How (un)Stable Are LLM Occupational Exposure Scores?

Evidence from Multi-Model Replication

Michelle Yin[1], Hoa Vu[2], & Claudia Persico[3]

Abstract

A rapidly growing literature estimates AI's labor-market effects using large language models (LLMs) to self-assess occupational exposure. We demonstrate these measures are highly fragile. Replicating the dominant rubric with three frontier models on identical tasks, we find a 3.6-fold divergence in mean exposure with agreement as low as 57%. This measurement instability alters downstream empirical conclusions: in a difference-in-differences estimates framework, individual-level coefficient magnitudes vary 2.4-fold across annotators, and county level estimates flip from a significant negative to an insignificant positive depending entirely on the annotator. We formalize this non-classical measurement error, highlighting the risks of treating evolving LLMs as static research instruments.

JEL: J23, J24, O33, C81

Keywords: occupational exposure, measurement error, large language models, AI and labor markets, task-based framework

[1] Corresponding author. School of Education and Social Policy, Northwestern University. Email: michelle.yin@northwestern.edu
[2] School of Education and Social Policy, Northwestern University. Email: hoa.vu@northwestern.edu
[3] School of Public Affairs, American University and IZA and also NBER. Email: cpersico@american.edu

Since 2023, a rapidly growing literature has studied how large language models (LLMs) affect employment, wages, and the structure of work. The modern study of technology and the labor market begins with the task-based framework of Autor, Levy, and Murnane (2003), which decomposes occupations into discrete tasks and asks which tasks are susceptible to computerization. Subsequent work operationalized this framework as occupational exposure indices. Frey and Osborne (2017) elicited expert assessments of automation probability. Brynjolfsson, Mitchell, and Rock (2018) constructed suitability-for-machine-learning scores. Webb (2020) linked patent text to occupational task descriptions. Felten, Raj, and Seamans (2018, 2023) mapped AI capabilities to occupational abilities using performance benchmarks. Each of these measures relies on human judgment or algorithmic text matching, and each captures a different dimension of technological exposure.

The introduction of capable LLMs created a new measurement possibility: using the technology itself as the rating instrument. Eloundou et al. (2024) developed the approach that has become dominant. Their rubric classifies each O*NET task into three tiers: E0 (no meaningful LLM assistance), E1 (direct assistance via a chat interface), and E2 (assistance via LLM-powered software). The rubric asks whether an LLM could reduce task completion time by at least 50% at equivalent quality. Applying this rubric with both human annotators and GPT-4, and aggregating task-level labels to occupational scores, Eloundou et al. produced exposure measures that are highly correlated across the two rating approaches. The GPT-4-generated scores have become the standard treatment variable in the downstream literature. Their paper, published in *Science*, has been cited in over 1,500 subsequent studies. Beyond economics, the International Labour Organization (ILO) replicated the rubric to estimate global AI exposure across International Standard Classification of Occupations (ISCO) categories (Gmyrek, Berg, and

Bescond 2023). The International Monetary Fund (IMF) extended it to compare advanced and emerging economies (Pizzinelli et al. 2023). The U.S. Bureau of Labor Statistics (BLS) incorporates the scores into its 10-year employment projections (Machovec, Rieley, and Rolen 2025). The Organization for Economic Co-operation and Development (OECD) and World Economic Forum use related measures to assess AI's workforce impact (Green 2024; World Economic Forum 2025).

The downstream economics literature now spans several subfields. Eisfeldt, Schubert, Zhang, and Taska (2023) construct firm-level generative AI exposure from GPT-3.5 Turbo classifications of O*NET tasks. Rock et al. (2025) extend the scores to firm-level analysis using Revelio Labs data. Acemoglu (2025) uses the Eloundou et al. exposure data directly in his calibration of the macroeconomic productivity effects of AI, estimating that AI could raise total factor productivity by no more than 0.66% over ten years. Humlum and Vestergaard (2025) use the scores to study AI adoption and labor-market outcomes in Denmark. International replications have appeared for China and OECD economies. Hui, Reshef, and Zhou (2024) study the impact on freelance labor markets.

A parallel stream of work has moved toward measuring actual usage rather than hypothetical capability. The Anthropic Economic Index (AEI) (Handa et al. 2025; Appel et al. 2026) analyzes millions of Claude conversations using an automated classification tool (Clio) to identify which O*NET tasks are performed with AI assistance. Massenkoff and McCrory (2026) construct a composite measure that combines the Eloundou et al. capability scores with Anthropic usage data and find that their composite predicts BLS employment projections, while the rubric-based scores alone do not. In each case, the exposure measure is generated by a single LLM, or at most two, and enters the analysis as a fixed occupational characteristic.

Unlike traditional labor variables derived from surveying workers and employers, the exposure score is generated by the very technology whose effects are being studied. If different models rate tasks differently, the scores are not fixed properties of occupations but joint products of the occupation and the model, and any model-specific bias propagates directly into downstream coefficient estimates. This concern is not unique to occupational measurement. Studies using LLMs for text classification (Gilardi, Alizadeh, and Kubli 2023), survey simulation (Argyle et al. 2023), and qualitative coding have documented inter-model disagreement and sensitivity to prompt framing. In the specific domain of occupational measurement, Xu et al. (2025) replicate the rubric in China using GPT-4, InternLM, and GLM, and report agreement patterns that vary across models, though they do not examine downstream consequences. Eisfeldt, Schubert, Zhang, and Taska (2023) use GPT-3.5 Turbo rather than GPT-4 to classify tasks, implicitly choosing a different annotator than Eloundou et al. without testing sensitivity to this choice.

We replicate the Eloundou et al. rubric using three frontier models available in early 2026: ChatGPT-5 (OpenAI), Gemini 2.5 (Google DeepMind), and Claude 4.5 (Anthropic). We refer to each rating model as an annotator, by analogy with human coders in content analysis, because it performs the same function of classifying tasks according to a rubric. Applying identical rubric language to identical tasks, we find that pairwise agreement falls to 56.9%, with Cohen's kappa as low as 0.36. Mean E1 exposure (direct LLM assistance, the margin most studies emphasize) ranges from 0.14 to 0.51 across annotators. The cross-tabulation of task-level classifications reveals that this disagreement is not noise: each model applies a systematically different threshold for what constitutes LLM-capable work, and the discrepancies concentrate at specific classification boundaries.

We show that this measurement instability changes the conclusions researchers draw. In a difference-in-differences (DiD) framework estimating the effect of LLM exposure on employment at the individual level, all annotators yield significant negative coefficients, but the Gemini coefficient is 2.4 times the original GPT-4 baseline. At the county level, the original GPT-4 scores and two of the three 2026 models produce no statistically significant association. Yet, one model (Claude 4.5) produces a statistically significant negative estimate. This reversal reflects how annotator-specific calibration bias interacts with geographic industrial structure. Conservative annotators concentrate regressor variance in a small subset of highly exposed professions clustered in technology hubs, compressing the employment-weighted exposure metric. Generous annotators assign non-zero exposure across a much broader range of occupations, smoothing geographic variance and yielding a different distribution of regional treatment intensity. A researcher's conclusion about whether LLM exposure reduces employment, and by how much, depends on an unreported and untested choice: which model rated the tasks.

Our contribution is methodological and empirical. We formalize the measurement problem within the econometrics of non-classical error (Bound, Brown, and Mathiowetz 2001). The bias is theoretically ambiguous because it operates through two channels: model-specific calibration tendencies that compress or amplify regressor variance, and an adoption-exposure feedback loop that generates correlation between the measurement error and the dependent variable. The sign instability we observe in the county-level estimates is consistent with this framework. We also identify a rubric-model alignment problem: the dominant rubric was designed for and calibrated with GPT-4, and its predictive validity deteriorates when applied by successor models to self-assess their own expanded capabilities. As an external benchmark, we

compare each annotator's scores against observed AI usage from the AEI, which measures the share of Claude conversations associated with each O*NET task. Usage-based measures have limitations, including platform selection and early-adopter composition bias, but they provide a diagnostic for how well rubric-based scores track actual behavior. All annotators correlate positively with usage (Spearman rho = 0.30 to 0.43), but the correlations are moderate and fall within a narrow range despite a 3.6-fold difference in mean exposure. The original GPT-4 scores predict usage best (R-squared = 0.169), with the 2026 models explaining less (R-squared = 0.037 to 0.104): the rubric was designed for GPT-4, and the fit deteriorates when newer models apply the same criteria. The measurement problem is therefore not only about which model rates the tasks but also about whether the rubric was designed for the rating model.

The implications extend beyond the specific coefficients we report. The downstream literature has already reached divergent conclusions. Brynjolfsson, Chandar, and Chen (2025) find concentrated employment declines among young workers in high-exposure occupations. Johnston and Makridis (2025) find wage and employment gains in exposed sectors. Both studies condition on the same exposure scores. Our results suggest that some of this divergence may reflect sensitivity to the exposure measure itself, not only differences in research design or data. More broadly, the practice of using LLMs as research instruments is expanding across economics. Occupational exposure measurement is a high-stakes test case. Until the field develops exposure measures whose properties are stable across rating instruments, or adopts multi-annotator sensitivity as a reporting standard, the rapidly growing empirical literature on AI and the labor market rests on a measurement foundation that is itself a function of the technology it seeks to evaluate.

