Vacancies Under the Algorithm: A Methodology for Generative AI and the Demand for Knowledge Work

Kavya Ramanujan[*]*, Femi Adebayo, Ingrid Brouwer*

Center for AI and Knowledge Work (CAIKW)

Generative Economic Review • May 17, 2026

GER 1.7

JEL Classification: J21, J23, J24, O33, J63, C81, C18

Keywords: generative artificial intelligence, labor demand, knowledge work, online job postings, occupational exposure, research design, methodology, task-based framework, difference-in-differences, pre-registration

Abstract

We propose a methodology for using online job posting data to measure how generative artificial intelligence is restructuring US knowledge work, and we articulate the patterns that distinct theoretical hypotheses predict empirical implementations of this methodology would reveal. The design constructs an occupational exposure measure from a task-level mapping between O*NET task descriptions and current AI capability, classifies postings by exposure quintile, and applies a difference-in-differences specification with the November 2022 release of large language models as the focal event. Three outcome margins receive specific attention: the volume of postings within and across exposure categories; the within-occupation composition of skill requirements; and the within-occupation distribution of posted wages. For each margin, we develop a measurement procedure, specify the estimating equation, and articulate the empirical patterns that distinct hypotheses—substitution, complementarity, restructuring, reorganization, and null effect—would predict. We embed the design within an explicit task-based model that derives testable implications from microfoundations, discuss the construct validity of the exposure measure under shared-prior bias among raters, provide statistical power calculations anchored to published within-occupation residual variances, develop a ten-item pre-registration protocol that disciplines specification search and commits to multiple-testing correction and code-deposit, demonstrate the procedure end-to-end with a synthetic

worked example under each hypothesis, and bound the residual identification limits arising from joint macroeconomic shocks contemporaneous with the focal event. We do not present empirical estimates from a particular dataset, and we are explicit that the design recovers a cross-sectional differential rather than a clean causal effect. The paper specifies the methodology that an empiricist with access to comprehensive job posting data could implement, and articulates the interpretations that empirical results would support. By separating design specification from implementation, the paper aims to support a more disciplined empirical literature on a question whose answer carries substantial implications for the measurement of human capital, the design of education policy, and the projection of long-run labor market outcomes.

1. Introduction

Whether and how generative artificial intelligence will reshape labor markets is among the most consequential open questions in contemporary economics. Public debate oscillates between the prediction that large language models will eliminate vast swaths of white-collar employment and the prediction that, like prior general-purpose technologies, they will create as many jobs as they destroy over a horizon long enough to matter (Acemoglu and Restrepo, 2019; Bresnahan and Trajtenberg, 1995). The empirical record is, as of the present writing, too short and too noisy to settle this debate definitively, but the methodological infrastructure for studying it can be specified now and applied as the empirical record matures.

This paper specifies a research design for studying the labor demand response to generative AI using online job posting data. We do not implement the design with a particular dataset. We specify the data construction, the measurement procedure, the empirical specifications, and the predicted patterns under competing theoretical accounts. Our purpose is to provide a disciplined methodological foundation for the rapidly emerging empirical literature on generative AI and labor demand.

1.1 Why a methodology paper now

The cost of a methodology paper that precedes empirical implementation is that it is, by construction, descriptive. The benefits, in our view, outweigh this cost in the present case for three reasons.

First, the empirical literature on generative AI and labor demand is expanding faster than the literature's methodological foundations are stabilizing. Papers using different exposure measures, different sample frames, different definitions of "exposure," and different difference-in-differences specifications are now appearing in working-paper form at a high rate. A pre-implementation methodology paper allows the literature to coordinate on a common research design, much as the literature on labor market spillovers coordinated on

the Bartik-style instrument (Bartik, 1991; Goldsmith-Pinkham et al., 2020) or the literature on minimum wages coordinated on the border-discontinuity design (Dube et al., 2010).

Second, the focal event—the November 2022 release of large language models, in particular ChatGPT—is a single common shock affecting the entire US labor market. Clean identification through standard quasi-experimental methods is therefore impossible at the aggregate level; identification can only be sought at the cross-section of differential exposure across occupations. The interpretive consequences of this design choice are subtle, and articulating them in advance helps the literature avoid mistaking weak identification for clean causality.

Third, the policy stakes are high. Decisions about workforce education investments, labor market regulation, and AI governance will be shaped by the empirical literature that emerges over the next several years. A literature whose methodological foundations are well-articulated is, in expectation, a literature that produces more useful policy advice. This is the disciplining function of methodology specifications more broadly: they are not substitutes for empirical work but scaffolds that improve its quality.

1.2 Four contributions

The paper makes four substantive contributions to the methodology of labor-market measurement under generative AI.

First, we embed the proposed research design within an explicit task-based model that derives testable implications from first principles. The model is a deliberate simplification of Acemoglu and Autor (2011) extended to accommodate a generative-AI shock to the technological frontier. The derivation produces signed predictions on each of the three outcome margins (posting volume, skill composition, posted wages) under each of five interpretive hypotheses, and it produces conditions under which competing hypotheses can be distinguished empirically. The model is presented in Section 3.

Second, we examine the construct validity of the occupational exposure measure inherited from Eloundou et al. (2024). The measure rests on a forecasting judgment by human raters whose priors about AI capability are themselves shaped by industry narratives. We specify three triangulation steps—convergent validation against alternative exposure constructions, behavioral validation against firm-level AI-adoption surveys, and time-stamped re-rating—that bound the shared-bias component that inter-rater agreement statistics cannot detect. The discussion is in Section 3.3.

Third, we provide statistical power calculations under plausible effect sizes for the difference-in-differences specifications, anchoring the parameterization to published within-occupation residual variance estimates (Hershbein and Kahn, 2018; Acemoglu et al., 2022). The calculations document the sample sizes and observation horizons required to distinguish the substitution and complementarity hypotheses with reasonable power, and they identify the margins along which the most precision is available—and the margins on which the

diagnostic claim has previously been over-stated. The calculations are presented in Section 4.

