

# AI and Code Clinic

## Corporate Behavior and Regulation Analysis (COBRA)

Luca Caprari



# Plan for today

1. Four AI-written papers on the same dataset (CES, V1 to V4).
2. A recent benchmark: Project APE.
3. Building a cross-country tax-code dataset with an agent.
4. What I take away from this so far.

You have already seen what these tools can do. My job today is the other side of the picture: where they still fall short, and why I think top-level research work cannot, for now, be replaced by an agent.

# Plan for today

1. Four AI-written papers on the same dataset (CES, V1 to V4).
2. A recent benchmark: Project APE.
3. Building a cross-country tax-code dataset with an agent.
4. What I take away from this so far.

You have already seen what these tools can do. My job today is the other side of the picture: where they still fall short, and why I think top-level research work cannot, for now, be replaced by an agent.

For risk-averse researchers: Settings  $\Rightarrow$  Privacy  $\Rightarrow$

Help improve Claude

Allow the use of your chats and coding sessions to train and improve Anthropic AI models. [Learn more.](#)



- ▶ The data is the ECB Consumer Expectations Survey: 11 euro-area countries, April 2020 to March 2026, around 146,750 unique respondents, 72 monthly waves.
- ▶ What makes the CES interesting is the probabilistic elicitation of inflation expectations, which lets you recover an individual subjective distribution.
- ▶ The dataset has already supported top-5 work, for example:
  - ▶ *Tell Me Something I Don't Already Know: Learning in Low- and High-Inflation Settings* (Econometrica);
  - ▶ *The Effect of Macroeconomic Uncertainty on Household Spending* (American Economic Review).

So the model should have strong priors on this data, and wrangling cannot be blamed for a weak result.

- ▶ I never wrote a line of code or LaTeX. I only sent prompts and read the output.

# Four versions, same data, escalating prompts

V.	Prompt change	Result
V1	"Write a paper for top econ/acct journals."	Panel with individual and wave fixed effects. Identification described as "within-respondent variation".
V2	Same chat. "The identification is weak. Re-think."	Moves to Bayesian updating with probabilistic beliefs. The mechanism sounds better; the identification is mechanical.
V3	Same chat. "Still not enough. Read top accounting journals and AER."	Staggered DiD on energy VAT cuts. The paper itself flags pre-trends and timing problems but does not solve them.
V4	New chat. Same literature folder from the start.	The cleanest prose of the four, but it goes back to a financial-literacy correlation. No identification.

Each step buys better writing. None of them buy an identification strategy.

# V1, the first attempt [▶ Open paper](#)

The first prompt was generic: write a paper that could go to a top economics or accounting journal, using this dataset.

The output is a 24-page paper, *Inflation Expectations and Consumer Spending: Panel Evidence from the Euro Area*. One percentage point more expected inflation raises a spending index by 0.019 ( $p < 0.01$ ), with heterogeneity by literacy, income, and housing tenure. The story is intertemporal substitution.

The entire identification section reads:

*“Conditional on individual and wave fixed effects, identification of  $\beta_1$  comes from within-respondent, relative-to-aggregate variation in inflation expectations.”*

No exogenous variation, no instrument, no shock. Expectations and spending intentions are jointly determined.

Same chat, escalation prompt. The model agreed and changed topic: *How Do Households Update Inflation Expectations? Bayesian Beliefs and Financial Literacy.*

- ▶ Nicer try: literacy implies more precise priors and leads to a sign reversal between individual and aggregate forecast errors.

**BUT** the model regresses “how much you changed your forecast” on “how wrong your forecast was,” with both containing the same variable (the old forecast), so a positive coefficient appears even if households don’t actually learn. The headline sign reversal across the two specifications has a much simpler explanation than Bayesian updating: literate respondents just report numbers more accurately, so their measurement noise is smaller.

## V3, “read the top-papers corpus”

► Open paper

Same chat again. This time I gave the agent the full PDFs of:

- the last six years of top accounting journals (TAR, JAR, JAE, CAR, EAR);
- the last five years of AER.

The instruction was to study them and reach that standard of identification.

Output: *Tax Salience and Household Inflation Expectations: Evidence from European Energy VAT Cuts*. A staggered DiD across five countries with France as a never-treated control.

Headline:  $\beta = 0.763$  on perceived inflation ( $t = 13.5$ ), and a much smaller  $\beta = 0.054$  on forward-looking expectations.