I. Measurement Error Framework

Let θ_i denote the true LLM exposure of occupation i, defined as the share of tasks where an LLM could reduce completion time by at least 50% at equivalent quality. The true structural relationship is:

$$(1) \qquad y_i = \beta\theta_i + \pi a_i + \mathbf{X}_i\Gamma + u_i$$

where y_i is the labor market outcome, a_i is the AI adoption rate in occupation i, $\mathbf{X}_i$ is a vector of controls, and u_i is an error term orthogonal to the regressors after conditioning on controls. In practice, θ_i is unobserved. Researchers instead condition on an annotator-specific proxy:

$$(2) \qquad \theta_i^m = \theta_i + \varepsilon_i^m$$

where m indexes the rating model. The literature sets $\varepsilon_i^m = 0$ for a single chosen m. Classical measurement error requires $E[\varepsilon_i^m|\theta_i] = 0$. Our data reject this condition. Claude classifies 2,728 tasks as E1 that Gemini classifies E0. Only 56 tasks move in the opposite direction. This directional asymmetry implies that the measurement error is model-specific and correlated with both the true exposure and observed adoption:

$$(3) \qquad \varepsilon_i^m = \delta^m\theta_i + \varphi^m a_i + \nu_i^m$$

The parameter δ^m governs the calibration channel: a generous rater ($\delta^m > 0$) systematically overstates exposure, while a conservative rater ($\delta^m < 0$) understates it. The parameter φ^m governs the adoption feedback channel: occupations with higher AI usage generate training data that causes subsequent models to rate those tasks more favorably ($\varphi^m > 0$). The term ν_i^m is classical noise, independent across annotators.

When a researcher estimates the naive specification $y_i = \beta^m \theta_i^m + \mathbf{X}_i\Gamma + \tilde{u}_i$, omitting the unobserved adoption rate, the probability limit decomposes into two sources of bias (derived in Appendix A):

(4) $\quad \text{plim}\ \hat{\beta}^m = \beta\lambda^m + \pi\Omega^m$

The first term, $\beta\lambda^m$, captures the calibration bias. The reliability factor λ^m depends on δ^m: when the rater is generous ($\delta^m > 0$), the denominator of λ^m grows faster than the numerator, attenuating the estimate toward zero. When the rater is conservative ($\delta^m < 0$), attenuation is weaker and amplification is possible. The second term, $\pi\Omega^m$, captures adoption confounding: OLS on θ_i^m picks up part of the adoption effect π through the feedback component φ^m in the error. Because π varies by outcome and time horizon (positive for short-run productivity, potentially negative for medium-run employment), the direction of the second bias is ambiguous.

Different annotators produce different (δ^m, φ^m) pairs, generating different biases. Because the two bias terms may have opposite signs, the direction of the total bias is theoretically ambiguous. This is the pattern we observe: coefficient magnitudes vary non-monotonically across annotators, and the qualitative conclusion at the county level depends on the annotator. The calibration channel predicts that generous raters should produce smaller individual-level coefficients, which we test in Section III. The adoption channel predicts that occupations with higher observed AI usage should show larger increases in measured exposure across model generations, which we test in Section III.E.

A further complication is that both channels evolve over time. As models become more capable, the calibration parameter δ^m shifts: a task that GPT-4 rated as E0 may be rated E1 by a successor model that genuinely can perform it faster. The adoption parameter φ^m increases as

models are trained on more user-generated data, creating a feedback loop between capability growth, and measured exposure. The implication for empirical work is that the measurement error in LLM exposure scores is not a static nuisance to be corrected once. It is a moving target that evolves with the technology itself.

From an identification standpoint, the standard approach in this literature treats exposure as a predetermined variable, fixed at the time of annotation and orthogonal to subsequent labor-market dynamics. Our framework suggests that this assumption fails on both counts. The exposure score is not predetermined because it depends on the annotator's vintage (through δ^{m}), and it is not orthogonal because the adoption channel (through φ^{m}) generates correlation with the outcome. These features distinguish the LLM exposure setting from classical measurement error problems in labor economics, where the mismeasured variable (e.g., education, income) is at least conceptually fixed and the error is typically assumed to be random.

II. Replication Design and Data

We apply the Eloundou et al. (2024) rubric without modification to the O*NET 30.0 task universe using three frontier LLMs: ChatGPT-5, Gemini 2.5, and Claude 4.5. Each model classifies approximately 18,797 tasks as E0, E1, or E2 using the original prompt language at temperature zero. We do not alter the rubric text, the task descriptions, or the category definitions. The prompt asks each model whether it, or software built on top of it, could reduce a worker's time to complete a given task by at least 50% while maintaining equivalent quality. This is a self-assessment: the model is asked to judge its own capability on each task. Occupational exposure scores are constructed by averaging task-level labels within each occupation, following the Eloundou et al. specification. We report both unweighted averages and

core-weighted averages that weight tasks by their O*NET importance ratings. The two approaches produce near-identical scores (correlation above 0.99), and all results reported in the main text use unweighted averages for comparability with the existing literature.

For the downstream analysis, we use the Current Population Survey for 2015 to 2024. The individual-level sample includes approximately 9.6 million person-year observations for workers aged 16 to 64. The dependent variable is an indicator for employment. We assign each worker an LLM exposure score based on their detailed occupation code, merging via the standard SOC crosswalk. The county-level sample aggregates employment rates to 4,825 county-year cells, constructed by computing the share of working-age adults who are employed within each county-year. Both specifications implement a DiD design in which occupational LLM exposure is the continuous treatment variable, interacted with a post-ChatGPT period indicator (2023 to 2024).

The individual-level specification is:

$$(3) \quad y_{ijct} = \alpha + \beta\, E1_j \times Post_t + \mathbf{X}_i'\Gamma + \eta_j + \lambda_t + \mu_c + \delta_s \times t + \mathbf{Z}_{st}'\Pi + u_{ijct}$$

where y_{ijct} is an employment indicator for individual i in occupation group j, county c, in year t. $E1_j$ is the occupational LLM exposure score. $Post_t$ equals one for 2023-2024; $\mathbf{X}_i$ is a vector of individual characteristics (age, gender, race/ethnicity, education, marital status, number of children). η_j are occupation group fixed effects, λ_t are year fixed effects, μ_c are county fixed effects, $\delta_s \times t$ are state-specific linear trends, and $\mathbf{Z}_{st}$ captures state-level COVID-19 cases and deaths. The parameter of interest is β, which measures the change in employment associated with a unit increase in LLM exposure after the introduction of ChatGPT. Standard errors are clustered at the occupation level.

The county-level specification aggregates to county-year cells:

$$(4) \quad \bar{y}_{ct} = \alpha + \beta \, \bar{E}1_c \times \text{Post}_t + \lambda_t + \mu_c + \delta_s \times t + \mathbf{Z}_{st}'\Pi + u_{ct}$$

where $\bar{y}_{ct}$ is the employment rate in county *c* in year *t,* and $\bar{E}1_c$ is the employment-weighted average occupational exposure in county *c*. County fixed effects absorb time-invariant local characteristics, and state-specific linear trends control for pre-existing diverging trajectories across states.

We assess the parallel trends assumption with the event study specification:

$$(5) \quad y_{ijct} = \alpha + \Sigma_{t \neq 2022} \, \beta_t \, E1_j \times \mathbf{1}(\text{Year} = t) + \mathbf{X}_i'\Gamma + \eta_j + \lambda_t + \mu_c + \delta_s \times t + \mathbf{Z}_{st}'\Pi + u_{ijct}$$

where the β_t coefficients trace the employment-exposure relationship in each year relative to 2022, the last pre-treatment year. Pre-period coefficients near zero would support the parallel trends assumption. Each specification is estimated four times, once per annotator, using an identical sample, controls, and estimator. The only object that changes across columns is the exposure score. This design isolates the contribution of annotator choice to the variation in estimated effects.

III. Results

A. Cross-model disagreement

Table 1 reports pairwise agreement rates and Cohen's κ statistics. The best-performing pair (Gemini-ChatGPT-5) achieves 73% agreement (κ = 0.56, classified as moderate under the McHugh 2012 conventions). The worst-performing pair (Gemini-Claude) achieves 56.9% (κ =

0.36, fair). For comparison, Eloundou et al. report approximately 76% agreement between human annotators and GPT-4 at the task level.

Figure 1 shows the cross-tabulation of task-level classifications. The disagreements are concentrated at specific boundaries. The E0/E1 boundary accounts for most of the Gemini-Claude divergence: Gemini assigns E0 to 3,073 tasks that Claude rates as at least E1. The E1/E2 boundary drives more of the Gemini-ChatGPT-5 divergence (2,378 tasks rated E2 by Gemini but E1 by ChatGPT-5). These patterns are consistent with model-specific thresholds rather than random classification noise: each model has a characteristic tendency toward generosity or conservatism in assessing LLM capability, and this tendency is stable across task types.

As shown in Panel C of Table 2, mean occupational E1 exposure is 0.51 (Claude), 0.30 (ChatGPT-5), 0.14 (Gemini), and 0.14 (original GPT-4). The 3.6-fold range between the highest and lowest estimates occurs on identical tasks, under identical rubric language, in the same observation window. No property of the occupations changed; only the rating model changed. The fact that Gemini replicates the original GPT-4 mean almost exactly while ChatGPT-5, its direct successor within OpenAI, more than doubles it indicates that the problem exists even within a single provider's model family and is therefore not attributable to cross-platform differences in architecture or training data alone.

B. Consequences for employment estimates

Table 2 reports DiD estimates of the effect of E1 exposure on employment. Panel A presents individual-level results (N = 9,596,366). Panel B presents county-year-occupation level results (N = 4,104,098). Each column corresponds to a different annotator.

At the individual level (Panel A), all four annotators produce significant negative coefficients. The original GPT-4 score yields −0.0066 (SE = 0.0018). ChatGPT-5 yields −0.0095 (SE = 0.0016). Claude yields −0.0097 (SE = 0.0011). Gemini yields −0.016 (SE = 0.0011). The Gemini estimate is 2.4 times the original baseline, and its standard error is large enough that the 95% confidence intervals of the Gemini and original estimates do not overlap. A researcher using the original scores would characterize the employment effect as small. A researcher using Gemini scores would characterize it as moderate. Both are significant, but the policy implications of a 0.66 percentage point effect differ from those of a 1.6 percentage point effect.