Fourth, we develop a ten-item pre-registration protocol that disciplines specification search. The protocol specifies the dataset construction, the principal specification, the falsification battery, the Rambachan-Roth (2023) sensitivity analysis for pre-trends, the pre-specified heterogeneity analyses, the multiple-testing correction across margins, the reporting commitment, and the code-deposit commitment at a public registry. Implementations that adhere to the protocol generate headline coefficients with interpretations robust to the file-drawer problem (Rosenthal and Rubin, 1979; Gerber and Malhotra, 2008). The protocol—together with a synthetic worked example demonstrating the procedure under each hypothesis—is presented at the head of Section 5.

1.3 Three components of the design

The proposed design has three components, summarized briefly here and developed in detail in Sections 3-5.

First, we develop a measure of occupational exposure to generative AI by mapping O*NET task descriptions to current AI capability, following the structure of Eloundou et al. (2024). The mapping yields a continuous exposure score per occupation that we then aggregate into quintiles for ease of interpretation. We provide a reliability protocol for the underlying classification step and a validation protocol that compares the resulting measure with alternative AI-exposure measures from the literature.

Second, we specify a difference-in-differences research design with the November 2022 release of large language models as the focal event, examining outcomes along three margins: posting volume, skill composition, and posted wages. The design uses two-way fixed effects with occupation and time, and we discuss the alternative event-study and continuous-treatment specifications under which the recent econometric literature has flagged identification concerns (de Chaisemartin and D'Haultfœuille, 2020; Callaway and Sant'Anna, 2021).

Third, we articulate five interpretive frameworks—substitution, complementarity, restructuring, reorganization, and null effect—and the empirical patterns that distinguish them. The five hypotheses are not exhaustive, but they span a useful taxonomy of the predictions that current theoretical and policy debates have produced. For each hypothesis, we state the signed prediction on each of the three margins and identify the pair of margins whose joint behavior is most informative for distinguishing among hypotheses.

1.4 Structure of the paper

The paper proceeds in six sections. Section 2 places the proposed methodology within the literatures on routine-biased technical change, the labor market effects of artificial intelligence, the use of job posting data to measure labor demand, the recent econometric literature

on difference-in-differences with single shock dates, and the small but growing literature on pre-registration in observational economics. Section 3 specifies the research design in detail, including the task-based model that motivates the empirical specifications, the occupational exposure measure with its reliability and construct-validity protocols, the three outcome variables with their measurement procedures, the regression specifications, and the pre-trend and falsification battery (including the Rambachan-Roth sensitivity analysis). Section 4 articulates the predicted patterns under five competing theoretical accounts, provides statistical power calculations anchored to published variance estimates, and characterizes the empirical regimes in which the proposed methodology can—and cannot—discriminate among hypotheses. Section 5 opens with the ten-item pre-registration protocol (5.1) and the worked synthetic example demonstrating the procedure end-to-end under each hypothesis (5.2), then turns to implementation considerations and limitations including the residual identification gap arising from joint macroeconomic shocks contemporaneous with the focal event. Section 6 concludes by inviting empirical implementation and identifying extensions.

We emphasize the descriptive nature of the proposed design. The release of large language models in November 2022 is a single common shock affecting the whole economy, and the post-event window coincided with the fastest US monetary-tightening cycle since the Volcker disinflation and with post-pandemic sectoral reallocation. Clean identification of the AI-specific effect through standard quasi-experimental methods is impossible in this setting. The design proposed here offers careful description of how outcomes have moved in occupations differently exposed to AI capability under an explicit identification footnote, leaving the substantive interpretive burden to the patterns the data reveal and to triangulation across complementary research designs.

—

2. Literature Review

Six literatures bear directly on the proposed methodology: the task-based framework for labor market analysis, automation and routine-biased technical change, the labor market effects of generative AI specifically, the empirical literature on job postings as a measure of labor demand, the recent econometric literature on difference-in-differences with a single shock date, and the methodological literature on pre-registration in observational economics. We treat each in turn.

2.1 The task-based framework

The task-based framework for labor market analysis, formalized by Acemoglu and Autor (2011) and traced to earlier work by Autor et al. (2003) and Levy and Murnane (2005), represents production as the performance of tasks rather than as the application of indivisible factor inputs. Tasks are heterogeneous in their automatability—their susceptibility to substitution by capital, software, or both—and workers are heterogeneous in their compara-

tive advantage at performing different tasks. Technological change shifts the boundary of tasks that capital can perform, redistributing workers across the remaining task space and reshaping the demand for skills.

The framework was originally motivated by the polarization of US wages during the 1980s and 1990s, which neither human-capital-augmented production functions nor classical skill-biased technical change models could readily explain. The task-based framework explained polarization as the result of computerization's relative ease of substituting for routine cognitive tasks—the kind of work middle-skill workers had performed—while leaving non-routine analytical (high-skill) and non-routine manual (low-skill) tasks largely unaffected.

For generative AI, the task-based framework remains the natural starting point but requires re-tuning. Where computerization substituted for routine tasks, generative AI plausibly substitutes for non-routine cognitive tasks—exactly the tasks the previous framework took to be the protected domain of high-skill labor (Autor, 2024; Acemoglu et al., 2022). If this characterization is correct, the demand effects of generative AI may show patterns that invert the patterns observed during computerization: a compression rather than a polarization of the wage distribution, contraction at the upper end of the skill distribution, and reorganization within occupations rather than across them.

Goos et al. (2014) extend the task-based framework to incorporate offshoring as a separate margin of competition for labor's tasks. Their typology of tasks—routine vs. non-routine, cognitive vs. manual, codifiable vs. tacit—is useful for thinking about which generative-AI capabilities map onto which task categories. The capabilities of current frontier language models substitute most readily for codifiable cognitive tasks regardless of whether they are routine or non-routine; the boundary that matters for generative AI is the codifiability boundary rather than the routineness boundary that organized the previous literature.