# V3, where it falls apart

The paper itself raises the issues, and then goes on:

*“The event study reveals that treated countries exhibit **trending inflation expectations in the pre-treatment period**. This is not surprising: the countries that adopted VAT cuts were precisely those experiencing more severe energy price pressures . . . ”*

- ▶ Parallel trends are visibly violated, then “handled” by adding fixed effects.
- ▶ Treatment timing is endogenous to the energy price shock, which is the very variable on the left-hand side.
- ▶ Goodman-Bacon and de Chaisemartin issues are cited but not addressed.
- ▶ France is not a clean control. It used direct subsidies, so it differs in policy and in exposure.

A careful referee rejects this on the identification page. The model notices the problems and writes them up anyway.

## V4, new chat, same literature [▶ Open paper](#)

The thought was: maybe V1 anchored the rest of the conversation. So I opened a fresh chat and gave it the same literature folder from the start.

Output: *Financial Literacy, Inflation Expectations, and Household Spending*, the cleanest prose of the four.

- ▶ It went back to the same correlation design as V1.
- ▶ Identification is again “within-individual variation in self-reported literacy”, which is mostly measurement noise.
- ▶ Given a free hand, the model produced the safest publishable-looking version, not the most credible one.

A fresh chat removes accumulated context. It does not remove the preference for mechanical fixed-effect designs.

# What I take from the four CES papers

- ▶ The agent is good at producing publication-shaped drafts: structure, tables, citations, robustness, even sections that anticipate referee comments.
- ▶ It is poor at the part that defines good work: noticing that an identification strategy is just OLS with extra steps, proposing a shock or instrument it has not already seen, or killing a project when the identification cannot be saved.
- ▶ The pattern is consistent across versions: the model imitates the surface of good papers without the underlying logic of why a given variation identifies a parameter.

The Autonomous Policy Evaluation Project (Social Catalyst Lab, University of Zurich) is the largest controlled experiment so far on AI-written economics papers.

## Set-up

- ▶ AI-written economics papers from end to end, using Claude Code (Opus 4.6). Idea, data, code, draft.
- ▶ Real data from public APIs only, fully reproducible replication packages, automated integrity checks. pass, a prose pass, three peer-style referees, revision.

## Output by mid-April

- ▶ About 1,000 finished papers, drawn from 3,299 candidate ideas.
- ▶ 43 human benchmark papers from recent AER and AEJ:Policy.
- ▶ More than 18,000 head-to-head matchups in a public tournament.

▶ [Open the APE leaderboard](#)

# How the tournament works

Every paper plays in a head-to-head tournament against other AI papers and against the human benchmarks.

- ▶ Each pair is given to a judge model (Gemini 3.1 Flash Lite), prompted to act as a senior editor at a top journal.
- ▶ The judge is non-Anthropic on purpose.
- ▶ Pairwise comparisons rather than absolute scores.
- ▶ Every pair is judged twice with the order swapped. If the judge just prefers whichever paper is shown first, that bias cancels out between the two runs.

The leaderboard is the result of many such pairwise votes, aggregated into ratings.

Each paper has three TrueSkill numbers (Microsoft's Xbox Live system, adapted here):

- ▶  $\mu$  (mu): the estimate of the paper's quality. Wins push  $\mu$  up, losses push it down; the move is bigger against a stronger opponent.
- ▶  $\sigma$  (sigma): how uncertain the system is about  $\mu$ . Starts high and shrinks as the paper plays more matches.
- ▶ **Cons.** =  $\mu - 3\sigma$ : the conservative rating. This is what actually determines the ranking.

A few other columns on the leaderboard:

- ▶ **Elo** is the chess-style rating. 1500 is average; a 400-point gap implies the higher-rated player wins about 90% of the time. AER papers sit around 1900, the APE average around 1264.
- ▶ **MP** (matches played) is the sample size behind a rating. More MP means a lower  $\sigma$  and a more reliable Cons.
- ▶ **Virtual losses** are integrity penalties applied outside the tournament. Every paper is auto-scanned for fabricated data, hard-coded results, and broken replications. Severe issues are applied as if the paper had lost a series of matches against a median opponent, dragging down Cons. even if real matchups are going well.

# What this evidence tells us

About 1,000 AI papers; average Elo  $\approx 1264$ . Human benchmarks sit at 1717–1817. The best AI papers reach the bottom of the human distribution; a small number compete with the median. The average is half a thousand Elo points behind.

- ▶ Volume is cheap: a small team produced a thousand finished drafts in a few months.
- ▶ Quality is different.
- ▶ The judges are also AI, and position-swap only helps marginally. None of the papers have been through a real peer review.
- ▶ The site itself says these papers should not be used for evidence-based policy.