The county-level results (Panel B) produce a qualitative reversal. The original GPT-4 coefficient is −0.0011, with a standard error of 0.0057. The estimate is not distinguishable from zero. Under ChatGPT-5, the coefficient is −0.0028 and also not statistically significant at conventional levels. Under Gemini, it is 0.0032, but also not statistically significant. Under Claude, it is −0.0077 ($p < 0.05$). Two of the three 2026 annotators fail to reject the null along with the original GPT-4. A researcher using the original measure would report no significant county-level relationship between LLM exposure and employment. A researcher using Claude would report a significant negative relationship. This is not a difference in degree. It is a difference in kind.

The ordering across annotators is informative. Claude produces the highest mean exposure (0.51) and the largest county-level coefficient among the 2026 models (−0.0077). Gemini produces the lowest mean exposure (0.14) but a positive county-level coefficient (0.0032). At the individual level, the pattern is consistent with the compression-amplification mechanism described in Section I: conservative raters preserve variation among highly exposed occupations and produce larger individual-level coefficients (Gemini: -0.016), while generous

raters compress variation and attenuate estimates (Claude: -0.0097). The county-level reversal reflects the interaction between annotator-specific calibration and geographic industrial structure. Conservative annotators like Gemini assign high exposure only to a narrow set of digitally intensive occupations that are disproportionately concentrated in technology and financial hubs. This concentrates regressor variance geographically, creating sharp between-county contrasts in the employment-weighted exposure metric. Generous annotators like Claude assign non-zero exposure across a much broader range of occupations, smoothing the geographic distribution and compressing between-county variation. The county-level specification, which identifies the effect from between-county differences in occupational composition, is therefore far more sensitive to the annotator's calibration tendencies than the individual-level specification, which exploits within-county, between-occupation variation. The fact that only one of three 2026 models rejects the null at the county level, despite 57 to 73% pairwise agreement, underscores that the choice of annotator determines not only the magnitude but also the significance and even the sign of the estimated effect.

C. Event study evidence

Figure 2 presents event study plots for the individual-level specification, overlaying all four annotators in a single panel. Pre-period coefficients exhibit a negative trend, which may partly reflect the choice of normalization year: 2021 coincides with severe COVID-19 labor market disruptions, making it an unstable baseline against which pre-period years are compared. Figure A1 addresses this by dropping 2020 and 2021 from the sample and normalizing at 2019 instead; pre-period coefficients are substantially smaller in magnitude and closer to zero, providing stronger support for the parallel trends assumption. The magnitude of the post-period break

varies across annotators in the direction predicted by the cross-sectional results: largest under Gemini, smallest under the original GPT-4. The visual difference across annotators in Figure 2 conveys the measurement problem more directly than any regression table can. The same data, the same specification, and the same time period produce event study trajectories that differ in magnitude by a factor of two, solely because a different model assigned the exposure scores. County-level event studies, reported in the online appendix Figure A2, show wider confidence intervals but a qualitatively similar pattern.

D. Heterogeneity across occupational groups

Figure 3 presents mean E1 exposure by SOC major group for each annotator, with occupation groups ordered by the original GPT-4 scores from Eloundou et al. (2024). The ordering reveals that even the broad rank ordering of sectors is not preserved across annotators: while Computer and Mathematical occupations are consistently the most exposed, other groups shift position substantially. Absolute levels diverge even more. Computer and Mathematical occupations range from approximately 0.42 under Gemini to 0.95 under Claude. Management occupations range from roughly 0.08 to 0.83. The cross-model spread is largest for white-collar occupations near the middle of the exposure distribution, where classification boundaries are most contested. Educational Instruction, Life and Physical Sciences, and Sales occupations all show spreads of 0.30 or more across annotators. For occupations at the extremes, the models agree because the tasks are unambiguously physical (Construction, Farming) or unambiguously linguistic (Computer Programming, Data Science).

This heterogeneity in the measurement error across sectors has direct consequences for applied work. Studies that examine differential exposure effects by industry or occupation group

are conditioning on a variable whose cross-sectional distribution depends on the annotator. A researcher studying whether high-exposure service-sector occupations are displacing workers faster than high-exposure manufacturing occupations would obtain different relative exposure rankings depending on the model. The measurement problem is not uniform across the economy; it is concentrated precisely in the sectors where the boundary between LLM-capable and non-LLM-capable work is most ambiguous, which are also the sectors where policymakers most need reliable estimates.

E. Evidence on the adoption channel

We test the feedback mechanism described in Section I by regressing the change in measured exposure on observed adoption. The dependent variable is the change in E1 scores, defined as the mean across the three 2026 models minus the original GPT-4 baseline, at the occupation level. We use two adoption measures. The first is the occupation-level AEI task coverage (Handa et al. 2025), which captures the share of Claude conversations associated with each occupation. The second is an industry-level proxy from the Census Bureau's Business Trends and Outlook Survey (BTOS), which reports the share of firms using AI by NAICS subsector. We weight these industry-level rates to occupations using the OES employment matrix. Both regressions control for the baseline GPT-4 exposure score.

The results, reported in Table A1, confirm that adoption feeds back into measured exposure through both channels. The Anthropic usage coefficient is 0.335 (SE = 0.154, $p < 0.05$), indicating that a one-standard-deviation increase in observed AI usage is associated with a meaningful increase in measured exposure change. The BTOS coefficient is 0.009 (SE = 0.003, $p < 0.01$). The Anthropic specification explains substantially more variation (R-squared = 0.252)

than the BTOS specification (R-squared = 0.084), consistent with the occupation-level measure being a more direct proxy for the mechanism. That is, workers in specific occupations generate the training data that shifts subsequent models' ratings, and this channel operates at the occupation level rather than the industry level. These findings support the second channel formalized in Section I: occupations with higher observed AI adoption show larger increases in measured exposure across model generations, even after controlling for baseline exposure.

IV. Robustness

Within-model test-retest reliability confirms that the cross-model disagreement we document reflects systematic calibration differences rather than stochastic noise. We re-ran each model three times on a 10% subsample of tasks at temperature zero. The results of this exercise are shown in Table A2. Within-model agreement exceeds 90% for all models tested: Claude 4.5 achieves 99.0% agreement across runs (kappa=0.98), ChatGPT-5 achieves 96.7% (kappa=0.93), and Gemini 2.5 achieves 90.8% (kappa=0.85). These within-model rates are far above the 57 to 73% cross-model rates reported in Table 1, confirming that the measurement problem is more between models, not within.

Composite exposure measures that combine E1 and E2 classifications narrow the cross-model spread: the max/min ratio of mean exposure falls from 3.6 for E1 alone to 1.9 for E1 + E2 (Table A3, Panel C). The narrowing reflects the fact that models agree more readily that software built on LLMs can assist a given task (E2) than that a bare LLM can (E1). The downstream coefficient instability, however, persists at the composite level (Table A3, Panels A-B). The measurement problem is attenuated but not resolved by broadening the exposure definition.

We benchmark each annotator against observed AI usage using the AEI. The results are presented in Table A4. Aggregating task-level coverage to the occupation level and computing Spearman rank correlations across 705 occupations, we find that all four annotators correlate positively with actual usage, but the original model correlates least well: ChatGPT-4 (rho = 0.298, $p < 0.001$), ChatGPT-5 (rho = 0.425, $p < 0.001$) Gemini 2.5 (rho = 0.406, $p < 0.001$), and Claude 4.5 (rho = 0.419, $p < 0.001$). Panel B uses full exposure (E1 + E2) and shows slightly different patterns: ChatGPT-4 (rho = 0.335), ChatGPT-5 (rho = 0.372), Gemini 2.5 (rho = 0.315), and Claude 4.5 (rho = 0.364), all significant at $p < 0.001$. Here, including indirect exposure lowers the correlation for Gemini 2.5 below that of ChatGPT-4, though ChatGPT-5 and Claude 4.5 continue to show the strongest alignment with observed usage.

However, OLS regressions of occupation-level Anthropic usage on E1 scores (Table A5) reveal a more complex pattern: the Eloundou et al. (2024) GPT-4 scores predict usage best (R-squared = 0.169), while Claude 4.5 predicts least (R-squared = 0.037). Among the three 2026 models, the most conservative rater (Gemini, R-squared = 0.104) predicts usage best and the most generous (Claude) predicts worst, consistent with the compression mechanism in Section I. A rubric-model alignment pattern is evident: the original Eloundou GPT-4 scores predict usage best (R-squared = 0.169), far exceeding the three 2026 models (R-squared = 0.037 to 0.104). The Eloundou rubric was designed for and calibrated with GPT-4, creating a natural fit between the rubric's thresholds and the model's self-assessment that breaks when newer models apply the same rubric to rate their own expanded capabilities. The measurement problem is not only about which model rates the tasks but also about whether the rubric was designed for the rating model. The composite exposure measure (E1 + E2) correlates less well with usage than E1 alone (rho =

0.315 to 0.372), suggesting that the broader measure captures hypothetical capability that has not translated into observed adoption.

Two features of these correlations are notable for the measurement problem. First, no annotator stands out as substantially better calibrated to actual behavior: the three correlations fall within a range of 0.02, which means the choice of annotator, despite producing a 3.6-fold difference in mean exposure, has almost no effect on the rank-order alignment with usage. Second, the correlations are moderate at best, indicating that all rubric-based measures, regardless of annotator, capture less than half of the variation in how workers actually use LLMs. These findings reinforce the case for usage-based measurement and suggest that the rubric-based scores are measuring a different object than adoption (Appendix Table A4).

