Measuring Estimation Uncertainty due to Researcher Degrees of Freedom with Agentic Artificial Intelligence*

Brett A. McCully

Collegio Carlo Alberto

First draft: April 10, 2026

This draft: June 2, 2026

Word count (main text): 2,141

Abstract

Researchers' sample and specification choices generate uncertainty not captured by standard errors. Measuring this uncertainty is costly, requiring recruiting and coordinating many independent research teams. I develop a scalable alternative using agentic AI: repeated AI agents receive the same research prompt and raw data, independently design empirical specifications, and estimate the target parameter. I benchmark this approach against a recent human multi-analyst study. AI agents' coefficient, standard-error, and sample-size distributions are broadly comparable to those of human analysts, but AI achieves this via systematically different specification choices. Agentic AI can help measure researcher-induced uncertainty, but remains an imperfect human substitute.

Keywords: Researcher degrees of freedom, multi-analyst studies, specification uncertainty, AI agents

JEL Codes: C12, C15, C18

*brett.mccully@carloalberto.org. I thank Yagan Hazard for helpful conversations. Any errors are my own.

1 Introduction

Empirical papers routinely report standard errors that quantify sampling uncertainty. These standard errors are conditional on a set of researcher decisions: which dataset to analyze, how to define the estimation sample, how to construct outcomes and treatments, and which specification and standard-error adjustments to use. A growing body of evidence shows that researchers answering the same question and using the same raw data make different decisions that can produce meaningfully different estimates (Silberzahn *et al.*, 2018; Huntington-Klein *et al.*, 2021; Breznau *et al.*, 2022). This implies an additional source of uncertainty beyond standard errors capturing sampling variation: researcher degrees of freedom.

Measuring parameter uncertainty due to researcher degrees of freedom at scale is challenging. Multi-analyst studies are expensive, requiring extensive recruitment and coordination. Specification-curve and multiverse analyses require the original analyst to enumerate and implement various alternatives (Simonsohn *et al.*, 2020; Steegen *et al.*, 2016).

I experiment with a novel approach: AI agents that perform empirical analyses at scale.[1] Each agent receives the same research prompt and raw dataset but independently generates a complete empirical specification—sample restrictions, outcome and treatment definitions, control set, and standard-error adjustments. The inherent stochasticity of the large language models which power modern AI agents ensures a range of varied research choices.

Do agent sample and specification choices yield a similar distribution of estimation outcomes? To answer this, I apply my AI agent approach to the question studied by many analysts in Huntington-Klein *et al.* (2025): how eligibility for the Deferred Action for Childhood Arrivals program affects the probability of full-time employment for eligible workers. Across 156 AI agent-generated specifications, I obtain an interquartile range from 0.014 to 0.099, resembling the human many-analyst coefficient distribution. Standard error and sample size distributions also resemble each other between AI and humans. The specification choices made—model

[1]There are at least 3 other concurrent, unpublished papers that attempt to do broadly what I propose in this article: Gao and Xiao (2026), Grundl (2026), and Huang *et al.* (2026). Given the timing of the release of these working papers and the work on my own public GitHub repository for this article dating back to January 2026 (see https://github.com/brettmcc/LLM-bootstrapping), this is a clear case of parallel invention. In any case, the only other of these papers to also look at Huntington-Klein *et al.* (2025), Grundl (2026), does not focus on agent specification choices.

type, control variables, weighting, and standard error adjustment—differ substantially, with AI converging on certain choices much more than human researchers. Overall my findings suggest that AI, while useful, is not a perfect substitute for multi-human-analyst studies.

This paper makes two main contributions. First, I introduce a scalable approach to multi-analyst designs that lets researchers quantify researcher-induced uncertainty. Second, I empirically test the AI agents against a human benchmark.

This paper proceeds as follows. Section 2 describes the AI agent pipeline and the DACA application; Section 3 reports the estimate distribution, specification heterogeneity, and the comparison to the human analyst benchmark.

2 Method

2.1 AI agents as empirical researchers

For a given empirical research question, social scientists have substantial degrees of freedom in shaping the analysis. Such choices include selecting which observations make it into their analysis sample, the estimation method used, which control variables to include, how to cluster standard errors, and so on.

For example, in the empirical application below on the effect of an immigration policy, Deferred Action for Childhood Arrivals (DACA), on the likelihood of full-time work among affected immigrants, researchers may make different defensible choices about age bounds of the sample, full-time work definitions, treatment eligibility proxies, and whether to estimate a linear or nonlinear model.