# The plan: a cross-country tax-code panel

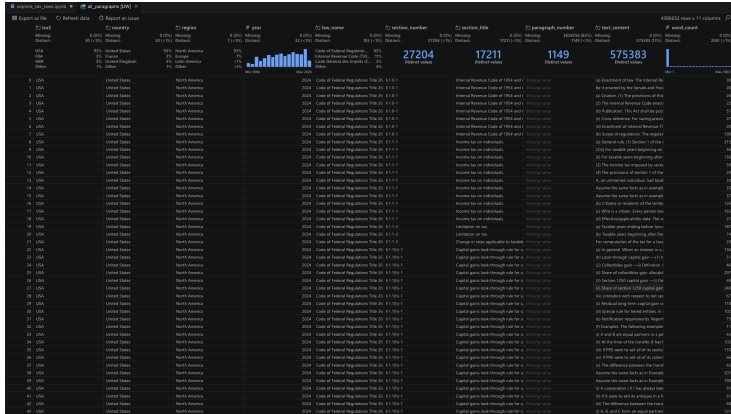
The idea was to build something I couldn't do in a year: a provision-level panel of statutory tax codes for around 30–40 countries, 1996–2026, from primary legal sources.

The first prompt was broad:

*“Collect the full text of personal and corporate income tax laws for 30 to 40 countries. Structure by section / subsection / paragraph. Build a panel-friendly database.”*

The agent went off and did exactly what I asked. It wrote eleven collectors (PwC summaries, the US Code, UK XML feeds, the German full text, JavaScript portals, historical archives), built a 2.8 GB SQLite database, and drafted a paper around an index “ProvComplex” built on top of that database.

# What the data looks like at first glance



country	region	year	law_name	section_number	section_title	paragraph_number	content
USA	North America	2024	Code of Federal Regulations Title 26, § 1.101-1	Internal Revenue Code of 1954 and 1954-1	Internal Revenue Code of 1954 and 1954-1	1	Section 1.101-1 of the Internal Revenue Code of 1954 and 1954-1
USA	North America	2024	Code of Federal Regulations Title 26, § 1.101-1	Internal Revenue Code of 1954 and 1954-1	Internal Revenue Code of 1954 and 1954-1	1	Section 1.101-1 of the Internal Revenue Code of 1954 and 1954-1
USA	North America	2024	Code of Federal Regulations Title 26, § 1.101-1	Internal Revenue Code of 1954 and 1954-1	Internal Revenue Code of 1954 and 1954-1	1	Section 1.101-1 of the Internal Revenue Code of 1954 and 1954-1
USA	North America	2024	Code of Federal Regulations Title 26, § 1.101-1	Internal Revenue Code of 1954 and 1954-1	Internal Revenue Code of 1954 and 1954-1	1	Section 1.101-1 of the Internal Revenue Code of 1954 and 1954-1
USA	North America	2024	Code of Federal Regulations Title 26, § 1.101-1	Internal Revenue Code of 1954 and 1954-1	Internal Revenue Code of 1954 and 1954-1	1	Section 1.101-1 of the Internal Revenue Code of 1954 and 1954-1

27,204 distinct law names, 17,211 “section numbers”, 1,149 “paragraph numbers”, 575k rows. The shape looks reassuring.

# What is wrong with the 30-country output

- ▶ Most non-US laws collapse into a single “section”. Belgium’s Code des impôts sur les revenus 1992 ends up with two paragraphs and six words.
- ▶ Latin American and Asian codes look identical in shape. The agent fell back on the same PwC summary template for every country, producing exactly 253 paragraphs and 913 words per country.
- ▶ “Section number” is whatever the first regex token captured; “paragraph number” is the second. The values depend on the source HTML, not on the legal structure of the law.

The table looks like a panel. It has rows for every country, low missingness, sensible variable names. But the unit of observation is not comparable across countries, so any cross-country regression on this dataset is fitting formatting artefacts.

My second prompt narrowed the scope: forget the 30-country panel, work only with the US Code (Title 26), all annual editions from 1994 to 2024.

The output is much more usable:

- ▶ 30 annual editions, 62,239 section–edition observations, 627 MB of plain-text statute.
- ▶ Reasonable section and subsection parsing for the IRC.

# The two prompts side by side

Broad (30–40 countries)	Narrow (US only)
Heterogeneous primary sources Sections and paragraphs collapsed; units not comparable “Panel” is mostly PwC summary boilerplate Cross-country regressions meaningless	One stable source (US Code, annual editions) Section / subsection structure preserved, but not consistent Real panel level, despite some missings Time-series within the US plausible
Not usable for research	Usable for first prototype, not for real scientific work

Scope was the variable that decided whether the data collection succeeded. Given a wide scope, the agent optimised for “return a row for every country” instead of “make sure the rows are the same object”.

