

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 9, 2026

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 whereby many AI agents receive the same research prompt and raw data, and separately design and estimate empirical specifications. 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 sample and specification choices. Agentic AI can help measure researcher-induced uncertainty, but remains an imperfect substitute for humans.

Keywords: Non-standard errors, multi-analyst studies, AI agents

JEL Codes: C6, C12, C15

*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: how to define the estimation sample, how to construct outcomes and treatments, which specification to use, and how to compute standard errors. 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 (Stevenson and Fischman, 2026). Measuring parameter uncertainty due to researcher degrees of freedom at scale is challenging. Multi-analyst studies are expensive, requiring extensive recruitment and coordination.

I assess 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 there is variation in research choices.

Do AI sample and specification choices yield a distribution of estimation outcomes similar to humans? 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 interquartile ranges in estimates, standard errors, and sample sizes that resemble the human many-analyst coefficient distribution. Specification choices—model 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

¹There 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 choices as I do.

substitute for multi-human-analyst studies.

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. “Non-standard errors” (Menkveld *et al.*, 2024) computed across many independent human analyst attempts to answer a research question can capture estimate variation due to researcher degrees of freedom, but are not easily scalable due to high recruitment and coordination costs. Stochastic artificial intelligence agents may provide a way around such scalability challenges.

Large language AI models, such as GPT or Claude, 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 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 reasonable research choices? Their underlying models are trained on huge corpora of digital data (OpenAI, 2025; Anthropic, 2026), likely including much economics research and associated code. The language in the prompt given to the AI agent (e.g., words such as “causal” and a focus on labor market outcomes) should nudge the AI towards producing output more in line with the conventions of applied economists, rather than, say, data science practice in industry. AI outputs may 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 deduce 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.

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 asked each to independently answer the same causal question using the same instructions and raw data: how did eligibility for the Deferred 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 a broad set of variables, a supplementary state policy dataset and codebook, and instructions adapted from [Huntington-Klein *et al.* \(2025\)](#).² The agent then writes a concise, structured file outlining their model specification, outcome definition, sample selection, and treatment definition. They proceed to write and execute a Python script to implement the estimation, working through any coding errors along the way, and reporting the resulting coefficient estimate, standard error, and sample size. Sampling and specification choices are then programmatically scraped from agents' reports and cross-referenced against the actual code they produced to ensure accuracy.

I use two frontier AI models—GPT-5.4 mini (OpenAI) and Claude Sonnet 4.6 (Anthropic)³—accessed through GitHub Copilot CLI. I primarily rely on GPT-5.4 mini, as it is cheaper to run than Claude Sonnet 4.6. When using a subscription such as GitHub Copilot, Codex, or Claude Code, the marginal cost of each run is modest, making large-scale

²Replication code and the AI agent results can be found at <https://github.com/brettmcc/LLM-bootstrapping/>.

³I exclude 40 exploratory runs with Claude's smaller and less advanced Haiku 4.5 model, which yielded several degenerate results, such as standard error of 0. Researchers applying this approach should therefore take heed that weaker models may not be suitable for this type of task.

replication affordable for individual researchers.