Estimate variation due to researcher degrees of freedom is not captured in standard measures of parameter estimate uncertainty, such as standard errors. "Non-standard errors' ' (Menkveld and many others, 2024) computed across many independent human analyst attempts to answer a research question do so, but are not easily scalable due to high coordination costs. Stochastic artificial intelligence agents may provide a way around such scalability challenges.

Large language AI models, such as GPT or Claude, by construction give non-deterministic

responses: the same prompt can generate a range of responses ex-ante. By submitting a prompt to estimate a given empirical parameter many times to independent AI agents, I propose to simulate at much lower cost a many-analyst approach for generating bounds on the degree of estimation uncertainty due to researcher degrees of freedom.

Why might AI agents produce plausible research choices? Their underlying models are trained on huge corpa of digital data, including academic articles, codebases, and webpages (OpenAI, 2025; Anthropic, 2026). These corpora likely includes many published and working papers in economics, their accompanying replication code, as well as handbook chapters and online textbooks. Much empirical work from outside economics, such as other social sciences and especially data science, also is present in the huge corpora (alongside nearly everything else every digitized). These other empirical fields have their own standards and norms. Still, the language in the prompt given to the AI agent (e.g., words such as "causal" and a focus on labor market outcomes) may nudge the AI towards producing output more in line with the conventions of applied economists.[2] AI agent outputs can therefore be interpreted as draws from a noisy representation of standard empirical social science practice.

On the other hand, AI agents may make research choices far outside disciplinary norms and make a set of choices which are less internally coherent. Yet due to the enormous complexity of frontier large language AI models (each having trillions of parameters), the secrecy of the labs regarding the weights of each parameter, and the randomness of the output, one cannot deduct how models will respond simply from theory. The degree to which AI agents hew to human patterns of research choices is therefore an empirical question. This paper replicates an existing human multianalyst study by Huntington-Klein *et al.* (2025) and compares the distribution of research choices and the resulting estimates.

[2]Large language models (LLM) work by predicting the next token in their output using a logit function. This logit function can have billions or even trillions of parameters. Not all of these parameters are 'activated' each run. Which parameters are activated depends on the context fed into the model; in my empirical application, the prompt and files provided to the AI agent. Parameters associated to econometrics (e.g., a Stata command to estimate a difference-in-differences model) may be given a higher weight when the prompt includes such words as "causal" and "identification", while data science parameters may be more heavily weighted when the prompt includes such words as "prediction".

2.2 Application

I task AI agents with the same question answered by many human analysts in Huntington-Klein *et al.* (2025). The authors retained 146 teams of economists and ask each to independently answer the same causal question using the same instructions and raw data: how did eligibility for the Defferred Action for Childhood Arrivals immigration policy affect the probability of full-time employment among eligible immigrants? Participants are given a data codebook and the research question but otherwise initially exercise broad discretion over sample construction, variable definitions, and specification choices. In this study I replicate this initial task from Huntington-Klein *et al.* (2025) in which participants had maximum flexibility to make research choices.

The computing pipeline works as follows. An AI agent is provided with the raw materials to answer the research question: an IPUMS extract and codebook with all variables used by human analysts in Huntington-Klein *et al.* (2025) (derived from their public Open Science Framework repository), a supplementary state policy dataset and codebook, and instructions adapted from those given to human analysts.[3] The agent then writes a concise, structured file outlining their model specification, outcome definition, sample selection, and treatment definition.[4] The agent then writes and executes a Python script to implement their estimation, working through any coding errors along the way, and reports the resulting sample size and preferred coefficient.

I use two frontier AI models—GPT-5.4 mini (OpenAI) and Claude Sonnet 4.6 (Anthropic)[5]—accessed through GitHub Copilot CLI. I primarily rely on GPT-5.4 mini, as it is cheaper to run than Claude Sonnet 4.6. Each run has 0 dollar marginal cost when using a

[3] All material passed to AI agents are available at https://github.com/brettmcc/LLM-bootstrapping/tree/master/NHK-replications/replication-materials.

[4] One example file from my data lists a specification with `full_time` on `treated_post`, treatment and post indicators, age terms, sex indicators, state fixed effects, and year fixed effects; defines full-time work as usual hours worked greater than or equal to 35; restricts the sample by Hispanic origin, Mexican birthplace, non-citizenship, pre-2008 immigration, childhood arrival, birth cohort, group-quarters status, year, and working age; and defines treatment using a birth-year eligibility proxy.