OLS regressions of occupation-level Anthropic usage on E1 scores provide a complementary measure of predictive power (Table A5). The R-squared values reveal two patterns. First, among the three 2026 models, predictive power decreases with generosity: Gemini 2.5 (R-squared = 0.104), ChatGPT-5 (0.076), Claude 4.5 (0.037). The most conservative annotator explains the most variation in observed usage; the most generous explains the least. This is consistent with the compression mechanism formalized in Section I: generous rating assigns high exposure to most occupations, destroying the between-occupation variation that predicts adoption. Second, the Eloundou et al. (2024) GPT-4 scores predict usage best of all (R-squared = 0.169). The original GPT-4 scores benefit from rubric-model alignment: the rubric was designed for GPT-4's capability level, producing a natural calibration that newer models applying the same rubric to their expanded capabilities do not share. Even so, the best-performing measure explains less than 17% of the cross-occupational variation in observed AI usage, confirming that rubric-based and usage-based measures capture different objects.

As a robustness check, we re-estimate the event study and DiD specifications normalizing to 2019 and excluding 2020-2021 to remove pandemic-related confounds from the pre-treatment window. Under this specification, the individual-level estimates are no longer statistically significant for most annotators (Table A6), and the pre-trend coefficients are smaller in magnitude and closer to zero (Figure A1). The cross-annotator pattern persists: coefficient magnitudes continue to vary substantially across rating models. Across three distinct DiD specifications (continuous E1 treatment in Table 2, composite E1+E2 treatment in Table A3, and 2019-normalized with pandemic years excluded in Table A6), the cross-annotator coefficient range remains wide and the county-level sign instability persists. That the qualitative conclusion about AI displacement depends jointly on the choice of annotator and the choice of pre-treatment normalization reinforces our central point: published estimates in this literature are conditional on modeling choices that are typically unreported and untested.

V. Discussion

Researchers across fields are attempting to understand the labor-market consequences of LLMs in real time. The primary measurement tool for this effort is a set of occupational exposure scores generated by a single LLM. We have shown that these scores are sensitive to the choice of rating model in ways that change empirical conclusions. The question is what we should do about it.

The most immediate step is to adopt multi-annotator sensitivity as a reporting standard. The practice of conditioning on a single model's scores without testing alternatives is analogous to reporting regression results from a single specification without robustness checks. Our county-level estimates make the case concretely: the original scores and two of three successor models

yield no significant effect, while one successor model yields a significant negative. At the individual level, all four annotators are significant but the magnitude spans a 2.4-fold range. No statistical procedure can determine which annotator is correct without access to the true exposure. But reporting the range of estimates across annotators communicates the precision of the conclusion. Where estimates diverge, researchers can treat the annotator-specific coefficients as bounds on a partially identified parameter (Manski 2003), formalizing the uncertainty that currently goes unreported.

More fundamentally, researchers should reconsider whether asking an LLM to assess its own capabilities is the right measurement strategy. The approach is circular: the technology being studied serves as the instrument that measures its own reach. The calibration biases we document are inherent to this design. The problem is compounded by rubric-model alignment: the Eloundou et al. rubric was designed for GPT-4, and the original GPT-4 scores predict actual usage far better (R-squared = 0.169) than any 2026 model's self-assessment (R-squared = 0.037 to 0.104). When successor models evaluate their own expanded capabilities using this static rubric, the resulting scores detach from observed adoption patterns. The field cannot simply update exposure scores by re-running an old rubric with a new model without introducing systematic bias relative to the original scores.

Usage-based measures avoid this circularity by grounding measurement in observed behavior. The AEI maps millions of conversations to O*NET tasks, measuring what workers actually do with LLMs rather than what a model predicts they could do. Our Spearman correlations in Section IV confirm that rubric-based scores are at best moderately correlated with observed usage (rho = 0.32 to 0.43). Furthermore, the original GPT-4 scores used throughout the downstream literature are the least well-calibrated to actual adoption in rank-order terms, though

the Eloundou GPT-4 scores predict usage levels best among all annotators (R-squared = 0.169), likely reflecting alignment between the rubric design and the model that generated the scores. Usage-based measures have their own limitations, including platform selection and early-adopter composition, but they do not inherit the calibration biases we document because they measure adoption, not hypothetical capability. The composite score developed by Massenkoff and McCrory (2026), however, incorporates the annotator-specific instability we document. They report that their composite measure predicts BLS employment projections whereas rubric-based scores alone do not, which independently validates our conclusion that rubric-based measures are insufficient as standalone treatment variables without usage anchoring. Yet the extent to which the instability propagates into their composite depends on the relative weight given to the rubric component versus the usage component, a question their analysis does not address.

Finally, the assumption that occupational exposure is time-invariant requires further assessment. Even within the OpenAI model family, ChatGPT-5 doubles the E1 estimates produced by GPT-4. Capabilities advance non-uniformly across task types, with coding and data analysis seeing larger gains than interpersonal or physical tasks. A study that treats 2023 exposure scores as fixed when analyzing 2025 outcomes conditions on a capability frontier that no longer exists and that has shifted unevenly across the occupational distribution. Treating exposure as time-varying and exploiting the within-occupation variation that repeated annotation would generate offers a path toward more credible identification.

Our analysis has several limitations. Our replication covers three 2026 models, and we compare these with the original GPT-4 annotations from Eloundou et al. (2024). The GPT-4 to ChatGPT-5 comparison provides within-provider temporal evidence: mean E1 more than doubled across model generations within OpenAI alone. A richer panel tracking exposure scores

across every major model release since 2023 would characterize the rate and direction of temporal drift more precisely. The adoption channel test uses both industry-level proxies and occupation-level Anthropic usage data, but the Anthropic measure captures only Claude conversations and may not represent adoption patterns for other platforms. Our outcome is binary employment; wage and hours responses may show different sensitivity. These caveats do not weaken the central result. Until the field develops exposure measures whose properties are stable across rating instruments, or until multi-annotator sensitivity becomes standard practice, published estimates of how LLMs affect the labor market should be interpreted as conditional on the model that produced them.

These measurement concerns extend beyond the academic literature. Occupational exposure scores are increasingly cited in policy documents, workforce board planning, and congressional testimony as evidence for the scale and distribution of AI-driven labor-market disruption. Many organizations, such as the Bureau of Labor Statistics, OECD, and others incorporate AI exposure assessments into their estimates and projections. If the scores are sensitive to annotator choice in the ways we document, the policy conclusions drawn from them are similarly fragile.

References

Acemoglu, Daron, and Pascual Restrepo. 2022. "Tasks, Automation, and the Rise in U.S. Wage Inequality." *Econometrica* 90 (5): 1973-2016.

Appel, Ruth, Maxim Massenkoff, Peter McCrory, et al. 2026. "Anthropic Economic Index Report: Economic Primitives." Anthropic Research.

Acemoglu, Daron. 2025. "The Simple Macroeconomics of AI." Economic Policy 40 (121): 13-58.

Appel, Ruth, et al. 2026. "Anthropic Economic Index Report: Labor Market Impacts." Anthropic Research.

Argyle, Lisa P., et al. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31 (3): 337-351.

Autor, David, Frank Levy, and Richard Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118 (4): 1279-1333.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics*, Vol. 5, edited by James J. Heckman and Edward Leamer, 3705-3843. Amsterdam: Elsevier.

Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock. 2018. "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?" *AEA Papers and Proceedings* 108: 43-47.

Brynjolfsson, Erik, Bharat Chandar, and Ruyu Chen. 2025. "Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence." Working Paper.

Colato, Javier, Lindsey Ice, and Sofia Laycock. 2024. "Industry and Occupational Employment Projections Overview and Highlights, 2023-33." Monthly Labor Review, November.

Eisfeldt, Andrea L., Gregor Schubert, Miao Ben Zhang, and Bledi Taska. 2023. "Generative AI and Firm Values." NBER Working Paper 31222.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. "GPTs Are GPTs: Labor Market Impact Potential of LLMs." *Science* 384 (6702): 1306-1308.

Felten, Edward W., Manav Raj, and Robert Seamans. 2018. "A Method to Link Advances in Artificial Intelligence to Occupational Abilities." *AEA Papers and Proceedings* 108: 54-57.

Felten, Edward W., Manav Raj, and Robert Seamans. 2023. "Occupational Heterogeneity in Exposure to Generative AI." SSRN Working Paper.

Frey, Carl Benedikt, and Michael A. Osborne. 2017. "The Future of Employment: How Susceptible Are Jobs to Computerisation?" *Technological Forecasting and Social Change* 114: 254-280.

Gilardi, Fabrizio, Meysam Alizadeh, and Mael Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120 (30): e2305016120.

Gmyrek, Pawel, Janine Berg, and David Bescond. 2023. "Generative AI and Jobs: A Global Analysis of Potential Effects on Job Quantity and Quality." ILO Working Paper 96.

Green, Anne. 2024. "Artificial Intelligence and the Changing Demand for Skills in the Labour Market." OECD Artificial Intelligence Papers, No. 14. https://doi.org/10.1787/88684e36-en

Handa, Kunal, et al. 2025. "Which Economic Tasks Are Performed with AI? Evidence from Millions of Claude Conversations." Anthropic Research.

Johnston, Andrew, and Christos Makridis. 2025. "The Labor Market Effects of Generative AI: A Difference-in-Differences Analysis." SSRN Working Paper 5375017.

Hampole, Menaka, Dimitris Papanikolaou, Lawrence D.W. Schmidt, and Bryan Seegmiller. 2025. "Artificial Intelligence and the Labor Market." NBER Working Paper 33509.

Hui, Xiang, Oren Reshef, and Luofeng Zhou. 2024. "The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market." Organization Science 35 (6): 1977-1989.