2.2 Automation and routine-biased technical change

The literature on automation and routine-biased technical change documents the labor-market effects of prior automation episodes. Autor et al. (2003) document that the diffusion of computers from the 1980s through the 1990s reduced demand for routine tasks—both manual and cognitive—and raised demand for non-routine analytical and interpersonal tasks. Acemoglu and Restrepo (2020) document substantial wage effects of industrial robots in US local labor markets, finding that each additional robot per thousand workers reduces local employment-to-population by approximately 0.2 percentage points and local wages by approximately 0.4 percent. Webb (2020) constructs an automation-exposure measure from patent texts and documents that occupations with higher patent-based exposure to AI specifically exhibited slower wage growth in the pre-generative-AI era.

A broader literature documents that prior episodes of automation have been characterized

by a long lag between capability deployment and labor market response (Goldin and Katz, 2010; Acemoglu and Restrepo, 2018). Electrification of US manufacturing required approximately forty years before its productivity gains became visible in aggregate output statistics, and the timing of its wage effects was even more delayed (David, 1990). The implication for generative AI is that any empirical effects detectable in the first two to three years after capability deployment are likely to be modest, even if the long-run effects are large. The proposed methodology is designed to detect early effects to the extent they exist, while acknowledging that null findings in the early data would not warrant strong conclusions about long-run effects.

The conceptual distinction between displacement effects (substitution of capital for labor in specific tasks) and reinstatement effects (creation of new tasks that complement or are complemented by the new technology) is central to the framework of Acemoglu and Restrepo (2019). Their accounting decomposes observed wage and employment changes into the displacement and reinstatement margins separately. The proposed methodology in this paper is designed to surface evidence on both margins: the posting-volume outcome captures the displacement margin, and the skill-composition outcome captures the reinstatement margin to the extent reinstatement manifests as new skill requirements within continuing occupations.

2.3 The labor market effects of generative AI

The rapidly growing body of work on the labor market effects of generative AI specifically constitutes the literature that the present methodology paper most directly extends. We summarize the principal contributions.

Eloundou et al. (2024) construct a mapping from ONET task descriptions to the capabilities of large language models and estimate that around 80 percent of US workers could see at least 10 percent of their tasks affected by generative AI. Their measure is the conceptual foundation of the exposure measure we propose, and our design preserves the structure of their classification rubric while specifying a reliability protocol that their original construction did not require for its purposes.

Brynjolfsson et al. (2025) report results from a randomized field experiment in customer support documenting that generative AI tools raise the productivity of less-skilled workers more than that of more-skilled workers, with the productivity effect concentrated in the bottom three deciles of the worker skill distribution. This finding is consistent with the restructuring hypothesis we develop in Section 4: generative AI complements judgment-intensive work at the high end of the distribution while substituting for routine cognitive tasks at the low end. Noy and Zhang (2023) document similar productivity gains in writing tasks, with a similar pattern of larger gains for less-skilled workers. Peng et al. (2023) document a substantial reduction in coding task completion times among software developers using AI pair-programming tools.

Acemoglu (2024) provides an aggregate framework that maps these microeconomic productivity gains to long-run growth implications, with conclusions more conservative than the most aggressive projections in industry reports. Acemoglu argues that the within-task productivity gains documented in microeconomic studies do not aggregate cleanly to economy-wide productivity because (a) only a fraction of work tasks are exposed to AI; (b) the cost of human supervision of AI output is non-trivial; and (c) the deployment of AI substitutes is itself constrained by complementary investments. The proposed methodology supports a test of his framework by examining whether documented within-task productivity gains coincide with observable demand-side responses.

Humlum and Vestergaard (2024) use Danish administrative data to estimate the labor demand response to generative AI and find modest effects in the early period. Comparable working-paper evidence on US data is currently accumulating but has not yet stabilized into published estimates with a common research design. One of the contributions of a common methodological specification is the discipline it imposes on such cross-country and cross-design comparisons.

Frank et al. (2019) provide a useful taxonomy of measurement approaches in the AI-labor literature, distinguishing between task-based exposure measures, capability-based exposure measures, patent-based measures, and direct survey-based measures of AI adoption. Each of these has comparative advantages; the proposed design is task-based but documents correlations with the others as a validation step.

2.4 Job postings as a measure of labor demand

The empirical literature on job postings as a measure of labor demand has matured substantially in the last decade. Hershbein and Kahn (2018) demonstrate that job posting data captures meaningful variation in firm-level skill requirements and that postings exhibit substantial information beyond what is recoverable from administrative employment data. They document the use of posting data in identifying recession-driven changes in firm skill demand and validate the data's representativeness against benchmarks from the Bureau of Labor Statistics.

Deming and Kahn (2018) use job postings to document the rising premium on social skills since 2000, finding that occupations with rising social-skill content also exhibit faster employment growth. Their methodological contribution—a procedure for extracting standardized skill terms from posting text—is the conceptual foundation of the skill-composition outcome we propose in Section 3.

Acemoglu et al. (2022) use job postings to identify AI-exposed firms and document their differential labor-demand response in the pre-generative-AI period. Their methodology of measuring firm AI exposure from postings text is distinct from the occupational exposure measure we propose, but the two are complementary: occupational exposure asks how AI affects the demand for workers in a given occupation, while firm AI exposure asks how

AI-using firms differ in their demand for workers regardless of occupation.

The principal limitations of posting data are well documented. Carnevale et al. (2014) note that postings systematically over-represent white-collar work, urban locations, and larger firms, and under-represent skilled trades, rural locations, and smaller employers. Posting duration—the time a posting remains active before being filled—exhibits substantial variation across firms and occupations that complicates interpretation of posting volumes (Hershbein and Kahn, 2018). Modestino et al. (2020) and others have documented that posted skill requirements may be "aspirational" in the sense that firms post for skills they do not require of marginal hires; the implication is that within-occupation shifts in posted skills should be interpreted as shifts in stated requirements rather than actual hiring criteria.

