the reason is the commonest way to be right and wrong at once.

The test was too clean to fail. A test that cannot fail is not a test.

The defaults were doing the work. Unnormalised columns, unstated thresholds, accidental collinearity — in each case the library, not the science, decided the answer.

> *Before trusting a result, state which convention it depends on. If the convention is unnamed, the result is unnamed too.*

10. Writing it up so it can be checked

A paper is not a story about what you did. It is an instrument for letting someone else decide whether to believe you.

Structure that works

Abstract: the finding, the data, the size of the effect, and the limitation — in that order, in under two hundred words.

Data: where it came from, how many records, what was excluded and why. A reader should be able to reconstruct the dataset exactly.

Methods: every estimator named, every threshold stated, every convention declared. This is where credibility is won or lost.

Results: numbers with units and uncertainties. Tables over prose.

Limitations: a real list, not a ritual one. The limitations section is the part editors read first.

Data and code availability: a repository, not a promise.

The reproducibility minimum

Fixed seeds, versioned dependencies, raw outputs retained, and a script that regenerates every number in the paper. If a reader cannot rerun it, the paper is a press release.

11. Publishing, licensing and being found

Publishing is now the easy half. Being found is the hard half.

Choose the outlet by scope, not by prestige

A short methodological note belongs in a letters journal or a preprint server, not in an article-shaped submission that will be rejected for length. Preprints are free, citable, and no longer career-limiting in most fields.

License for reuse

Code under MIT or Apache-2.0; text under CC BY 4.0 unless a publisher requires otherwise. A permissive licence increases citations, because people can build on what they can legally use.

Make the artefact the advertisement

A working tool with a README is more convincing than any claim in a paper, because it can be run before it is trusted. Where possible, publish the instrument, not just the result.

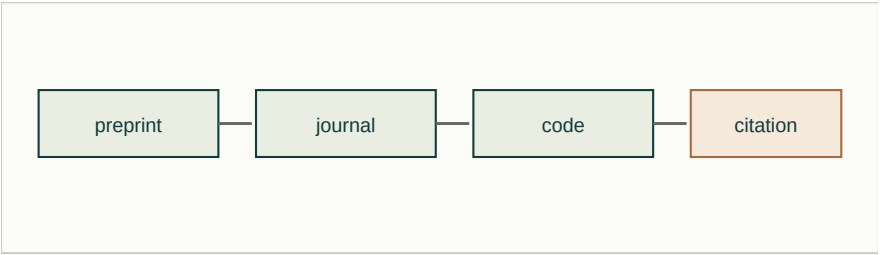


Figure 4. The citation chain. Code is the least glamorous link and the one most often skipped.

12. What to do when the answer is no

Most tests fail, most claims narrow, and most of the value is in what survives.

A failed test is a result if it was designed to be informative. "The effect did not appear under conditions X and Y" is a finding, provided X and Y were pre-specified and the test could have detected the effect.

A failed test is not a result if it was underpowered, if the negative control fired, or if the metric never had a chance to move. Distinguishing these is the difference between a negative result and a wasted week.

> *The discipline is not to be right. It is to arrange your work so that being wrong is informative, cheap, and published.*

13. Working with a machine collaborator

The default use of an AI in research is to ask it for ideas. That is the least valuable thing it can give you, and it is the one it is most fluent at. This chapter is about extracting the useful services instead.

Ask for tests, not ideas

"Give me twenty hypotheses" produces volume. "Given this claim, what is the cheapest observation that would refute it?" produces a design. The second question is harder, slower, and worth more.

Make it argue against you

Ask explicitly for the strongest case that you are wrong: the prior art, the alternative mechanism, the artefact that would produce your result without your effect existing. A collaborator who only agrees is a mirror with a keyboard.

Ask it to build the negative control

Machines are good at constructing the case where the effect must be absent, because it is a mechanical transformation of the positive case. That single habit would have caught two failures in the programme behind this book.

Use two independent implementations

Have the same quantity computed twice, in different languages or by different routes, and compare. The disagreement is the diagnostic. A 28% error in a headline statistic surfaced only this way.