[5] I exclude 40 exploratory runs with Claude's smaller and less advanced Haiku 4.5 model from the analysis sample. That cohort yielded one run with standard error of 0 and two more with extremely high coefficient estimates. Researchers applying my approach should therefore take heed that smaller models with insufficient reasoning may not be suitable for this type of task.

subscription such as GitHub Copilot, Codex, or Claude Code, making large-scale replication affordable for individual researchers.[6]

3 Results

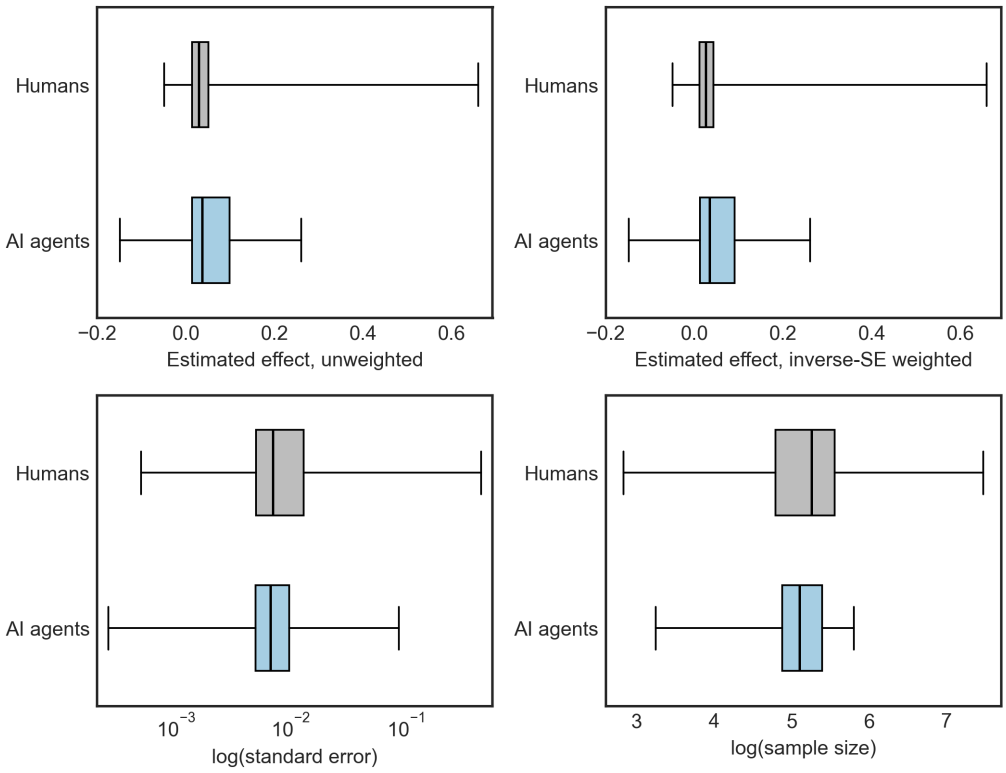
I compare 145 human analyses from Huntington-Klein *et al.* (2025) to 156 AI agent analyses for the same researh question and using the same raw data. Both humans and agents had complete freedom to select their estimation sample, define dependent and independent variables, and choose a specification.

3.1 Distribution of estimates

Figure 1 depicts box-and-whisker plots for both human and AI analyses for weighted and unweighted effects, standard errors, and sample size. The left-most whisker depicts the minimum value, the box depicts the 25th percentile value, median, and 75th percentile, with the right-most whisker showing the maximum reported value. Exact values are reported in Table 1.

[6] Of course, the cost is that each run counts against one's 5-hour and weekly usage limits.

Figure 1: Human and AI estimation outcomes



Notes: Panels compare unweighted point-estimate distributions (top left), point-estimate distributions weighted by inverse standard error (top right), standard errors (bottom left), and log sample sizes (bottom right) for both the human researcher sample of Task 1 from Huntington-Klein *et al.* (2025) and the AI agent sample from this paper. Boxes show the interquartile range, center lines show medians, and whiskers show minimum and maximum values.

Across all four outcomes, means and medians are fairly similar between human and AI researchers. Moreover, both humans and AI agents exhibit substantial dispersion in their estimation outcomes, underling the stochasticity of AI outputs.

For the unweighted effect size, AI obtains a mean of 0.064 and a median of 0.037 and humans a mean of 0.053 and median of 0.030. The 25th percentile estimates are identical at 0.014, while the 75th percentile effect is substantially larger for AI agents at 0.099 relative to humans' 0.051. AI agents also obtained a much lower minimum estimate of -0.148 compared to humans' -0.049, while the AI's maximum estimate is also much lower than humans'. For standard errors, the minimum, 25th percentile, and median are identical across humans and AI, while humans exhibit higher standard errors at the 75th percentile and maximum. As a result, AI agents obtain coefficients statistically significantly different from 0 about 90% of the time, compared to 78% time among humans.