Humlum, Anders, and Emilie Vestergaard. 2025. "Large Language Models, Small Labor Market Effects." NBER Working Paper 33777.

Machovec, Christine, Michael J. Rieley, and Emily Rolen. 2025. "Incorporating AI Impacts in BLS Employment Projections: Occupational Case Studies." Monthly Labor Review, February. https://doi.org/10.21916/mlr.2025.1

Manski, Charles F. 2003. *Partial Identification of Probability Distributions.* New York: Springer.

Massenkoff, Maxim, and Peter McCrory. 2026. "Labor Market Impacts of AI: A New Measure and Early Evidence." Anthropic Research.

McHugh, Mary L. 2012. "Interrater Reliability: The Kappa Statistic." *Biochemia Medica* 22 (3): 276-282.

Rock, Daniel, et al. 2025. "Extending 'GPTs Are GPTs' to Firms." *AEA Papers and Proceedings* 115: 51-55.

Webb, Michael. 2020. "The Impact of Artificial Intelligence on the Labor Market." SSRN Working Paper.

World Economic Forum. 2025. The Future of Jobs Report 2025. Geneva: World Economic Forum.

Xu, Xingcheng, Yong Tan, Songqin Ke, and Jianguo Wei. 2025. "Large Language Models at Work in China's Labor Market." *China Economic Review*.

Tables and Figures

Table 1. Inter-rater agreement among frontier LLMs

Model pair	Agree %	Cohen's Kappa	Primary disagreement locus
ChatGPT-5 vs. Gemini 2.5	73.0	0.56	E1/E2 boundary
ChatGPT-5 vs. Claude 4.5	70.0	0.51	E0/E1 and E1/E2 boundaries
Gemini 2.5 vs. Claude 4.5	56.9	0.36	E0/E1 boundary

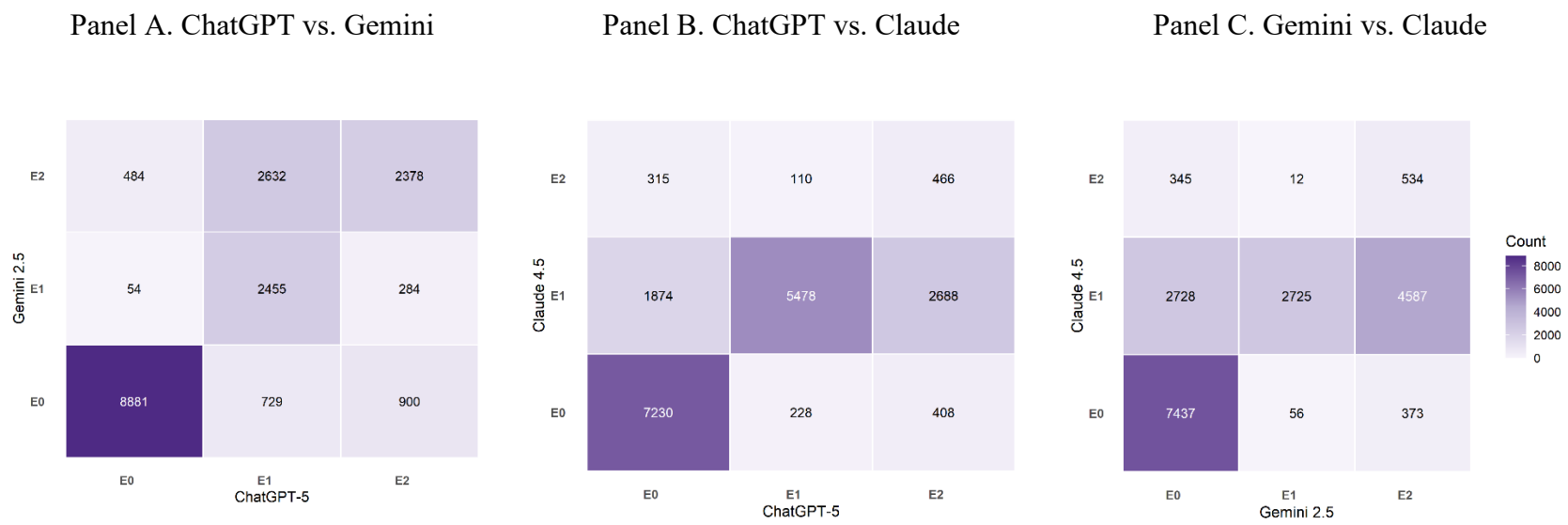
*Notes: N = 18,797 O*NET tasks per model pair. Three 2026 frontier language models, ChatGPT-5 (OpenAI), Gemini 2.5 (Google DeepMind), and Claude 4.5 (Anthropic), each classified 18,797 O*NET task descriptions using the Eloundou et al. (2024) rubric into E0 (not exposed: the model alone could not reduce task completion time by at least 50 percent), E1 (directly exposed: the model alone could achieve this speedup), or E2 (exposed with complementary software). This table reports the percentage of tasks receiving identical classifications from each model pair and Cohen's kappa measuring agreement beyond chance. "Agree" reports the share of tasks receiving identical E0/E1/E2 classifications from both annotators. Cohen's kappa is computed from the full 3x3 cross-tabulation. Kappa values range from 0.36 (fair) to 0.56 (moderate), following McHugh (2012) benchmarks. The primary disagreement column identifies which classification boundary accounts for the most disagreement in each pair.*

Table 2. LLM exposure and employment: difference-in-differences by annotator

	ChatGPT-4 (original)	ChatGPT-5	Gemini 2.5	Claude 4.5
Panel A. Individual-level				
E1	-.0066***	-.0095***	-.016***	-.0097***
	(.0018)	(.0016)	(.0028)	(.0011)
Observations	9,596,366	9,596,366	9,596,366	9,596,366
R-squared	.079	.079	.079	.079
Panel B. County-level				
E1	-.0011	-.0028	.0032	-.0077**
	(.0057)	(.0048)	(.0089)	(.0032)
Observations	4,104,098	4,104,098	4,104,098	4,104,098
R-squared	.1049	.1049	.1049	.1049
Panel C. Exposure summary				
Mean E1	0.14	0.30	0.14	0.51
Max/Min ratio	3.6			
Coefficient range (individual)	2.4-fold			

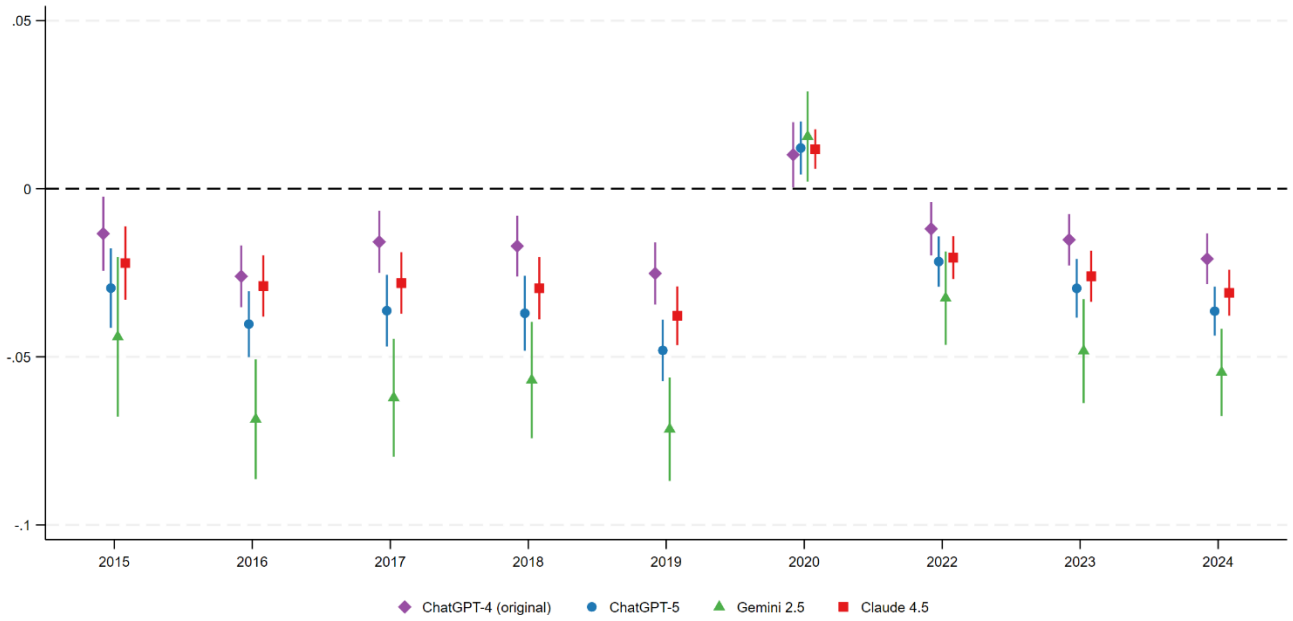
*Notes: Panel A reports individual-level difference-in-differences estimates of the effect of occupational E1 exposure on binary employment status using Current Population Survey data (N = 9,596,366 person-years, 2015-2024, ages 16-64). Panel B reports county-level estimates (N = 4,104,098 county-year-occupation cells). Dependent variable: binary employment indicator. Treatment: occupational E1 exposure (share of O*NET tasks classified as directly exposed to the language model by the indicated annotator) interacted with a post-2022 indicator. E1 tasks are those where the model alone could reduce completion time by at least 50 percent. Each column uses a different annotator (rating model) to construct the treatment variable: the original GPT-4 from Eloundou et al. (2024) or one of three 2026 frontier models. Controls: county, year, and occupation group fixed effects, state-specific linear trends, individual demographics (age, gender, race/ethnicity, education, marital status, number of children), and state-level COVID-19 cases and deaths. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Panel C reports the mean composite exposure across occupations and the ratio of the highest to lowest annotator mean.*