Never let it paraphrase a source it has not been given

Fluent summary is the failure mode with the longest half-life, because it reads exactly like knowledge. If a claim depends on a paper, put the paper in front of it or check the claim yourself.

Request	What you get	Use it for
"Twenty hypotheses"	volume, mostly prior art	exploration, never conclusions
"What would refute this?"	a test design	the actual work
"Where is this an artefact?"	candidate confounds	study design
"Compute it two ways"	a discrepancy	defect finding
"Search for this claim"	prior art, or an honest blank	before anything else

14. Credit, disclosure and the ethics of assistance

AI assistance is now ordinary in research and still unevenly disclosed. Three rules keep you out of trouble and, more importantly, keep the record honest.

Disclose the assistance, not the tool

State what the machine did — drafted, summarised, computed, checked — rather than which product was used. Tools change; the division of labour is the thing a reader needs.

Never attribute a claim to a source you have not read

Generated citations are the most damaging artefact of this era: they are plausible, formatted correctly, and sometimes do not exist. If a reference matters, open it.

You are responsible for the errors

No reader will accept "the model produced it" as an explanation, and no reviewer should. Assistance changes how the work is done, not who answers for it.

The disclosure test

Would you be comfortable if the assistance were printed beneath the title? If not, the problem is not the disclosure — it is the work.

> *Ethics here is not a compliance exercise. It is the same question as everywhere else in this book: can the reader check what you did?*

15. Three worked cases

These came from real computations on public earthquake records, and each is small enough to reproduce.

Case 1 — A boundary that swallowed 28% of a result

A threshold computed as `bin_edge + 0.05` is not 4.4 in binary but 4.400000000000001, so a selection written `m >= mc` excluded every record sitting exactly at 4.4 — 2,060 of them, in the densest bin.

Selection	Records	Value
`m >= 4.400000000000001`	6,351	0.980
`m >= 4.4`	8,411	1.252

Lesson: round the threshold and compare with a tolerance. The densest bin sits at the threshold, so the error is largest exactly where the statistic is most sensitive.

Case 2 — A pooled number that belonged to nothing

Five magnitude scales appear in the same file. Computed separately: duration 0.728, moment 0.995, local 1.047, regional moment 1.562, body wave 1.742 — a spread of **2.39×**. The pooled value, 1.252, is stable to ±2% across three years and belongs to no single population.

Lesson: a mixture can be stable while its components move by a factor of two.

Case 3 — The penalty belongs to the parameters, not the measurements

Three similar sensitivities on five temperatures: as two parameters the condition number is **26.9**; with the third carrying its own free parameter it rises to **2,001.8**. The same third measurement added as a redundant observation improves conditioning instead.

Lesson: count parameters, not data points. The correction is narrower and more useful than the claim it replaced.

What the three share

None produced an error message. All three looked like results. In each case the defect surfaced by computing the same quantity a second way, not by reading the code again.

16. Objections, answered

Six objections arrive reliably whenever this material is discussed. Each has a short answer, and the short answer is usually a version of the same point.

"The models will get better, so this will date."

The specific failures will change; the structure will not. A system that generates fluent text about the world cannot, by generating more, tell you what is already true in it. Verification is not a limitation of the current generation of tools; it is what verification means.

"Peer review will catch it."

Review catches implausibility, not arithmetic. The 28% threshold error described in chapter 15 produced a complete, coherent, publishable table. No reviewer reading the manuscript would see it, because the defect is invisible in the output and obvious only in the selection statement.

"I don't have data."

Public archives exist in most quantitative fields, and reanalysis is real research. The worked cases in this book use two open datasets totalling under eighty thousand records, both free to download. The obstacle is usually not access but the belief that only new data counts.

"My field is different."

That is what the sixty-domain exercise tested. The same failure appeared under different names in seismology, metrology, manufacturing and conservation science, because it is a property of measurement rather than of a discipline. Field differences change the vocabulary, not the counting.

"Aren't you just doing statistics?"