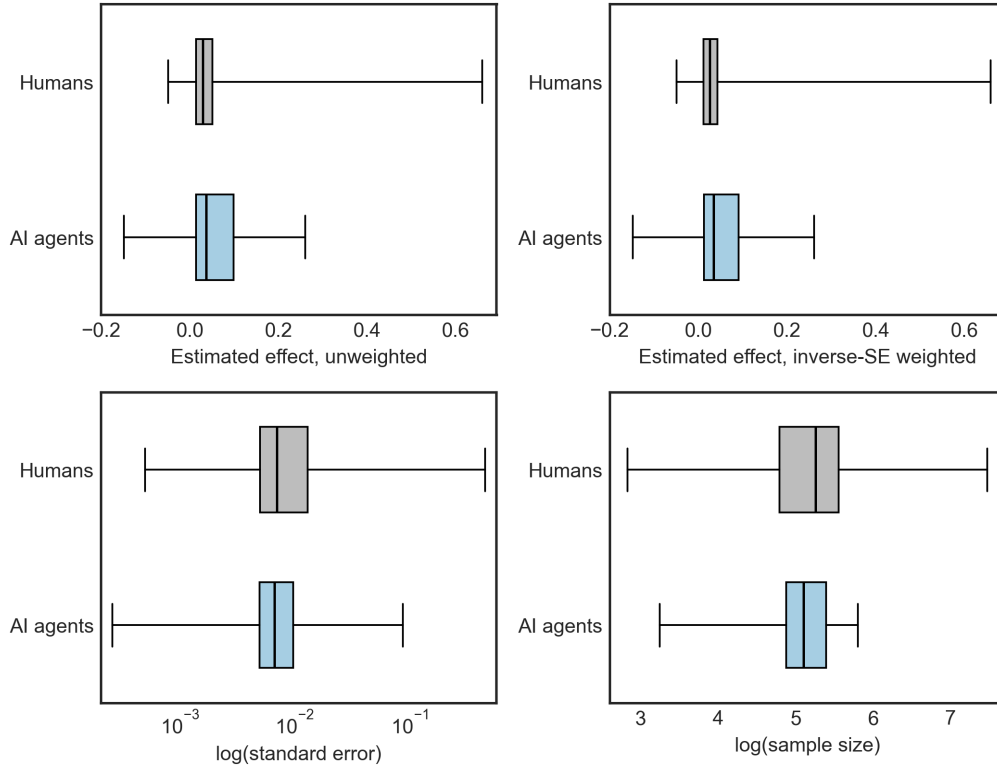
3 Results

I compare 145 human analyses from [Huntington-Klein *et al.* \(2025\)](#) to 156 AI agent analyses. 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. Boxes show the interquartile range, center lines medians, and whiskers minimum and maximum values. Exact values are reported in Table 1.

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, underlining 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'.

Human coefficient estimates are more spread, exhibiting a 13% higher standard deviation in effect size. Still, 92% of AI point estimates are within the min-max range of humans, with about 30% within humans' interquartile range.

For standard errors, the minimum, 25th percentile, and median are identical across humans

Table 1: Summary statistics on estimation outcomes

	N	Mean	SD	Min	Pctl. 25	Median	Pctl. 75	Max
<i>Panel A: AI-agent estimates</i>								
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	155	65,518	28,865	7,717	36,477	74,231	81,508	169,320
<i>Panel B: Human estimates</i>								
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.

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% of the time among humans.

The range of sample sizes chosen by human researchers—between 681 and over 29 million observations—is several orders of magnitude 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.

3.2 Specification heterogeneity

Did AI and human researchers arrive at their estimates through similar specification and sampling choices? Table 2 displays the prevalence of model specification choices between AI and human runs.

Table 2: Estimation choices by researcher type

Category	Choice	AI		Humans	
		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 exclusively choose linear 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 highlighted by [Huntington-Klein *et al.* \(2025\)](#).

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, and AI almost never controlled for education while humans did in nearly half of specifications. 92% of AI agents included year fixed effects, compared to just 24% of human runs. Similarly, 92% of AIs included state fixed effects compared to 36% of human runs. Estimating a fully-saturated difference-in-differences specification diminishes the downsides of linear probability models, the choice of every AI agent, as also noted by [Huntington-Klein *et al.* \(2025\)](#). Overall, AI agents agreed on the set of controls much more often than human researchers. Just 25% of AI agent runs used a set of controls unique among all other AIs, compared to 64% of humans.

Table 3: Control variable choices by researcher type

Category	Control	AI		Humans	
		N	Share (%)	N	Share (%)
AGE	Linear Age	82	52.6	164	37.5
AGE	Age FE	19	12.2	36	8.2
AGE	Age Quadratic	69	44.2	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	143	91.7	103	23.6
STATE/YEAR	State FE	144	92.3	155	35.5
STATE/YEAR	State FE x Year FE	132	84.6	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 specifications which included various controls. Data from [Huntington-Klein *et al.* \(2025\)](#) include more restricted human runs (Tasks 2 and 3) in which the research design was more tightly specified and precleaned data was provided.

Next, Table 4 compares sample and treated group filters. Rows underneath each variable heading are roughly ordered by the ‘correctness’ for defining the treated group, as in [Huntington-Klein *et al.* \(2025\)](#). This is based on DACA’s eligibility criteria, which are listed in the AI’s prompt. The prompt does not specify, however, how to construct the control group.