Figure 1. Cross-Tabulation of Task-level Classifications: pairwise task-level disagreement among frontier LLMs (heat map version)



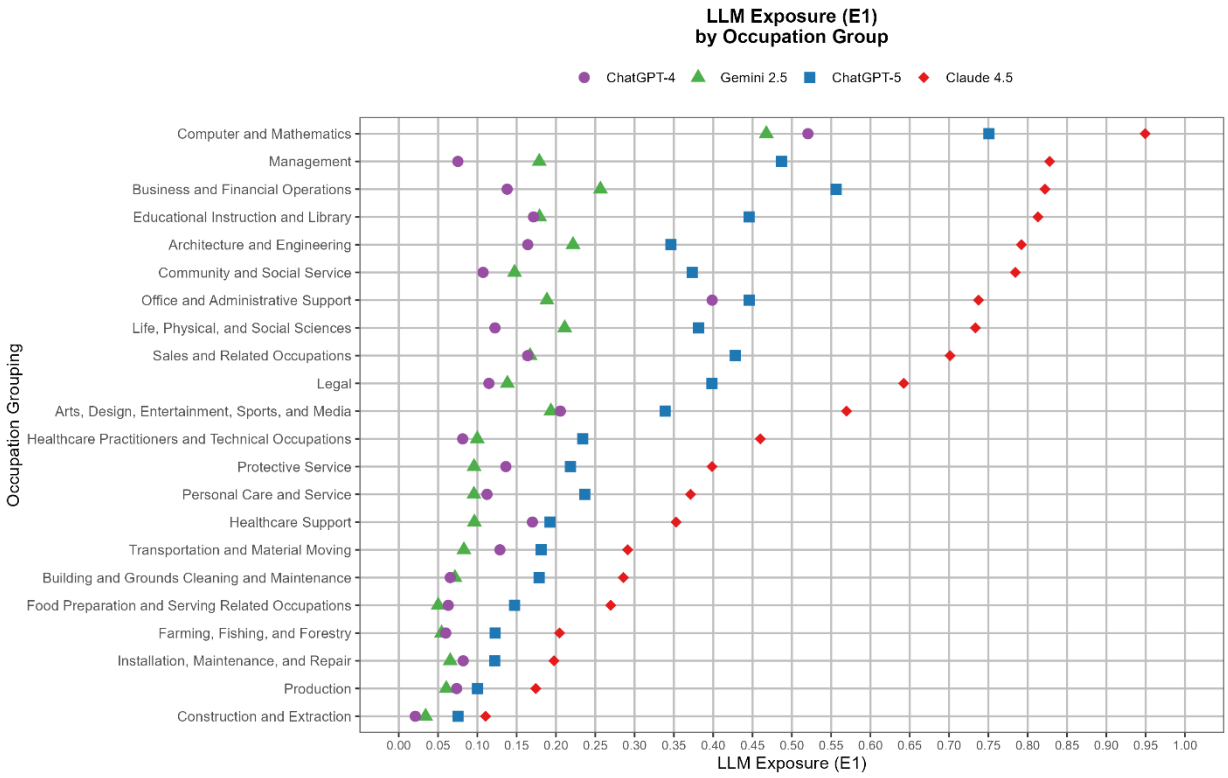
*Notes: Each panel displays a 3×3 heat map of task-level E0/E1/E2 classifications for a pair of annotator models applied to 18,797 O*NET task descriptions using the Eloundou et al. (2024) rubric. Rows correspond to classifications by the first annotator; columns correspond to classifications by the second. Diagonal cells indicate agreement; off-diagonal cells indicate disagreement. Panel A: ChatGPT-5 vs. Gemini 2.5. Panel B: ChatGPT-5 vs. Claude 4.5. Panel C: Gemini 2.5 vs. Claude 4.5.*

Figure 2. Event study: individual-level employment effects by annotator model



Notes: The figure plots coefficients on year-by-exposure interactions from equation (5), with 2021 as the omitted base year. All specifications include county fixed effects, year fixed effects, occupation group fixed effects, state-specific linear trends, individual characteristics (age, gender, race/ethnicity, education, marital status, number of children), and state-level COVID-19 cases and deaths. Capped vertical bars denote 95% confidence intervals based on standard errors clustered at the state level.

Figure 3. Mean E1 exposure by SOC major group and annotator



*Notes: The figure displays mean E1 (direct exposure) scores by Standard Occupational Classification (SOC) major group, separately by annotator model. E1 exposure is the share of O*NET tasks within each occupation classified as directly exposed by the indicated language model using the Eloundou et al. (2024) rubric. Occupational scores are averaged within SOC major groups using O*NET employment weights. The four annotators are ChatGPT-4 (original, from Eloundou et al., 2024), ChatGPT-5 (OpenAI), Gemini 2.5 (Google DeepMind), and Claude 4.5 (Anthropic).*

Online Appendix A

Table A1: LLM Automation Exposure Gains and Observed Usage

	Anthropic Usage	BTOS Usage
	(1)	(2)
Change in direct LLM exposure score (E1)	.335**	.009***
	(.154)	(.003)
Observations	705	710
R-Squared	0.252	.084

*Notes. The dependent variable is the change in LLM exposure score, defined as the mean E1 score across three frontier models (ChatGPT-5, Gemini 2.5, Claude 4.5) minus the GPT-4 baseline score from Eloundou et al. (2024). Column 1: Anthropic Usage is the occupation-level share of Claude conversations classified as work-related using the Clio automated classification tool (Anthropic Economic Index, March 27, 2025 release). Column 2: BTOS Usage is the share of workers in the occupation reporting use of AI or machine learning tools in the past week (U.S. Census Bureau Business Trends and Outlook Survey). All specifications control for the original GPT-4 exposure rating. OLS estimates with heteroskedasticity-robust standard errors in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.1. N = 705 (Column 1) and 710 (Column 2) occupations.*

Table A2: Within-Model Stability Test

Model	Pair	Within-Model Agreement (%)	Cohen's Kappa
Claude 4.5	Run 1 vs 2	0.990	0.982
Claude 4.5	Run 2 vs 3	0.992	0.986
Claude 4.5	Run 1 vs 3	0.989	0.980
ChatGPT-5	Run 1 vs 2	0.967	0.934
ChatGPT-5	Run 2 vs 3	0.968	0.936
ChatGPT-5	Run 1 vs 3	0.966	0.932
Gemini 2.5	Run 1 vs 2	0.910	0.851
Gemini 2.5	Run 2 vs 3	0.914	0.857
Gemini 2.5	Run 1 vs 3	0.901	0.835

Notes. A single prompt was repeated three times per model (ChatGPT-5, Gemini 2.5, Claude 4.5) on a 10% random subsample of tasks at temperature zero. Within-model agreement (%) and Cohen's kappa were calculated to measure consistency and account for chance agreement.

Table A3: Composite Exposure (E1 + E2) and Employment: Difference-in-Differences by Annotator

	ChatGPT-4 (original)	ChatGPT-5	Gemini 2.5	Claude 4.5
Panel A. Individual-level				
E1+E2	-.0102***	-.0103***	-.0154***	-.01***
	(.0012)	(.0014)	(.0018)	(.0011)
Observations	9,596,366	9,596,366	9,596,366	9,596,366
R-squared	.079	.079	.079	.079
Panel B. County-level				
E1+E2	-.0013	.0009	.0008	-.0019
	(.003)	(.0045)	(.0056)	(.0032)
Observations	4,104,098	4,104,098	4,104,098	4,104,098
R-squared	.015	.013	.013	.016
Panel C. Exposure summary				
Mean E1+E2	0.55	0.40	0.29	0.54
Max/Min ratio	1.9			
Coefficient range (individual)	1.5-fold			

*Notes: This table replicates the Table 2 specification using composite exposure (E1 + E2) as the treatment variable instead of E1 alone. Occupational composite exposure is the share of tasks classified as either E1 (direct LLM assistance) and E2 (assistance via LLM-powered software) by the indicated annotator. Working age (16-64) population sample. All specifications include county fixed effects, year fixed effects, occupation group fixed effects, state-specific linear trends, individual characteristics (age, gender, race/ethnicity, education, marital status, number of children), and state-level COVID-19 cases and deaths. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Sample period: 2015-2024. Panel C reports the mean composite exposure across occupations and the ratio of the highest to lowest annotator mean, for comparison with the 3.6-fold E1 ratio reported in Table 2.*

Table A4: Associations Between Frontier Model Exposure Scores and Anthropic Usage

	ChatGPT-4 (original)	ChatGPT-5	Gemini 2.5	Claude 4.5
Panel A. Direct Exposure E1				
Spearman rho	0.298	0.425	0.406	0.419
Bootstrap std. errors	0.035	0.033	0.033	0.032
p-value	<0.001	<0.001	<0.001	<0.001
N	705	705	705	705
Panel B. Composite Exposure E1 + E2				
Spearman rho	0.335	0.372	0.315	0.364
Bootstrap std. errors	0.035	0.033	0.035	0.034
p-value	<0.001	<0.001	<0.001	<0.001
N	705	705	705	705

Notes: This table reports occupation-level Spearman rank correlations between LLM exposure measures and Anthropic usage based on the March 27, 2025 release. Panel A shows correlations using direct exposure (E1), and Panel B uses composite exposure (E1 + E2). Each column corresponds to a different LLM measure: ChatGPT-4 (original, from Eloundou et al., 2024), ChatGPT-5, Gemini 2.5, and Claude 4.5. p-values are based on Spearman rank correlation tests.