Partly, and that is the point. Most of the expensive errors in the programme behind this book were statistical rather than physical: a threshold comparison, a pooled mixture, a parameter counted as data. The exotic-sounding failures — symmetry breaking, quantum noise — were the benign ones.

"Who is accountable for machine-assisted work?"

The person who signs it. Assistance changes how the work is produced and not who answers for it, which is the subject of chapter 14 and the reason disclosure is a requirement rather than a courtesy.

> *Every answer reduces to one question: can the reader check what you did? Where the answer is yes, the objection disappears; where it is no, the objection is correct.*

17. Techniques: the arithmetic of not fooling yourself

Most measurement error is arithmetic before it is physical. These five techniques take an hour to learn and remove a large share of the errors that reach publication.

Build the uncertainty budget first

List every independent contribution as a standard uncertainty and combine in quadrature. If one term dominates by three to one, improving anything else is wasted effort. The budget tells you where the money goes, and it usually contradicts the plan.

> **Common mistake:** adding contributions linearly for safety. Independent terms combine in quadrature; linear addition overstates the total and hides which term dominates.

State the coverage factor

A 95% interval at five degrees of freedom is about 2.6 standard uncertainties, not 2.0. Quoting the smaller number makes a result look twice as certain as it is.

Detection limits are not detection thresholds

Three standard deviations of the blank is a critical level: the point above which a false positive is unlikely. It is not the point at which a real signal is reliably found. The two differ by roughly a factor of two,

and quoting the first as the second is the commonest metrological error in print.

Round once, at the end

Round for presentation, never for storage. Means of rounded values inherit a systematic bias that interacts badly with thresholds sitting on round numbers.

Keep the raw numbers

Every table should regenerate from a stored original. Lose the raw data and the analysis becomes an anecdote with a decimal point.

> **Technique:** compute the same quantity two ways and print both. The difference is your real error bar, and it is usually larger than the propagated one.

18. Techniques: designing a measurement that can decide

A measurement is a decision procedure. If two hypotheses predict the same reading, the apparatus cannot separate them however precise it is.

Contrast beats repetition

Weight observations so the unknowns you care about are weighted differently. Ten nearly collinear observations condition a fit almost as badly as one; two separated operating points are worth more than ten at the same point.

Count before you fit

Parameters must be fewer than independent measurements, with margin. At equality the system is solvable in principle and unusable in practice: condition numbers in the billions are routine at the boundary.

Vary what should not matter

Change instrument, operator, day, batch. If the answer moves, you have found an unbudgeted term. It is the cheapest confound test there is and it is almost never done systematically.

Design the null first

Build the case where the effect must be absent and run it before the positive case. If the null fires, you have found a defect rather than a discovery.

Design choice	Buys	Costs
Repeats at one point	precision	nothing about bias
Two separated points	conditioning	time; wider model assumption
Swap the instrument	artefact detection	a second calibration
Pre-register the rule	credibility	the option to change your mind

19. Techniques: public datasets, honestly used

Reanalysis is real research and the cheapest entry point available. It has its own failure modes, most of them invisible in the download.

Record provenance at download time

URL, query parameters, date, record count and a hash. A dataset without a retrieval record cannot be defended a year later when the archive has been revised.

Read the field list before the numbers

Which column is preferred? Are several measurement types in one file? Is any column derived from another? Five magnitude scales shared one file in the catalogue used here, and pooling them produced a statistic belonging to none of them.

Find the completeness edge

Every catalogue is incomplete below some threshold. Plot the histogram before modelling; the modal bin and the shape of the low end tell you where the data stop being a sample.

Check the digits

Last-digit distributions reveal rounding, mixed sources and transcription. A large chi-square against uniformity is a warning that the column was assembled rather than measured.

> **Rule:** if you cannot state the query that produced your file, you do not have a dataset — you have a spreadsheet.

20. Building your own error corpus

Every mistake you catch is worth more than the result it destroyed, provided you write it down. A personal corpus turns isolated embarrassment into a checklist that fires early.

The entry format

Four fields: the claim as you believed it, what the test showed, the mechanism of the error, and the check that would have caught it. Only the fourth compounds, because it becomes a procedure.