Table 4: Sample and treated-group restriction choices by researcher type

Variable	AI				Humans			
	All		Treated		All		Treated	
	N	Share (%)	N	Share (%)	N	Share (%)	N	Share (%)
Hispanic	156		156		144		144	
... Hispanic-Mexican	154	98.7	154	98.7	105	72.9	109	75.7
... Hispanic-Any	2	1.3	2	1.3	17	11.8	17	11.8
... Hispanic-Mex or Mex-Born	0	0.0	0	0.0	1	0.7	2	1.4
... None	0	0.0	0	0.0	21	14.6	16	11.1
Birthplace	156		156		145		145	
... Mexican-Born	154	98.7	154	98.7	103	71.0	112	77.2
... Hispanic-Mex or Mex-Born	0	0.0	0	0.0	2	1.4	2	1.4
... Non-US Born	0	0.0	0	0.0	4	2.8	4	2.8
... Central America-Born	0	0.0	0	0.0	1	0.7	1	0.7
... None	2	1.3	2	1.3	35	24.1	26	17.9
Citizenship	156		156		145		145	
... Non-Citizen	153	98.1	153	98.1	83	57.2	117	80.7
... Foreign-Born	2	1.3	2	1.3	2	1.4	2	1.4
... Non-Cit or Natlzd post-2012	0	0.0	0	0.0	4	2.8	7	4.8
... Other	1	0.6	1	0.6	11	7.6	11	7.6
... None	0	0.0	0	0.0	45	31.0	8	5.5
Age at Migration	156		156		145		145	
... < 16	56	35.9	112	71.8	21	14.5	105	72.4
... ≤ 16	0	0.0	0	0.0	10	6.9	25	17.2
... Other	36	23.1	12	7.7	24	16.6	11	7.6
... None	64	41.0	32	20.5	90	62.1	4	2.8
Age in June 2012	156		156		145		145	
... Year-Quarter Age	6	3.8	11	7.1	40	27.6	117	80.7
... Year-Only Age	148	94.9	143	91.7	18	12.4	21	14.5
... Other	0	0.0	0	0.0	2	1.4	0	0.0
... None	2	1.3	2	1.3	85	58.6	7	4.8
Year of Immigration	156		156		145		145	
... < 2007	0	0.0	0	0.0	15	10.3	43	29.7
... ≤ 2007	78	50.0	133	85.3	13	9.0	52	35.9
... < 2012	0	0.0	0	0.0	3	2.1	1	0.7
... ≤ 2012	0	0.0	0	0.0	2	1.4	4	2.8
... Any Year	44	28.2	9	5.8	7	4.8	4	2.8
... Other	26	16.7	9	5.8	5	3.4	3	2.1
... None	8	5.1	5	3.2	100	69.0	38	26.2
Education/Veteran	156		156		145		145	
... HS Grad or Veteran	0	0.0	0	0.0	0	0.0	3	2.1
... 12th Grade or Veteran	0	0.0	0	0.0	0	0.0	0	0.0
... HS Grad	0	0.0	0	0.0	13	9.0	21	14.5
... HS Grad or Non-Veteran	0	0.0	0	0.0	0	0.0	0	0.0
... Other	0	0.0	0	0.0	3	2.1	6	4.1
... None	156	100.0	156	100.0	129	89.0	115	79.3
Years Continuous in USA	156		156		145		145	
... Used YRSUSA	1	0.6	4	2.6	23	15.9	55	37.9
... No YRSUSA	155	99.4	152	97.4	122	84.1	90	62.1

Notes: The table compares AI-agent restriction choices with the published Task 1 counts from Table 6 of [Huntington-Klein *et al.* \(2025\)](#). Treated-group restrictions inherit those used to restrict the whole sample.

Once again AI agents overwhelmingly converged on several sample and treated-group construction choices. Virtually all AI agents restricted the sample to ethnically Hispanic-Mexicans born in Mexico who are not US citizens.⁴

⁴I suspect this results from the AI becoming highly focused on the first clause of the AI prompt (copied

Moreover, AI consistently made two subtle mistakes in coding treatment more frequently than the human researchers. First, AI was much less likely to condition treatment on age derived from both birth year and birth quarter (7% vs. humans’ 81%), instead preferring age measured only in years. This matters because the DACA eligibility age cutoff is June 15th, 2012, as conveyed in the AI’s prompt. Second, AI inappropriately includes immigrants arriving in 2007 as part of the treated group (85% vs. humans’ 36%), when in fact it should have included only those arriving strictly before 2007. AI never conditions treatment on arrival prior to 2007 (0% compared to humans’ 30%), rather defining it conditional on arrival up to and including 2007 (85% of the time vs. humans’ 36%).

4 Conclusion

This paper develops a novel, scalable approach to multi-analyst studies leveraging new breakthroughs in agentic artificial intelligence. I then test whether current frontier models can match humans in outcomes and choices following the multi-analyst study of [Huntington-Klein *et al.* \(2025\)](#).

The resemblance of AI agent estimation outcome distributions to the human distributions is striking. The sharp disagreement in specification choices between AI and humans is equally striking. AI agents converged much more on particular specification and sampling choices relative to human researchers. This finding is consistent with concurrent work by [Huang *et al.* \(2026\)](#), who find in a different setting that AI choices tend to cluster on certain sampling and specification choices. This suggests that AI is at best an imperfect substitute for humans in multi-analyst designs aimed at quantifying the degree of uncertainty due to researcher degrees of freedom.

Two limitations are worth emphasizing. First, the [Huntington-Klein *et al.* \(2025\)](#) benchmark materials were public before my AI runs, so model training data may have included parts of the original study or human submissions. Still, the convergence of many AI choices in

verbatim from the human task), “**Among ethnically Hispanic-Mexican Mexican-born people living in the United States**, what was the causal impact of eligibility for the Deferred Action for Childhood Arrivals (DACA) program (treatment) on the probability that the eligible person is employed full-time (outcome),” emphasis mine.

contrast to the wide divergence of humans suggests the AI was not merely copying human work. In any case, a novel research question simultaneously assigned to human researchers and AI agents would best address this concern. Second, results may depend on the AI model and agent harness. Future work should explore the degree to which outcomes vary by model-harness pair.

References

- ANTHROPIC (2026). System card: Claude sonnet 4.6.
- 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). The 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).
- MENKVELD, A. J., DREBER, A., HOLZMEISTER, F. *et al.* (2024). Nonstandard errors. *Journal of Finance*, **79** (3).
- OPENAI (2025). GPT-5 system card.

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.

STEVENSON, M. T. and FISCHMAN, J. B. (2026). Humans in the loop: The next frontier in the credibility revolution, I4R Discussion Paper Series, DP No. 296.