

COBRA - AI and Code Clinic for Research and Teaching

Pitfalls when Writing Papers with Claude Code - Luca Caprari, 26 May 2026

My presentation outlines the limitations of agentic AI and explains why, in my opinion, top-level research work cannot currently be replaced by an agent. I build the argument from two experiments of my own, plus a third in the appendix, and one external benchmark.

I first asked an agent to write papers, end to end, on the ECB Consumer Expectations Survey. It is a dataset that I know well, and it has already supported two top-5 papers (*"Tell Me Something I Don't Already Know: Learning in Low- and High-Inflation Settings,"* Econometrica; *"The Effect of Macroeconomic Uncertainty on Household Spending,"* AER), so the model should know the context and the dataset already from its training. I wrote no code and no LaTeX, only prompts. The first version ("V1") came from a generic prompt asking the model to produce a paper aimed at the top economics or accounting journals. The result was a 24-page panel-FE paper whose identification section was one sentence about within-respondent variation. When I pushed back criticizing the identification strategy, the model pivoted obediently to a Bayesian-updating story ("V2") whose update regression turned out to have the prior forecast on both sides; the sign reversal it announced was measurement noise dressed up as learning. After I fed in six years of top accounting journals and five years of the AER, in the third article "V3", it produced a staggered DiD on European energy VAT cuts that flagged, with apparent self-awareness, its own violated parallel trends and endogenous timing, and then proceeded to its conclusions anyway. A fresh chat with the same literature gave me the cleanest prose of the four, called "V4", and slipped back to the correlation design of the first. Each prompt was generating better writing, but the identification strategies were not satisfactory in any of the four.

The Autonomous Policy Evaluation (APE) project at the University of Zurich's Social Catalyst Lab is useful for cross-checking findings on a large scale. They ran Claude Code through the full research pipeline and produced approximately 1,000 finished papers. As a benchmark, they selected 43 recent AER and AEJ: Policy papers. The evaluation is a public head-to-head tournament: every paper is paired against AI and human peers and handed to a non-Anthropic judge (Gemini 3.1 Flash Lite), prompted to act as a senior editor at a top journal. More than 18,000 pairwise comparisons the picture is consistent with what I saw with the CES: the best AI papers reach the bottom of the human distribution and a handful compete with the median, but the average sits about 500 Elo points below, a gap that, in chess terms, implies the human side wins almost every time. It's now cheap to produce a high volume of papers, but quality, so far, has not improved.

The same lesson surfaced when I switched from producing finished papers to data collection. A broad prompt ("build me a provision-level tax-code panel for thirty to forty countries from 1996 to 2026") produced a 2.8 GB large database that, at first glance, looked exactly like a usable dataset. On closer reading it was largely PwC summary boilerplate with units not comparable across countries, and any cross-country regression on it would have been fitting formatting artefacts. Narrowing the scope to the U.S. tax code yielded a more usable result. What decided success was not the model or the phrasing of the prompt, but the scope I gave it.

Five tendencies recur across these episodes: surface imitation, refusal to stop, scope amplification, cross-document bleed, and polish over substance. I do not think these are bugs in the colloquial sense. Sikka and Sikka (2025) give the cleanest account of why: a transformer's per-response $O(N^2 \cdot d)$ compute budget bounds what it can verify, and any task that exceeds that budget (checking parallel trends, judging cross-country comparability, deciding that a project should be abandoned) produces hallucination as a matter of complexity. The limit looks structural, and structural limits do not usually fall to a bigger model. For now, the agent is a fast, tireless collaborator that amplifies what one researcher can do; it does not replace the researcher, provided the researcher stays the one doing the thinking and the managing.

Appendix: A follow-up question could be whether an agent can read a folder of empirical papers and pull out their information for a meta-analysis. The answer depends sharply on how it is framed. The workflow that actually produces a usable meta-dataset is the opposite of a token-convenient one: one subprocess per paper, fresh context, a hard schema pinned in every prompt, page and table references on every cell, and a second pass that re-reads the source PDF and rejects any row whose cited location does not contain the claimed number. The shortcut of dropping the whole folder into one context is the thing that destroys the dataset.

Documents in this folder

- **COBRA_AI_Workshop.pdf**: the slide deck (24 main slides + 4 backup slides on meta-study extraction).
- **CES_V1.pdf**: Fully AI-generated article #1. “Inflation Expectations and Consumer Spending: Panel Evidence from the Euro Area.” Identification is within-respondent variation only.
- **CES_V2.pdf**: Fully AI-generated article #2. “How Do Households Update Inflation Expectations? Bayesian Beliefs and Financial Literacy.” Mechanical update regression; sign reversal driven by measurement noise.
- **CES_V3.pdf**: Fully AI-generated article #3. “Tax Salience and Household Inflation Expectations: Evidence from European Energy VAT Cuts.” Staggered DiD; the paper itself flags violated parallel trends, endogenous timing, and an unclear control.
- **CES_V4.pdf**: Fully AI-generated article #4. “Financial Literacy, Inflation Expectations, and Household Spending.” Cleanest prose; reverts to a V1-style correlation design.
- **Sikka2025.pdf**: Sikka & Sikka (2025), Hallucination Stations. It makes the complexity-theoretic argument: self-attention is $O(N^2 \cdot d)$, so any task above that budget cannot be performed or verified by an LLM. It frames the failures I describe in the talk as structural rather than fixable bugs.