Review it quarterly

Read the corpus before starting anything new. Errors cluster by type; if four of your last ten were threshold comparisons, that is where the next check belongs.

Publish it

A public corrections register is the cheapest credibility available to an unknown researcher, because almost nobody does it. It also makes the corpus impossible to quietly revise, which is the point.

> *An error you have written down is a technique. An error you have forgotten is a habit.*

21. A year of practice

Skill here is built by repetition on small complete problems, not by reading about method. One full cycle per month, each ending in something published.

Month	Exercise	Teaches
1	Reanalyse one public dataset; publish a one-page note	provenance, completeness
2	Reproduce a published number from its stated method	where descriptions fail
3	Build an uncertainty budget for an instrument you use	where error lives
4	Run the seven defects against your own last analysis	self-audit
5	Search twenty claims from your field for prior art	the base rate
6	Design and run a null that could fail	test design
7	Write a correction to something you published	the register habit
8	Build a small tool others can run	artefact over claim
9	Review a paper formally, in writing	reading suspiciously
10	Reproduce your own result from raw data alone	reproducibility
11	Present one result to a hostile audience	where the argument is weak
12	Consolidate the corpus into a checklist you will use	compounding

Twelve finished pieces beat one ambitious project that never ships. This is chapter 2 applied to a career: what compounds is not volume

but the completion of checkable things.

23. Where this came from, where it is, where it is going

A short history is useful not because history predicts, but because it tells you which parts of the current arrangement are load-bearing and which are accidents of the moment.

How we got here

The first era of machine text was statistical: counts of words, smoothed, ranked. It summarised and it translated, badly but usefully, and it never appeared to think because it was visibly counting.

The second era was neural and still narrow. Models learned representations rather than counts, and a supervised model could label, classify and extract at human-adjacent accuracy on a single task. The characteristic failure was brittleness: excellent inside the training distribution, useless just outside it.

The third era is the current one. Models trained to predict the next token across an enormous range of text acquire, as a side effect, the ability to follow instructions, summarise, translate, write code and explain. Nobody designed a "research assistant"; it fell out of scale plus instruction tuning, and then reinforcement from human feedback made it pleasant to use.

Era	What it could do	Characteristic failure	What researchers used it for

Statistical	count, rank, retrieve	no understanding	search, indexing
Neural, narrow	one task, well	brittleness off-distribution	labelling, extraction
Generative	write, explain, code, summarise	fluent confabulation	drafting, exploring, explaining
Agentic	act, call tools, iterate	silent compounding errors	pipelines, automation

Where it is now

Three things are true simultaneously, and confusing them is the source of most bad judgement about these tools.

> **It is genuinely useful.** For breadth — enumerating failure modes, drafting a protocol, explaining an unfamiliar method, writing the boring 80% of a script — the leverage is real and large. Work that took a week of reading can take an afternoon.

> **It is confidently unreliable.** It will produce a citation that does not exist, a number that is plausible and wrong, and an explanation that fits the data because it was generated from the same assumptions as the claim. None of these announce themselves.

> **It is not a source of truth.** It has read a great deal and verified nothing. The distinction between having read and having checked is the whole of research, and it is the one distinction the tools cannot make for you.

The practical consequence is a division of labour that sounds obvious and is rarely practised: let the machine generate breadth and prose; keep verification entirely for yourself. The failure mode is that generation feels like verification because the output is fluent.

Where it is going

Three trajectories seem safe to describe, and one warning.

Longer chains of action. Models increasingly call tools, run code, read files and iterate. That multiplies capability and multiplies error: an error introduced at step two is propagated confidently through step nine, and the final answer still looks finished.

Better grounding. Retrieval against your own documents reduces invention and increases the temptation to trust without reading. A quotation from your own file is checkable; a quotation that was nearly from your own file is worse than no quotation.

Cheaper and more local. Capability per dollar keeps rising, which moves these tools from a subscription into a habit and into instruments. Expect analysis in the loop of measurement rather than after it.

> **The warning:** as tools improve, the failure modes migrate from the generation to the verification. When the prose is consistently excellent, the only remaining defect is in the numbers — and numbers are precisely what nobody re-reads.