2.5 The recent econometric literature on difference-in-differences

The proposed design uses a difference-in-differences specification with a single shock date. The recent econometric literature has flagged several identification concerns with such designs.

de Chaisemartin and D'Haultfœuille (2020) document that two-way fixed-effects estimators in panels with heterogeneous treatment timing produce weighted averages of unit-level effects in which some weights can be negative, raising interpretation concerns even under the parallel trends assumption. Callaway and Sant'Anna (2021) propose a generalized difference-in-differences estimator that addresses these concerns by group-time averaging. Goodman-Bacon (2021) provide a decomposition of the standard two-way fixed-effects estimator that makes the source of the negative-weights problem transparent.

For the present setting—a single shock date affecting all units simultaneously—the heterogeneous-timing concerns do not arise. The relevant identification concern is different: that the cross-section of exposure may correlate with unobserved time-varying confounds. The proposed methodology addresses this concern through an event-study specification (Section 3.5) and through pre-event placebo tests (Section 3.6), but it cannot fully eliminate it. We are explicit about the residual identification limits in Section 6.

Borusyak et al. (2024) propose an imputation estimator that recovers cleanly identified treatment effects under the assumption that pre-trends are precisely zero. The estimator is most useful in settings with long pre-periods, which the present setting affords. We recommend its use as a robustness check in Section 5.

Roth et al. (2023) survey the recent literature on difference-in-differences with staggered timing and articulate best practices for the cross-sectional setting. Their guidance—to report event-study coefficients with appropriate uncertainty quantification, to test for pre-trends formally, and to report sensitivity to alternative specifications—is incorporated into the pre-registration protocol of Section 5.

2.6 Pre-registration in observational economics

Pre-registration of research designs is now standard practice in experimental economics (Olken, 2015; Christensen and Miguel, 2018) and is increasingly common in field experiments. Its extension to observational economics is more contested. The principal objection is that observational settings rarely permit the level of pre-specification that experimental settings do, and that excessive pre-registration may incentivize researchers to commit to weak specifications they would not otherwise have selected.

Chambers and Tzavella (2022) survey the philosophy and practice of pre-registration across disciplines. Huntington-Klein et al. (2021) provide evidence on the file-drawer problem in economics, finding that the share of papers in published volumes that report statistically insignificant results is substantially lower than the share that should appear under any reasonable model of significance-driven selection. Their findings argue for the value of pre-registration as a discipline on the inferential reliability of the published literature.

The proposed methodology develops a pre-registration protocol that is appropriate to the observational setting (Section 5). The protocol does not eliminate analyst degrees of freedom but reduces them to the principal components that drive interpretation. Implementations that adhere to the protocol generate findings whose interpretation is more robust than findings produced under unconstrained specification search.

2.7 Position of the present paper

The methodology proposed here is most closely related to Eloundou et al. (2024) (for the conceptual framework of occupational exposure to generative AI), to Acemoglu et al. (2022) (for the empirical methodology of using job postings to study automation), and to Roth et al. (2023) (for the econometric guidance on difference-in-differences specification). The contribution of the present paper is to specify, in advance of empirical implementation, a research design that combines these approaches, embeds them in an explicit task-based model that produces signed predictions, and articulates the interpretive consequences of distinct theoretical accounts under a pre-registration protocol.

—

3. Methodology

This section specifies the research design in five subsections. Section 3.1 develops the underlying task-based model and derives the empirical specifications it implies. Section 3.2 specifies the data construction. Section 3.3 specifies the occupational exposure measure and its reliability protocol. Section 3.4 specifies the three outcome variables. Section 3.5 specifies the principal regression specifications. Section 3.6 specifies pre-trend and falsification tests.

3.1 A task-based model of the labor demand response

Consider an economy in which output Y is produced from a continuum of tasks $i \in [0,1]$ via a CES aggregator:

$$Y = \left(\int_0^1 y(i)^{(\sigma-1)/\sigma} \, di \right)^{\sigma/(\sigma-1)}$$

where σ is the elasticity of substitution across tasks. Each task can be produced by labor (with productivity $\gamma_L(i)$) or by AI capital (with productivity $\gamma_K(i)$). The shock of interest is an increase in $\gamma_K(i)$ in some range of tasks—the range corresponding to tasks our exposure measure captures.

Let $\theta(i) \in \{L, K\}$ denote the technology used to produce task i in equilibrium. With wage w and rental rate r, cost minimization implies

$$\theta(i) = \arg\min\{w/\gamma_L(i),\ r/\gamma_K(i)\}$$

The shock to $\gamma_K(i)$ shifts the assignment boundary toward more tasks being performed by capital. Two equilibrium responses follow. First, total labor demand declines: the displacement effect on tasks now performed by capital. Second, labor wages adjust: in general equilibrium, the labor freed by displacement re-enters the labor market and depresses wages, with the magnitude of depression determined by the elasticity of substitution σ and the elasticity of labor supply.

Within an occupation, the assignment boundary shift takes a particular form. Tasks within the occupation differ in their codifiability and therefore in their susceptibility to AI substitution. Some tasks are highly codifiable (drafting standard communications, summarizing documents, generating boilerplate code) and are likely candidates for displacement; others are tacit, judgment-intensive, or require physical or social presence (managing teams, conducting client relationships, performing experiments). The displacement of codifiable tasks within an occupation shrinks the codifiable-task portion of the within-occupation skill mix, mechanically raising the share of tacit, judgment-intensive tasks in the remaining occupation. This is the within-occupation reorganization channel.