Table 1: Summary statistics on estimation outcomes

	N	Mean	SD	Min	Pctl. 25	Median	Pctl. 75	Max
Panel A: AI-agent estimates								
Effect Size (unweighted)	156	0.064	0.084	-0.148	0.014	0.037	0.099	0.261
Effect Size (weighted by inverse SE)	156	0.060	0.080	-0.148	0.013	0.035	0.091	0.261
Standard Error	156	0.008	0.007	0.000	0.005	0.007	0.010	0.088
Sample Size	156	165,965	123,732	1,764	75,319	125,655	248,270	636,722
Treated-Group Size	152	65,005	28,807	7,717	36,075	73,180	81,508	169,320
Panel B: Human estimates								
Effect Size (unweighted)	145	0.053	0.095	-0.049	0.014	0.030	0.051	0.660
Effect Size (weighted by inverse SE)	138	0.044	0.092	-0.049	0.012	0.026	0.043	0.660
Standard Error	139	0.019	0.055	0.000	0.005	0.007	0.013	0.460
Sample Size	145	828,318	3,056,037	681	61,600	179,960	356,787	29,536,580
Treated-Group Size	141	96,395	648,493	270	17,950	34,435	52,581	7,727,201

Notes: Panel A reports outcomes from the AI agent sample. Panel B restates the published Task 1 panel of Table 3 of Huntington-Klein *et al.* (2025). Treated-group sample size not available for all AI agent runs since its collection was added after the initial runs.

The range of sample sizes chosen by human researchers—between 681 and over 29 million observations—is several orders of magnitudes larger than what the AI agents chose. Still, the interquartile range is comparable: 75 to 248 thousand for the AI, and 61 to 357 thousand for the humans. A potential explanation for the divergence in sample size at the upper end of the distribution is better compliance among AI agents with the reseach question. Given the unbounded nature of the problem, Huntington-Klein *et al.* (2025) note that "some researchers used nearly the entire ACS sample, including people very unlike the DACA-eligible [treatment] group." (p. 24) It is plausible that AI agents better adhered to the research instructions about constructing a control group. For both humans and AI agents, the research questions starts with the proviso, "Among ethnically Hispanic-Mexican Mexican-born people living in the United States…"

3.2 Specification heterogeneity

Did AI and human researchers arrive at their estimates through similar specification choices? We answer this question by examining the prevalence of model specification choices between AI and human runs, summarized in Table 2.

Table 2: Estimation choices by researcher type

		AI		Humans	
Category	Choice	N	Share (%)	N	Share (%)
Method	Linear Regression	156	100.0	358	81.9
	Logit/Probit	0	0.0	57	13.0
	Matching	0	0.0	11	2.5
	New DID Estimator	0	0.0	7	1.6
	Other	0	0.0	4	0.9
S.E. Adjustment	Cluster (State)	127	81.4	118	27.0
	Cluster (State & Year)	1	0.6	58	13.3
	Cluster (ID/Strata/Other)	0	0.0	65	14.9
	Het-Robust	26	16.7	76	17.4
	Other/Bootstrap	0	0.0	23	5.3
	None	2	1.3	98	22.4
Weights	No Sample Weights	7	4.5	329	75.3
	Sample Weights	149	95.5	109	24.9

Notes: Estimation choices are inferred from each generated model specification and execution metadata. Data from Huntington-Klein *et al.* (2025) includes more restricted human runs (Tasks 2 and 3) in which the research design was more tightly specified and precleaned data was provided.

I find that AI agents make notably different model specification choices relative to human economists. AI agents overwhelmingly favor linear models, and never chose nonlinear models, in contrast to 13% of human runs using nonlinear models. AI agents also overwhelmingly converged on clustering standard errors at the state level, with 81% of agents choosing to do so compared to 27% of human runs. The share of runs featuring heteroskedasticity robust standard errors was quite similar between AI and humans.

AI agents almost exclusively used sample weights, while a majority of human researchers did not. Using weights is in line with IPUMS instructions, a fact also highlighted by Huntington-Klein *et al.* (2025).