# Patterns I see across the cases

Five recurring problems, in roughly the order they hurt the most:

1. **Surface imitation.** The output borrows the vocabulary of good papers (DiD, event study, IV) without the logic that makes those designs work.
2. **No willingness to stop.** The model will not say “this cannot be identified, drop the project”. It rewrites the section instead.
3. **Scope amplification.** Asked for 30 countries, it returns 30, even when only a few have parseable sources.
4. **Cross-document bleed.** A single context window behaves like a single dataset, and provenance is lost.
5. **Polish above substance.** Tables, plots, robustness sections, references, everything looks finished. The weakest link is usually the underlying claim.

# Why these failures are not just bugs

**A recent paper (Sikka & Sikka, 2025) gives a name to what we saw:**

*“If a task requires more computational steps than an LLM can perform in one response, the LLM will unavoidably hallucinate.”*

**LLMs have prompt “thinking budget”:**  $\sim (\text{input tokens})^2 \times (\text{model size})$ .

It cannot work longer on a harder problem. Tasks that need exponential, cubic, or factorial steps (e.g. truly verifying a claim) exceed that budget  $\Rightarrow$

**Hallucinations are mathematically inevitable.**

**What this means for my cases:**

- ▶ **Surface imitation:** The model mimics a DiD paper’s vocabulary but cannot run the actual verification steps needed to check parallel trends.
- ▶ **No willingness to stop:** Saying “this cannot be identified” would require a meta-analysis beyond the model’s step budget.
- ▶ **Scope amplification:** Returning 30 rows for 30 countries looks correct, but checking cross-country comparability is a higher-order task.

These are not “fixable” bug, they are structural limits of transformer-based models, for now.

# Where I think we are

- ▶ In its current form, agentic AI is an extremely fast, great and tireless collaborator. It is strong on local tasks and weak on the judgment calls that turn data into a paper.
- ▶ My four CES drafts and the data-collection experiment show the same pattern on a smaller scale.
- ▶ Project APE makes the same point: a thousand papers, a handful close to top-5 quality, an average half a thousand Elo points behind.
- ▶ Most of this will improve. The structural parts (surface imitation, scope amplification, cross-document bleed) will not go away just because the next release is more capable.

For now, the agent does not replace the researcher. It changes what one researcher can do, provided the researcher stays the one doing the thinking.



**Thank you.**

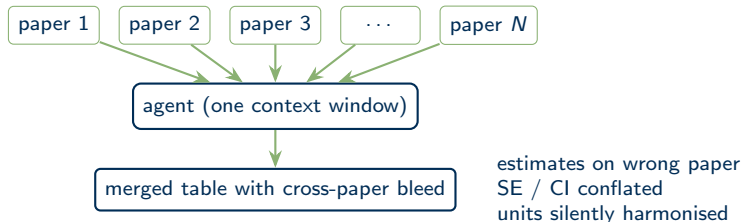
# Meta-study extraction

The task is the obvious use case for agentic AI in applied work: extract point estimates, standard errors, sample sizes, and the identification strategy from a folder of empirical papers with different layouts.

If this works, it saves a year of RA time. So I tried it.

In my experience, the quality depends on whether the agent is allowed to see one paper at a time or many at once. One paper at a time, with a clear schema, works decently, after a clear project description. A folder of papers does not.

# Why one folder is worse than many sessions



A single context window behaves like a single working memory. The agent wants to compare and combine. For a meta-study, you want the opposite: each paper an independent extraction with explicit provenance.

# What goes wrong with a folder of papers

- ▶ The agent reads all PDFs at once, builds an internal “summary” across them, and then back-fills the per-paper fields from that summary.
- ▶ Concrete problems I saw:
  - ▶ estimates from paper A attributed to paper B;
  - ▶ standard errors confused with confidence-interval widths when conventions differ across papers;
  - ▶ “main specification” picked by table position, not by what the paper itself calls main;
  - ▶ units silently harmonised (percentage points versus log points) because the agent “noticed” they were close.
- ▶ Forced to process papers one at a time, the agent still loses the column schema between papers unless I pin it in the prompt.

# What I ended up doing

- ▶ One subprocess per paper, fresh context, no access to its siblings.
- ▶ A hard schema pinned in every prompt: `paper_id`, `table_id`, `row_id`, `coef`, `se`, `n`, `units`, `spec_label`, `source_page`.
- ▶ Every cell has to carry a page and table reference from the source PDF. An empty cell is preferred over a guessed one.
- ▶ A second pass re-reads the source PDF and rejects rows whose cited location does not contain the claimed number.

This is much slower and consumes many more tokens. The shortcut of dropping the whole folder into one context is the thing that destroys the dataset.