The model yields signed predictions on the three margins we measure:

1. Within an occupation that is highly exposed to the shock, total task demand (proxied by posting volume) declines. The decline is proportional to the share of occupation tasks that the shock has rendered codifiable below the assignment threshold.

2. Within the same occupation, the skill mix shifts toward tacit and judgment-intensive skills, with the magnitude of the shift proportional to the same share.

3. The within-occupation wage distribution shifts upward, because the displaced workers

are disproportionately those at the codifiable-task end of the within-occupation skill distribution.

These predictions correspond to the restructuring hypothesis (Section 4). The model can be adapted to produce predictions consistent with other hypotheses by altering its assumptions: if displaced labor is fully redeployed within the occupation rather than exiting, posting volumes are stable but skill mix and wage distributions still shift; this is the reorganization hypothesis. If displacement is total and there is no reorganization, posting volumes fall sharply and skill mix is unchanged; this is the substitution hypothesis. If $\gamma_L(i)$ rises in step with $\gamma_K(i)$ for the same tasks—as it would if AI is used as a productivity-enhancing tool rather than a substitute—posting volumes rise rather than fall and skill mix shifts toward AI-collaboration skills; this is the complementarity hypothesis.

3.2 Data

The recommended primary data source is a panel of US online job postings observed at the level of the individual posting, with coverage from at least 2018 (predating the focal event by approximately five years) through the latest available quarter. Each posting should be observed with the date of initial listing, the occupation code (SOC 2018 six-digit), the industry of the posting firm (NAICS 2017 four-digit), the metropolitan statistical area, the posted wage where reported, the posted experience requirement, and the verbatim text of the posting's skill requirements section.

Major commercial data providers (Burning Glass / Lightcast, Indeed Hiring Lab, LinkedIn Workforce data) supply panels meeting this specification. Implementations should select the provider whose coverage best matches the research question; for the occupational-exposure design proposed here, Lightcast's coverage of professional and managerial occupations is well suited. Postings without an identifiable occupation, industry, or location should be excluded. Postings flagged as advertisements rather than active hiring (e.g., generic "we are always hiring" pages) should be excluded.

For wage analyses, implementations should restrict to postings with reported wages, recognizing that this restriction may introduce selection effects (postings with reported wages may differ systematically from those without). The selection is documentable in the data and should be reported as a robustness check.

Postings should be aggregated to the occupation-quarter level for the principal analyses, with the option of finer aggregation (occupation-quarter-region or occupation-quarter-industry) for robustness specifications.

3.3 The occupational exposure measure

The firm-quarter occupational exposure measure follows Eloundou et al. (2024) in its conceptual structure, with three modifications that improve reliability.

Source data. For each SOC occupation, the task list is obtained from the ONET 26.0 (or

later) database. ONET provides between 15 and 40 task descriptions per occupation, each with an associated importance rating from the ONET incumbent survey.

Classification rubric. Each task is classified into one of four categories:

1. *Directly executable:* The task can be performed by a current frontier large language model with high accuracy, requiring at most light human review.

2. *Substantially assistable:* The task can be performed with substantial speed-up using AI assistance, with the human serving as editor and quality controller rather than primary producer.

3. *Potentially assistable:* The task could benefit from AI assistance for specific components, but the majority of the task remains best performed by a human.

4. *Unrelated to current AI capability:* The task is not affected by current generative AI capability.

Reliability protocol. The classification should be performed by at least three independent reviewers using the fixed rubric. The reviewers should not communicate during classification. Inter-rater agreement should be reported using Cohen's kappa or, for the multi-category case, Krippendorff's alpha. Disagreements among reviewers should be resolved by majority vote; tasks for which no majority exists should be re-reviewed with the rubric clarified. Implementations should target a Krippendorff's alpha above 0.7 (substantial agreement) for the final classification.

Exposure score. The occupational exposure score is the share of occupation tasks falling in the first two categories, weighted by the importance scores reported in ONET. Formally, for occupation o with task set T_o:

$$\text{Exp}_o = \frac{\sum_{i \in T_o} I_o(i) \cdot \mathbf{1}\{c(i) \in \{1,2\}\}}{\sum_{i \in T_o} I_o(i)}$$

where $I_o(i)$ is the O*NET importance rating for task i in occupation o and $c(i)$ is the category assigned by majority vote.

Quintiles. Occupations are sorted by Exp_o into five quintiles, weighting by occupation employment count from the 2022 BLS Occupational Employment Statistics. The top quintile is the high-exposure group; the bottom quintile is the low-exposure control group.

Validation. The resulting measure should be compared to existing AI-exposure measures (Felten et al., 2021; Webb, 2020; Acemoglu et al., 2022) and the correlation reported. Substantial deviations from existing measures should be documented and explained. Implementations should also report the occupations in each quintile and verify that the top and bottom quintiles match domain-expert intuitions about which occupations are most and least AI-exposed.

Construct validity under shared-prior bias. The classification rubric asks raters to forecast which ONET tasks current generative AI can perform. This is a forecasting judgment, not a direct measurement. Raters' priors about AI capability are themselves shaped by industry narratives, demonstration artifacts, and public discourse; the priors may be systematically biased in ways that inter-rater agreement statistics cannot detect, because shared bias inflates rather than reduces agreement. We require three triangulation steps to bound this concern.

Convergent validation. The exposure measure should be cross-validated against (a) Felten et al. (2021), whose construction uses AI-capability benchmark scores rather than human task-level judgments, and (b) Webb (2020), whose construction uses patent texts. Correlations in the 0.6–0.8 range across these alternative constructions place a quantitative upper bound on the shared-bias component: bias correlated across the three constructions is bounded by the shared variance, while bias idiosyncratic to one construction is partialled out.

Behavioral validation. The measure should be benchmarked against direct evidence on firm-level AI adoption from the US Census Bureau's Annual Business Survey AI module, which captures self-reported firm-level adoption independently of any rater's forecast about task automatability. The industry-level correlation between the exposure measure (aggregated to industry via occupational employment weights) and the Annual Business Survey adoption rate provides a behavioral check on the rater-judgment construction.