Table A5: OLS Regressions of Occupation-Level Anthropic Usage and LLM Exposure

	ChatGPT-4 (original)	ChatGPT-5	Gemini 2.5	Claude 4.5
Panel A. Direct Exposure E1				
E1 Exposure	0.119***	0.062***	0.118***	0.033***
	(0.010)	(0.008)	(0.013)	(0.006)
N	705	705	705	705
R-squared	0.169	0.076	0.104	0.037
Panel B. Composite Exposure E1 + E2				
E1+E2 Exposure	0.026***	0.035***	0.032***	0.026***
	(0.006)	(0.007)	(0.007)	(0.006)
N	705	705	705	705
R-squared	0.024	0.038	0.030	0.025

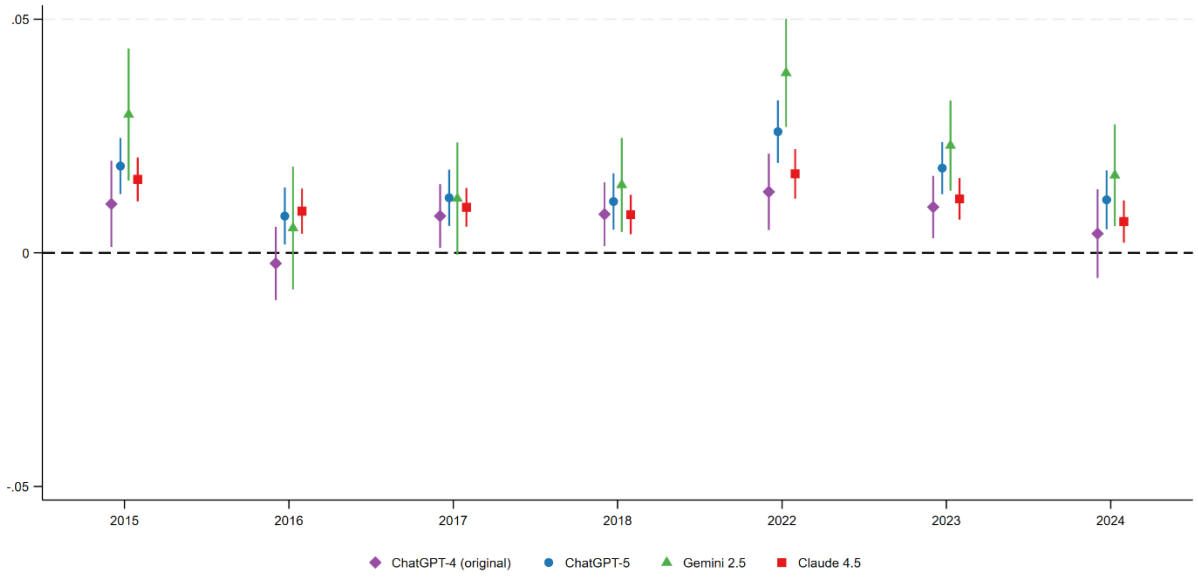
*Notes: This table reports OLS regression results relating LLM exposure measures to Anthropic usage based on the March 27, 2025 release. The dependent variable in all regressions is Anthropic usage. Panel A uses direct exposure (E1), while Panel B uses composite exposure (E1 + E2). Each column corresponds to a different LLM measure: ChatGPT-4 (original, from Eloundou et al., 2024), ChatGPT-5, Gemini 2.5, and Claude 4.5. Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1*

Table A6. LLM Exposure and Employment: Robustness with 2020-2021 exclusion

	ChatGPT-4 (original)	ChatGPT-5	Gemini 2.5	Claude 4.5
Panel A. Individual-level				
E1	.0013	.0025	.0028	-.0001
	(.0022)	(.002)	(.003)	(.0015)
Observations	7,734,472	7,734,472	7,734,472	7,734,472
R-squared	.0771	.0771	.0771	.0771
Panel B. County-level				
E1	.0024	.0035	.0159*	-.0024
	(.006)	(.0045)	(.0092)	(.0031)
Observations	3,155,042	3,155,042	3,155,042	3,155,042
R-squared	.1027	.1027	.1027	.1027
Panel C. Exposure summary				
Mean E1	0.14	0.30	0.14	0.51
Max/Min ratio	3.6			
Coefficient range (individual)	28-fold			

*Notes: Years 2020 and 2021 are excluded due to COVID-19 disruptions. All specifications include county fixed effects, year fixed effects, occupation group fixed effects, state-specific linear trends, individual characteristics (age, gender, race/ethnicity, education, marital status, number of children), and state-level COVID-19 cases and deaths. Robust standard errors in parentheses. Working age (16-64) population sample. Sample period: 2015-2024. Panel C reports the mean composite exposure across occupations and the ratio of the highest to lowest annotator mean. *** p<0.01, ** p<0.05, * p<0.1.*

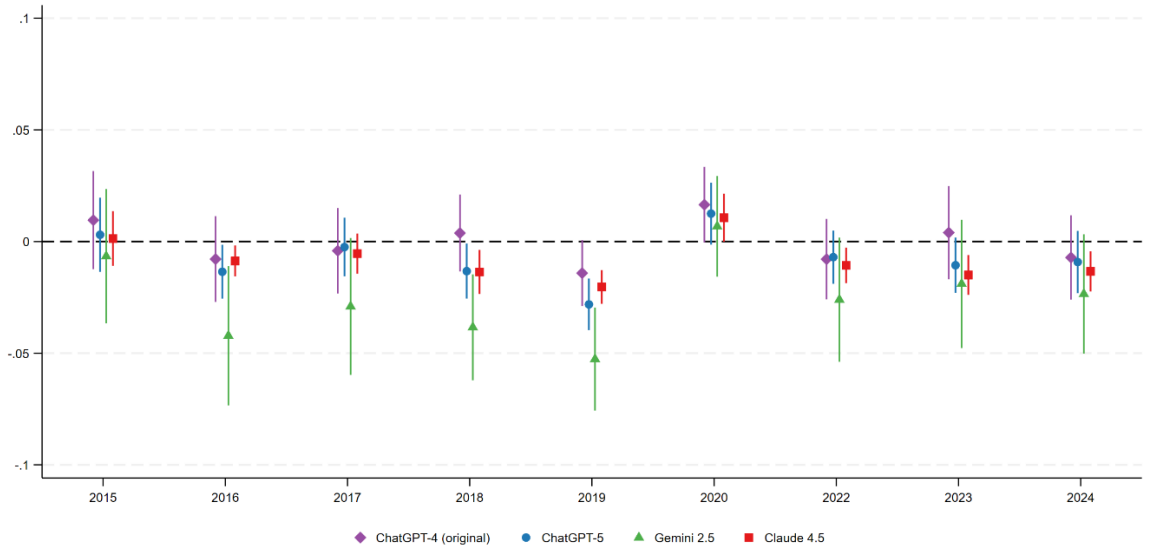
Figure A1: Individual-level Event Study - Robustness to Excluding 2020–2021, 2019 Normalization



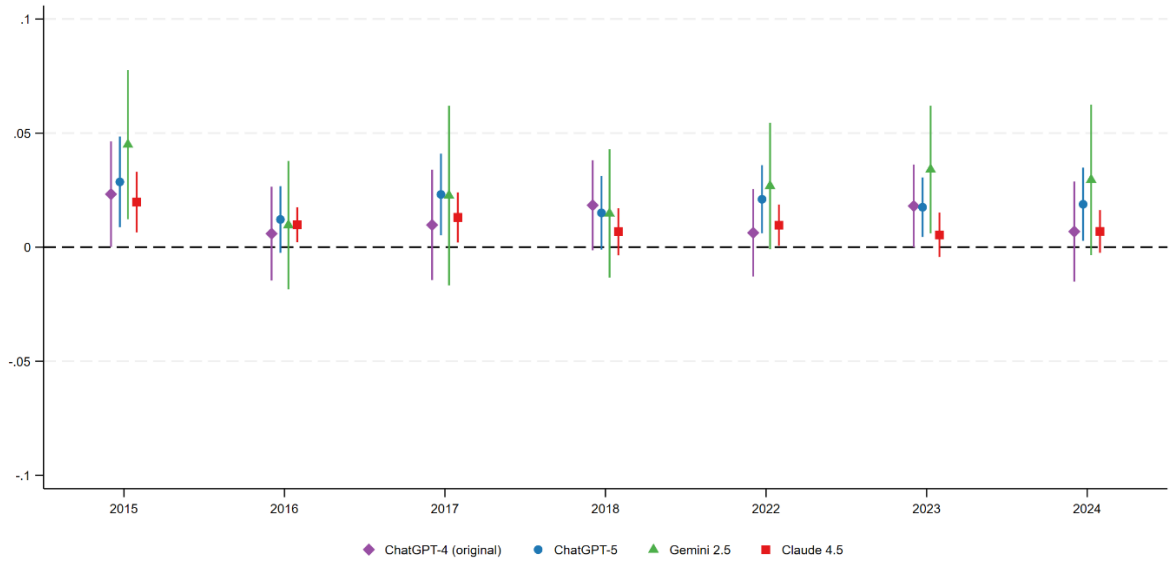
*Notes: The figure replicates Figure 2 excluding years 2020 and 2021, with 2019 as the omitted base year. Sample: CPS 2015–2019 and 2022–2024, ages 16–64. All specifications, controls, and fixed effects are identical to Figure 2. Capped vertical bars denote 95% confidence intervals based on standard errors clustered at the state level. *** p<0.01, ** p<0.05, * p<0.1.*

Figure A2: County-level Event Study

Panel A. All Years, 2021 Normalization



Panel B. Excluding 2020–2021, 2019 Normalization



Notes: The figure plots coefficients on year-by-exposure interactions from the county-level specification (equation 5 applied to county-year-occupation cells). Each series corresponds to a different annotator model. Sample: county-year-occupation cells, 2015–2024. Dependent variable: log employment count. Controls and fixed effects are identical to Table 2, Panel B. Vertical bars denote 95% confidence intervals based on standard errors clustered at the state level.