What does not change

The epistemics are fixed. A claim is not true because it is well written, a number is not verified because a program printed it, and a test is not a test unless it could have failed. None of these were ever different, and no amount of capability alters them.

> *The tools change what is cheap. They do not change what counts as evidence. Every durable advantage in this domain comes from doing the expensive part anyway.*

What to do about it, this week

Three habits, ordered by return.

Ask for tests, not ideas. One good discriminating test is worth a gross of hypotheses, and the machine is competent at generating tests when asked directly.

Verify every number twice. Two implementations, two routes, one comparison. The discrepancy is the only error bar that has ever caught a real defect in this work.

Write down what you got wrong. The register is the only asset here that appreciates. Capability will be rented from vendors forever; a record of your own errors is the one thing that cannot be bought.

24. Doing research without a laboratory

The assumption that research requires an institution is mostly false and entirely convenient. What an institution actually supplies is not equipment but access, insurance and a name on a form. Each of those can be obtained other ways, at a price, and knowing the price is the difference between a project that happens and one that does not.

What you are really buying when you rent a lab

Instrument time is the visible cost. The invisible costs are set-up, calibration, sample preparation, operator expertise and repeats after a failure. A quoted hourly rate typically understates the true cost per usable result by a factor of two to five. Ask for the cost of *one finished measurement*, not one hour.

Access route	Typical cost shape	Best for	Main trap
University core facility	hourly, sometimes waived for collaboration	standard characterisation	queue time; you get the operator you get
National laboratory beamline	free at point of use via competitive proposal	unique large instruments	months of lead time; proposal writing
Commercial testing house	fixed price per test	standard, documented results	methods optimised for industry, not for your question
Contract research organisation	project priced	whole studies	you lose the detail that matters

Shared equipment marketplace	hourly, self-operated	quick iterations	training burden falls on you
Borrowed academic collaboration	co-authorship	long projects	credit dilution; their priorities

The three questions before any booking

What can this instrument not do? Ask for the failure modes before the specifications. The operator knows them and will tell you if asked directly; the brochure never will.

What is the calibration chain? An uncalibrated instrument produces numbers, not measurements. Ask when it was last calibrated, against what, and whether the certificate travels with your data.

What does a null look like here? Run a standard or a blank first. Without one, a result you cannot interpret is worse than no result, because it costs the same and looks like progress.

> **Technique:** buy a small pilot before a campaign. One sample, one run, documented, and ask for the raw files. Many facilities will credit a pilot against a larger booking, and the pilot tells you whether the question is answerable at all.

Alternatives that are genuinely credible

Reanalysis. Public archives hold more usable data than most institutions generate. It is real research, it is publishable, and it costs nothing but time.

Simulation against published data. A model validated on someone else's measurements is a contribution; a model validated on its own output is not.

Instrument sharing. A measurement you cannot afford alone is often affordable across three groups with adjacent questions. The coordination cost is real but usually below the equipment cost.

Hiring the measurement, not the lab. If the question is standard, a testing house is cheaper than a relationship.

> **The honest limit:** some questions require an apparatus that does not exist yet, and no amount of cleverness substitutes. The skill is recognising early which kind of question you have.

25. Money, time and the shape of a small research practice

Research as an individual is a small business with an unusual cost structure: almost all of the cost is your time, and almost none of the revenue comes from the research.

Where the money actually comes from

Activity	Pays?	Compounds?
Publishing papers	no (often costs money)	slowly, over years
Preprints and code	no	citation trail
Consulting and audits	yes, within weeks	reputation and referrals
Tools others use	rarely directly	strongly, via credibility
Teaching and workshops	yes	yes

The practical arrangement for most independents is that services fund research, and research makes the services credible. The order matters: the service comes first because it pays now, and the research keeps the service from being a commodity.

Time allocation that survives contact with reality

One day in five on selling, two on delivery, two on the work that compounds. A week with no selling is a month with no income, and this is the failure mode that ends independent research practices, not bad science.