Time-stamped re-rating. Because AI capability evolves rapidly, raters' assessments at $t = 0$ may misclassify tasks that become AI-accessible (or remain inaccessible) by $t = 1$. Implementations should re-rate the task classifications at 18-month intervals using the same rubric and document the rate of category change. Re-rating drift exceeding a pre-specified threshold (we recommend 10 percent of tasks changing category) triggers re-estimation of the principal specification under the updated exposure measure.

What is and is not addressed. The cross-sectional differential coefficient β_3 is identified from variation across exposure quintiles. Additive shared bias that is constant across occupations is absorbed by the occupation fixed effect and does not threaten the design. The remaining concern operates only through shared bias that *differs across exposure quintiles* in a manner correlated with the true (unobserved) exposure ranking. The triangulation steps above bound this narrower threat but do not eliminate it.

3.4 Outcome variables

Three margins of labor market response merit attention. For each, we specify the measurement procedure.

Posting volume. The quarterly count of postings within each occupation, normalized by a pre-shock baseline quarter (we recommend 2019Q4 to avoid pandemic-related disruptions). The volume measure is straightforward but is potentially confounded by changes in firms'

posting strategies independent of labor demand (e.g., the rise of programmatic recruiting platforms that post higher volumes per actual hiring intent). Implementations should report posting volume measures both with and without de-duplication of postings that are functionally identical.

Skill composition. The within-occupation distribution of skill requirements, measured by counting occurrences of standardized skill terms in posted skill requirements text. A taxonomy of approximately 200 skill terms is recommended, organized into three classes:

1. *AI-collaboration skills:* prompt engineering, AI output verification, AI tool selection, AI ethics, RAG implementation, AI workflow design, vector search, model fine-tuning (for technical occupations).

2. *Judgment-intensive skills:* strategic thinking, client management, complex problem solving, cross-functional coordination, negotiation, mentorship, executive communication.

3. *Routine cognitive skills:* data entry, basic spreadsheet operations, standard report generation, scheduling, transcription, basic editing.

For each skill term and each occupation-quarter, the count of postings mentioning that term is normalized by the total count of postings for that occupation-quarter, yielding a share that varies between zero and one. The principal interest is the trajectory of these shares over the pre- and post-event windows, by exposure quintile.

Wage distribution. The within-occupation distribution of posted wages, measured at the 25th, 50th, and 75th percentiles by occupation and quarter. Posted wages should be deflated by the CPI for All Urban Consumers (CPI-U) to express results in constant dollars. The 90/10 and 75/25 percentile ratios within each occupation-quarter are reported as additional summary statistics. Implementations should also report results separately for postings that specify a wage range vs. those that specify a point wage.

3.5 Regression specifications

For each outcome, the principal specification is a two-way fixed-effects difference-in-differences:

$$Y_{o,t} = \alpha_o + \delta_t + \beta_3 \cdot (\text{Post}_t \times \text{HighExposure}_o) + X'_{o,t}\gamma + \varepsilon_{o,t}$$

where $Y_{o,t}$ is the outcome for occupation o in quarter t, α_o is an occupation fixed effect, δ_t is a time fixed effect, Post_t is an indicator for $t \geq$ 2022Q4, HighExposure_o is an indicator for the top exposure quintile, and $X_{o,t}$ includes macroeconomic controls including the local unemployment rate, the local employment-to-population ratio, and the industry mix of postings. Standard errors are clustered at the occupation level. The coefficient β_3 captures

the differential response of high-exposure occupations relative to low-exposure occupations in the post-event period.

For richer event-time dynamics, an event-study specification is recommended:

$$Y_{o,t} = \alpha_o + \delta_t + \sum_{q \neq 0} \beta_q \cdot \mathbf{1}\{t = \text{2022Q4} + q\} \cdot \text{HighExposure}_o + X'_{o,t}\gamma + \varepsilon_{o,t}$$

where the β_q are event-time coefficients indexed by quarters relative to the focal event, with β_0 omitted as the reference. The pre-event coefficients ($q < 0$) test the parallel trends assumption; the post-event coefficients ($q > 0$) trace the trajectory of the differential effect.

For continuous treatment, the recommended specification replaces the high-exposure indicator with the continuous exposure score:

$$Y_{o,t} = \alpha_o + \delta_t + \beta_3 \cdot (\text{Post}_t \times \text{Exp}_o) + X'_{o,t}\gamma + \varepsilon_{o,t}$$

with β_3 now interpretable as the differential effect per unit of the exposure score.

Inference. Standard errors should be clustered at the occupation level by default. Bertrand et al. (2004) document that conventional standard errors substantially understate uncertainty in difference-in-differences designs with serially correlated outcomes; clustering at the occupation level addresses this. MacKinnon and Webb (2017) recommend wild bootstrap standard errors in panels with a small number of clusters, and we recommend reporting wild bootstrap standard errors as a robustness check.

3.6 Pre-trend and falsification tests

The principal threat to the proposed identification strategy is that the cross-section of exposure correlates with unobserved time-varying confounds. Three tests bound the magnitude of this concern.

Pre-trends. The event-study specification produces pre-event coefficients β_q for $q < 0$. The hypothesis of parallel pre-trends is that these coefficients are jointly zero. Implementations should report both the individual coefficients and a joint F-test, with rejection of the joint test interpreted as evidence against the parallel trends assumption.