Table 3: Control variable choices by researcher type

		AI		Humans	
Category	Control	N	Share (%)	N	Share (%)
AGE	Linear Age	77	49.4	164	37.5
AGE	Age FE	19	12.2	36	8.2
AGE	Age Quadratic	64	41.0	33	7.6
EDUC	Linear Education	3	1.9	122	27.9
EDUC	Education FE	1	0.6	32	7.3
EDUC	Education Transform	0	0.0	61	14.0
STATE/YEAR	Linear Year	16	10.3	79	18.1
STATE/YEAR	Year FE	135	86.5	103	23.6
STATE/YEAR	State FE	138	88.5	155	35.5
STATE/YEAR	State FE x Year FE	126	80.8	56	12.8
STATE/YEAR	State FE x Linear Year	15	9.6	23	5.3

Notes: The table presents the number and share of AI and human estimation specificaitons which included various controls. Data from Huntington-Klein *et al.* (2025) includes more restricted human runs (Tasks 2 and 3) in which the research design was more tightly specified and precleaned data was provided.

Choice of control variables differed substantially between humans and AI, as shown in Table 3. AI agents included an age control more frequently than humans did, while AI almost never controlled for education while humans did in over a third of specifications. Nearly 90% of AI agents included year fixed effects, compared to just 24% of human runs. Similarly, nearly 90% of AIs included state fixed effects compared to 36% of human runs. Using a fully-saturated

difference-in-difference diminishes the downsides of linear probability models, the choice of every AI agent. Overall, AI agents agreed on the set of controls much more often than human researchers. Just 25% of AI agent runs used a unique set of controls, compared to 64% of humans.

The resemblance of AI agent estimation outcome distributions to the human distributions is striking. There was, however, sharp disagreement in specification choices between AI and humans. In these disagreements, AI agents often better hewed to the field's empirical standard—such as using two-way fixed effects in a linear probability model and using weights with the ACS data. This suggests that AI is a substitute, albeit an imperfect one, for humans in multi-analyst designs aimed at quantifying the degree of uncertainty due to researcher degrees-of-freedom.

4 Conclusion

This paper develops a novel, scalable approach to multi-analyst studies leveraging new breakthroughs in agentic artificial intelligence. Across runs, frontier AI agents generate a dispersion of typically reasonable research choices. These choices differ systematically from those of human researchers. The multi-AI agent approach introduced in this paper may therefore not be a perfect substitute for multi-human analyst studies. Given the high coordination and time cost difference, however, many-AI agent runs may still be useful to quickly and cheaply get a handle on estimation uncertainty due to researcher degrees-of-freedom—so-called "non-standard errors."

Two limitations are worth emphasizing. First, the Huntington-Klein *et al.* (2025) benchmark materials—including codebooks, instructions, and human submissions—were public on the Open Science Framework website well before my AI agent runs. It is therefore plausible that the AI models I used were trained in part on the original human researcher choices, causing the prompt fed to the AI to surface the code produced by the human researchers, making agent choices too close to the human ones. Future work using a novel research question simultaneously delivered to human researchers and AI agents would help clarify the

importance of this issue. Second, the choice of AI model (e.g., GPT-5.4 mini) and harness (e.g., GitHub Copilot CLI) may be important—other models and harnesses may resemble humans better, or worse. For example, while I dropped runs using the less-sophisticated Claude Haiku model due to several degenerate runs, several of those Haiku runs used nonlinear models compared to 0 of the retained AI runs. Further research should explore how variations in model choice and harness matter for AI agent decisions.[7]

References

Anthropic (2026). Anthropic system card. Anthropic-hosted PDF, last modified March 6, 2026.

Breznau, N., Rinke, M. E., Wuttke, A. *et al.* (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, **119** (44), e2203150119.

Gao, R. and Xiao, S. C. (2026). Nonstandard errors in AI agents. SSRN working paper, date written: March 16, 2026.

Grundl, S. (2026). A comparison of agentic AI systems and human economists. SSRN working paper, date written: April 9, 2026.

Huang, W., Menkveld, A. J. and Yu, S. (2026). AI "errors". SSRN working paper, date written: March 13, 2026.

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T. *et al.* (2021). Influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, **59** (3), 944–960.

—, Portner, C. C., McCarthy, I. and The Many Economists Collaborative on Researcher Variation (2025). *The Sources of Researcher Variation in Economics*. I4R Discussion Paper 209, Institute for Replication (I4R).

[7] I would have done more of this, but doing so becomes quite costly to obtain a sufficient number of runs per model-harness pair.

Menkveld, A. J. *et al.* (2024). Nonstandard errors. *Journal of Finance*, **79** (3).

OpenAI (2025). Gpt-5 system card. OpenAI-hosted PDF, last modified August 19, 2025.

Silberzahn, R., Uhlmann, E. L., Martin, D. P. *et al.* (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, **1** (3), 337–356.

Simonsohn, U., Simmons, J. P. and Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, **4**, 1208–1214.

Steegen, S., Tuerlinckx, F., Gelman, A. and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, **11** (5), 702–712.