> **Technique:** price the smallest useful unit of work — a one-day audit, a fixed-scope review — and name the price publicly. Unpriced services attract people who will never buy.

Appendix A — A one-week research protocol

Day	Activity	Output
1	State the claim as one sentence; search for it, against it, and for reviews	prior-art note with dates
2	Write the six possible shapes; pick one; define the decisive observation	one-page protocol
3	Build the negative control first; verify the test can fail	control result
4	Run the positive case; compute the quantity two independent ways	result + discrepancy note
5	Check the seven defects against your own pipeline	defect list
6	Write limitations before results; decide what would change your mind	limitations section
7	Publish the artefact, the data, and the corrections you made	preprint + repository

Appendix B — The claim checklist

Apply to every claim before it is believed, including your own:

1. Is it one sentence, with a mechanism and a test?
2. Which of the six shapes is it?
3. What observation would destroy it?
4. Has anyone published it? When?
5. Is the effect separable from the instrument?
6. Does the estimator choice change the answer?
7. Is the statistic pooled over a mixture?
8. Are there more parameters than measurements?
9. Does the negative control behave?
10. Which convention does the result depend on, and is it named?

Appendix C — Fifty-eight tools

The companion library, *research_tools*, provides 58 dependency-free helpers in nine groups — robust statistics, outliers, tails, uncertainty, signal handling, sampling, regression, data hygiene and provenance — plus seven catalogue checks in a separate module. Every function is short enough to read before trusting it, which is the design rule that makes auditing possible.

Appendix D — Glossary

Term	Meaning in this book
Artefact	An apparent effect produced by the instrument or the analysis rather than the sample.
Condition number	Ratio of largest to smallest singular value; how much a fit amplifies error. Above ~10^3 the answer is usually unusable.
Count rule	Identifiability requires at least as many independent measurements as unknowns.
Derived column	A variable computed from another, which cannot serve as independent evidence.
Estimator spread	Disagreement between standard ways of computing the same quantity. Systematic, and it does not shrink with more data.
Negative control	A case in which the effect must be absent; if the test fires, it measures something else.
Pooled statistic	A number computed across a mixture of populations, belonging to none of them.
Prior art	Work already published on a claim. Found by searching the claim, not the topic.
Rank deficiency	Fewer independent measurements than unknowns; the parameters cannot be separated at all.
Reproducibility minimum	Fixed seeds, stated versions, retained raw output, and a script that regenerates every number.
Threshold comparison	Selection at a computed boundary, which can silently exclude values sitting exactly on it.

The evidence behind this book

The book's central empirical claim — that generated hypotheses are largely prior art — rests on **eight checked claims, all eight of which turned out to be published.** That is a narrow base and it is stated here rather than in a footnote: four is enough to justify the warning and not enough to establish a rate. A reader should treat "most generated hypotheses are already known" as a strong prior from this experience, not as a measured proportion.

Everything else in the book is of a different kind: methods, definitions, and results that came from computations on public data and can be rerun. The three worked cases in chapter 15 are the same numbers reported elsewhere in this work, and the scripts that produced them ship with it.

How this book was made

Written in collaboration with an AI assistant, which drafted and revised the text at the author's direction. That fact is stated here because the book argues, at length, that generated material must be checked before it is believed — and a page making that argument has no business concealing its own provenance.

What was generated: the prose, the structure and the figures. What was checked: every number, against a script that can be rerun. Three claims from the first draft were removed or weakened on review, and one figure was redrawn because it implied a precision the underlying computation did not support. The corrections are recorded in the same register the book describes.

Colophon

Discovery with Machines: A Field Guide to Research with AI. First edition, 2026. Written and typeset as a single self-contained HTML document, converted to PDF and EPUB from the same source. Illustrations are vector figures drawn inline so that they remain sharp at any size and can be edited in place.

The claims in this book about how research fails are drawn from a programme of 6,000 generated hypotheses, 58 tools, 16 published corrections and two real datasets totalling 79,095 records. Where a number appears, it came from a computation that can be rerun.

Licence: text CC BY 4.0. Companion code: MIT. Corrections are welcome and will be recorded.