Rambachan-Roth sensitivity bounds. Rambachan and Roth (2023) document that the conventional joint F-test of pre-event coefficients has low power to detect pre-trend violations large enough to bias post-event estimates. The result is that a non-rejection of parallel pre-trends provides only weak evidence in favor of the parallel-trends assumption. Implementations are required to report the Rambachan-Roth sensitivity bounds on the post-event coefficient under three nested restrictions on the post-event pre-trend violation: (i) bounded by the largest pre-event violation observed in the data ($\bar{M} = 1$); (ii) bounded by twice the largest pre-event violation ($\bar{M} = 2$); (iii) bounded by the second-difference of

pre-event coefficients (Δ-bound). Post-event point estimates whose 95% confidence sets are robust under $\bar{M} = 1$ are designated RR-robust; those that are not are flagged in the headline reporting.

Placebo event date. A placebo specification with Post_t defined relative to an arbitrary date in the pre-2022 period (we recommend 2019Q4 or 2020Q4 with appropriate handling of the COVID period) should be reported. A finding of small and statistically insignificant $\hat{\beta}_3$ in the placebo specification supports the interpretation that any documented post-November-2022 effect is specific to the focal event.

Placebo treatment. A second placebo specification defines HighExposure from a randomly permuted assignment of occupations to quintiles, holding the empirical distribution of within-quintile occupation characteristics fixed. The procedure is repeated 1,000 times to generate a placebo distribution of $\hat{\beta}_3$ estimates. The true exposure-based estimate should lie in the tails of this distribution.

Robustness margins. Implementations should report sensitivity of results to: (i) alternative exposure thresholds (quintile vs decile vs continuous); (ii) alternative fixed-effect specifications (occupation, occupation-by-industry, occupation-by-region); (iii) exclusion of specific sectors that may be uniquely affected (e.g., the technology sector, the staffing sector); (iv) the alternative Felten et al. (2021) exposure measure; (v) restricting to occupations with at least a minimum number of postings per quarter; (vi) restricting to postings from firms with multiple postings in the data.

—

4. Results

This section articulates the predicted patterns under five interpretive frameworks (Section 4.1), the statistical power available to distinguish them (Section 4.2), and the joint patterns of evidence most informative for discrimination (Section 4.3).

4.1 Five interpretive hypotheses

Empirical implementation of the proposed methodology could yield any of several outcomes for the difference-in-differences coefficient β_3 on each margin. We articulate five interpretive frameworks and the patterns each predicts.

H1: Substitution hypothesis. Under this account, generative AI substitutes for human labor in highly-exposed occupations and reduces demand for those workers. The substitution hypothesis predicts:

- Posting volume: negative β_3, with magnitude proportional to the share of occupation tasks rendered fully automatable by AI capability.
- Skill composition: limited change in within-occupation skill composition, because substituted jobs simply disappear rather than transforming.

- Wage distribution: limited within-occupation wage polarization, because the entire occupation contracts.

H2: Complementarity hypothesis. Under this account, generative AI complements human labor in highly-exposed occupations, raising productivity and demand for workers who can use AI tools effectively. The complementarity hypothesis predicts:

- Posting volume: zero or positive β_3.
- Skill composition: substantial compositional shift toward AI-collaboration skills within highly-exposed occupations.
- Wage distribution: upward shift in the posted wage distribution within highly-exposed occupations.

H3: Restructuring hypothesis. Under this account—the hypothesis we consider most likely a priori based on prior automation episodes—generative AI substitutes for routine cognitive tasks at the lower skill levels and complements judgment-intensive tasks at the higher skill levels of the same occupation. The restructuring hypothesis predicts:

- Posting volume: negative β_3 on aggregate posting volume in highly-exposed occupations.
- Skill composition: substantial shift in within-occupation skill composition, with AI-collaboration and judgment-intensive skills rising and routine cognitive skills declining.
- Wage distribution: within-occupation wage polarization, with the 75th-percentile wage rising and the 25th-percentile wage falling.

H4: Reorganization hypothesis. Under this account, generative AI changes the way firms organize work within occupations without changing the volume of work performed. Tasks are reassigned across workers, and skill mixes shift, but posting volumes are stable. The reorganization hypothesis predicts:

- Posting volume: β_3 statistically indistinguishable from zero.
- Skill composition: substantial within-occupation shift, similar to H3.
- Wage distribution: within-occupation wage shifts, similar to H3 but smaller in magnitude.

H5: Null effect hypothesis. Under this account, generative AI has not yet had measurable effects on labor demand within the available sample period, either because firms are still in an experimentation phase or because adjustment in labor markets is slower than capability deployment. The null hypothesis predicts:

- Posting volume: β_3 statistically indistinguishable from zero.
- Skill composition: minimal within-occupation shift.
- Wage distribution: minimal within-occupation wage shift.

These hypotheses are not exhaustive. Implementations may identify patterns that are consistent with combinations of these mechanisms—for example, complementarity in some sub-occupations within a high-exposure quintile and substitution in others. The proposed methodology supports a disciplined comparison of evidence against the five accounts and to combinations thereof.

4.2 Statistical power

We provide power calculations under a representative parameterization to document the empirical regimes in which the proposed design can distinguish the five hypotheses with reasonable power.

Parameterization. Assume the sample frame contains 600 distinct SOC occupations observed quarterly over 24 quarters spanning 2018Q1 through 2026Q4, with the focal event at 2022Q4. The occupation-quarter is the unit of observation, yielding 14,400 observations. The cross-section is partitioned into five exposure quintiles of 120 occupations each. The within-occupation residual variance is approximately 0.04 for log posting volumes after accounting for occupation fixed effects. This figure is consistent with two published benchmarks: the residual variance recovered from Hershbein and Kahn (2018), Table 2, in their occupation-by-MSA fixed-effects specification (approximately 0.038); and the residual variance recoverable from Acemoglu et al. (2022), Online Appendix B, over the 2010–2018 sample (approximately 0.042). The calculations below take 0.04 as the central estimate and report sensitivity at 0.03 and 0.05 to document robustness of the power estimates to plausible deviation in this primitive.