Data Appendix B

Overall Exposure Rubric Summary
Following Eloundou et al. (2024), our prompt is as follows for each type of occupation:

Consider the most powerful **Gemini/OpenAI/Claude** large language model (LLM). This model can complete many tasks that can be formulated as having text input and text output.

Assume you are a worker with an average level of expertise in your role trying to complete the given task. You have access to the LLM as well as any other existing software or computer hardware tools mentioned in the task. You also have access to any commonly available technical tools accessible via a laptop (e.g., a microphone, speakers, etc.). You do not have access to any other physical tools or materials.

Please label the given task according to the taxonomy below.

E0 – No exposure

Label tasks E0 if direct access to the LLM through an interface like **Gemini/OpenAI/Claude** cannot reduce the time it takes to complete this task with equivalent quality by half or more.

If a task requires a high degree of human interaction (for example, in-person demonstrations), then it should be classified as E0.

E1 – Direct exposure

Label tasks E1 if direct access to the LLM through an interface like **Gemini/OpenAI/Claude** can reduce the time it takes to complete the task with equivalent quality by at least half. This includes tasks that can be reduced to:

- Writing and transforming text and code according to complex instructions
- Providing edits to existing text or code following specifications
- Writing code that can help perform a task that used to be done by hand
- Translating text between languages
- Summarizing medium-length documents
- Providing feedback on documents
- Answering questions about a document
- Generating questions a user might want to ask about a document

E2 – Exposure by LLM-powered applications

Label tasks E2 if having access to the LLM alone may not reduce the time it takes to complete the task by at least half, but it is easy to imagine additional software that could be developed on top of the LLM that would reduce the time it takes to complete the task by half. This software may include capabilities such as:

- Summarizing documents longer than 2000 words and answering questions about those documents
- Retrieving up-to-date facts from the Internet and using those facts in combination with the LLM capabilities
- Searching over an organization's existing knowledge, data, or documents and retrieving information

Examples of software built on top of the LLM that may help complete worker activities include:

- Software built for a home goods company that quickly processes and summarizes their up-to-date internal data in customized ways to inform product or marketing decisions
- Software able to suggest live responses for customer service agents speaking to customers in their company's customer service interface
- Software built for legal purposes that can quickly aggregate and summarize all previous cases in a particular legal area and write legal research memos tailored to the law firm's needs
- Software specifically designed for teachers that allows them to input a grading rubric and upload the text files of all student essays and have the software output a letter grade for each essay
- Software that retrieves up-to-date facts from the internet and uses the capabilities of the LLM to output news summaries in different languages

E3 – Exposure given image capabilities

Suppose you had access to both the LLM and a system that could view, caption, and create images. This system cannot take video media as inputs. This system cannot accurately retrieve very detailed information from image inputs, such as measurements of dimensions within an image. Label tasks as E3 if there is a significant reduction in the time it takes to complete the task given access to a LLM and these image capabilities:

- Reading text from PDFs
- Scanning images
- Creating or editing digital images according to instructions

Annotation examples:

- **Occupation:** Inspectors, Testers, Sorters, Samplers, and Weighers
  Task: Adjust, clean, or repair products or processing equipment to correct defects found during inspections.
  Label (E0 / E1 / E2 / E3): E0
  Explanation: The model does not have access to any kind of physicality, and more than half of the task (adjusting, cleaning, and repairing equipment) described requires hands or other embodiment.
- **Occupation:** Computer and Information Research Scientists
  Task: Apply theoretical expertise and innovation to create or apply new technology, such as adapting principles for applying computers to new uses.

Label (E0 / E1 / E2 / E3): E1
Explanation: The model can learn theoretical expertise during training as part of its general knowledge base, and the principles to adapt can be captured in the text input to the model.

- **Activity:** Scheduled dining reservations
  Label (E0 / E1 / E2 / E3): E2
  Explanation: Automation technology already exists for this (e.g., Resy) and it's unclear what an LLM offers on top of using that technology (no-diff). That said, you could build something that allows you to ask the LLM to make a reservation on Resy for you.
- **Activity:** Negotiate purchases or contracts
  Label (E0 / E1 / E2 / E3): E2
  Explanation: You could have each party transcribe their point of view and then feed this to an LLM to resolve any disputes (E3). That said, many people would need to buy into using new technological tools to accomplish this (system).
- **Occupation:** Allergists and Immunologists
  Task: Prescribe medication such as antihistamines, antibiotics, and nasal, oral, topical, or inhaled glucocorticosteroids.
  Label (E0 / E1 / E2 / E3): E2
  Explanation: The model can provide guesses for different diagnoses and write prescriptions and case notes. However, it still requires a human in the loop using their judgment and knowledge to make the final decision.

Now, apply the above rubric to the example below:

Appendix C: Derivation of the OLS Bias

This appendix derives equation (4) in the main text. All variables are assumed demeaned after partialling out controls $\mathbf{X}_i$.

Setup.

The true structural relationship is:

(A1) $y_i = \beta\theta_i + \pi a_i + u_i$

where u_i is orthogonal to (θ_i, a_i). The researcher observes:

(A2) $\theta_i^m = \theta_i + \varepsilon_i^m = \theta_i + \delta^m\theta_i + \varphi^m a_i + v_i^m = (1 + \delta^m)\theta_i + \varphi^m a_i + v_i^m$

where $v_i^m \sim \text{iid}(0, \sigma^2_{v,m})$ is a classical error term, independent of (θ_i, a_i, u_i).

OLS probability limit.

The naive OLS estimator regresses y_i on θ_i^m, omitting a_i. The probability limit is:

(A3) $\text{plim}\ \beta^m = \text{Cov}(\theta_i^m, y_i) / \text{Var}(\theta_i^m)$

Numerator.

$\text{Cov}(\theta_i^m, y_i) = \text{Cov}((1+\delta^m)\theta_i + \varphi^m a_i + v_i^m,\ \beta\theta_i + \pi a_i + u_i)$

$= \beta(1+\delta^m)\sigma^2_\theta + \pi(1+\delta^m)\sigma_{\theta a} + \beta\varphi^m\sigma_{\theta a} + \pi\varphi^m\sigma^2_a$

$= \beta[(1+\delta^m)\sigma^2_\theta + \varphi^m\sigma_{\theta a}] + \pi[\varphi^m\sigma^2_a + (1+\delta^m)\sigma_{\theta a}]$

using the independence of v_i^m from all other terms and orthogonality of u_i.

Denominator.

$$\text{Var}(\theta_i^{*m}) = (1+\delta^m)^2\sigma^2_\theta + (\varphi^m)^2\sigma^2_a + 2(1+\delta^m)\varphi^m\sigma_{\theta a} + \sigma^2_{v,m}$$

Decomposition.

Combining numerator and denominator and defining $D^m \equiv \text{Var}(\theta_i^{*m})$:

(A4) $\quad \text{plim}\ \beta^{*m} = \beta\lambda^m + \pi\Omega^m$

where:

(A5) $\quad \lambda^m = [(1+\delta^m)\sigma^2_\theta + \varphi^m\sigma_{\theta a}] / D^m$ $\quad$ (Calibration Reliability)

(A6) $\quad \Omega^m = [\varphi^m\sigma^2_a + (1+\delta^m)\sigma_{\theta a}] / D^m$ $\quad$ (Adoption Confounding)

Properties.

Calibration reliability. In the absence of adoption feedback ($\varphi^m = 0$ and $\sigma_{\theta a} = 0$), the reliability simplifies to $\lambda^m = (1+\delta^m)\sigma^2_\theta / [(1+\delta^m)^2\sigma^2_\theta + \sigma^2_{v,m}] = 1/[(1+\delta^m) + \sigma^2_{v,m}/((1+\delta^m)\sigma^2_\theta)]$. When $\delta^m > 0$ (generous rater), the first term in the denominator increases, reducing λ^m below the classical attenuation factor. When $\delta^m < 0$ (conservative rater), the denominator decreases, and λ^m can exceed the classical level, producing weaker attenuation or amplification.

Adoption confounding. When $\varphi^m > 0$, the second term $\pi\Omega^m$ contributes bias whose sign depends on π. If early adoption raises productivity and wages ($\pi > 0$), the confounding is positive, inflating the estimated displacement effect. If adoption reduces employment in the medium run ($\pi < 0$), the confounding is negative. The magnitude of Ω^m increases with φ^m, which we expect to be larger for more recent models trained on more extensive user data.

Total bias is annotator-dependent. Because different annotators occupy different points in the (δ^m, φ^m) parameter space, the total bias $\beta(\lambda^m - 1) + \pi\Omega^m$ varies across annotators in both magnitude and potentially in sign. This is the formal basis for the non-monotonic pattern of

coefficient estimates across annotators observed in the main text (Table 2): the downstream conclusion about AI's employment effect depends on the combined annotator's position (δ^m, φ^m).

Special cases.

1. *Classical error*: $\delta^m = \varphi^m = 0$. Then $\lambda^m = \sigma^2_\theta / (\sigma^2_\theta + \sigma^2_{v,m}) < 1$ and $\Omega^m = 0$.

2. *Calibration only* ($\varphi^m = 0$, $\sigma_{\theta a} = 0$): $\lambda^m = 1/[(1+\delta^m) + \sigma^2_{v,m}/((1+\delta^m)\sigma^2_\theta)]$. Generous raters ($\delta^m > 0$) attenuate more than the classical case. Conservative raters ($\delta^m < 0$) attenuate less.

3. *Adoption only* ($\delta^m = 0$): λ^m and Ω^m both depend on φ^m and the covariance $\sigma_{\theta a}$. The researcher's estimate conflates the true exposure effect with the adoption effect, even if the rating model is perfectly calibrated on average.