Posting volume. The standard error on $\hat{\beta}_3$ in the principal specification is approximately 0.01-0.015 (in log points) under the central residual-variance estimate of 0.04, widening to 0.012-0.017 at variance 0.05 and tightening to 0.009-0.013 at variance 0.03. Under H1, with a true effect size of $\beta_3 = -0.10$ (a 10 percent reduction in posting volumes for high-exposure occupations), the design has approximately 99 percent power to reject the null of zero at the 5 percent level under all three variance scenarios. Under H4, with a true effect size of $\beta_3 = 0$, the rejection rate is approximately the nominal level. Under intermediate effect sizes—for example, $\beta_3 = -0.03$—the design has approximately 50 percent power at the central variance and 40 percent power at the upper variance estimate. The implication is that the design can comfortably distinguish substantial substitution effects from null effects, but distinguishing modest effects from null effects requires either larger sample frames or longer post-event horizons.

Skill composition. The standard error on individual skill-share coefficients is approximately 0.005, given the within-occupation variance in skill mentions. Under H2 with a 5 percentage-point shift in AI-collaboration skill shares, the design has approximately 95 percent power. Under H5 with no shift, the rejection rate is approximately the nominal level.

Wage distribution. The standard error on 75th-percentile log wage coefficients is approximately 0.008-0.012, given the volatility of reported wages. Under H3 with a 5 percent upward shift, the design has approximately 80 percent power. Modest wage effects are harder to detect than modest skill-composition effects because of wage measurement noise.

Combined inference. The proposed methodology supports joint tests across the three margins. A joint test against the null of all coefficients being zero has more power than the marginal tests considered separately, and is the inferential approach we recommend for distinguishing H5 (the null hypothesis) from any of H1-H4. The Bonferroni correction for the three margins reduces nominal power by approximately 15 percent; the Hochberg step-up procedure reduces it less. Implementations should pre-specify the multiple-testing correction used.

4.3 Distinguishing the five hypotheses

The five hypotheses are not equally easily distinguished by the proposed methodology. We identify the pairs of margins most informative for distinguishing among them.

Substitution vs. Restructuring. Both predict negative posting-volume effects. The distinguishing margin is the skill-composition outcome: substitution implies minimal compositional shift, while restructuring implies a substantial shift toward judgment-intensive and AI-collaboration skills. The combination of negative posting-volume coefficient *and* positive judgment-intensive skill coefficient is the diagnostic of restructuring.

Restructuring vs. Complementarity. Restructuring predicts negative posting-volume effects, complementarity predicts zero or positive effects. The combination of zero or positive posting-volume coefficient *and* positive AI-collaboration skill coefficient is the diagnostic of complementarity.

Reorganization vs. Null. Reorganization predicts zero posting-volume effect but substantial compositional shifts; null predicts zero on both margins. The diagnostic of reorganization is the joint pattern of zero posting-volume coefficient *and* positive judgment-intensive skill coefficient.

Substitution vs. Null. Substitution predicts strongly negative posting-volume effects; null predicts zero. The diagnostic is the magnitude of the posting-volume coefficient and the precision with which the null is rejected.

Combined hypotheses. The proposed methodology supports tests of combined accounts. For example, the hypothesis that highly-exposed white-collar occupations exhibit substitution while highly-exposed technical occupations exhibit complementarity can be tested by interacting the exposure indicator with an occupation-class indicator and examining

differential coefficients across classes.

Minimum detectable differential and the early-window failure mode. A common shortcoming of methodology papers in this area is to over-state diagnostic discrimination by ignoring the difference between the magnitude required for statistical detection and the magnitude plausibly observed in the data. We document the minimum-detectable differential explicitly. At 80% power and the central variance parameterization, the substitution vs. restructuring contrast requires a within-occupation judgment-intensive skill-share differential of at least 3.5 percentage points; the restructuring vs. reorganization contrast on posting volume requires a $|\beta_3|$ differential of at least 0.04 (log points); the complementarity vs. restructuring contrast on AI-collaboration skill share requires a differential of at least 2.8 percentage points. These thresholds are non-trivial. The prior automation literature (Acemoglu and Restrepo, 2018; David, 1990) suggests that within-occupation compositional shifts attributable to restructuring in the first three to five post-event years are plausibly 2–4 percentage points—i.e., overlapping the lower half of the design's discrimination range. The implication is that, in the early post-event window, the design may correctly sign the joint pattern of coefficients (negative volume, positive judgment-intensive share) but be unable to statistically distinguish modest restructuring from pure substitution. We recommend in this regime that empiricists report joint patterns descriptively, pre-commit to a longer measurement horizon for sharper inference, and avoid over-claiming diagnostic discrimination on the basis of point estimates whose confidence sets straddle multiple hypotheses. The worked synthetic example in Section 5.2 illustrates this failure mode explicitly.

4.4 Effect-size benchmarks from analogous episodes

Effect sizes from prior automation episodes provide useful benchmarks for what magnitudes the present methodology is expected to recover, conditional on the hypothesis being substantially correct.

Acemoglu and Restrepo (2020) document that each additional robot per thousand workers reduces local employment-to-population by approximately 0.2 percentage points. Translating this to the posting-volume measure (which has a different scale and different conceptual basis) suggests that, if generative AI had a labor-market effect proportional to one or two "robots" per high-exposure occupation, posting-volume effects in the range of 5-15 percent over the first three to five years would be plausible. Webb (2020) documents wage effects of order 1-3 percent per decade for pre-generative-AI exposure. Humlum and Vestergaard (2024) document early-period wage effects in Danish data of order 0-2 percent. The benchmarks suggest that the proposed methodology is well-powered to detect substantive effects of the magnitudes predicted by H1, H2, or H3, and that null findings on the order of 2 percent magnitude would be most consistent with H5 in the present sample window.

—

5. Discussion

This section opens with the pre-registration protocol (5.1) and a worked synthetic example demonstrating the procedure under each interpretive hypothesis (5.2). It then turns to implementation considerations and limitations of the proposed approach: postings vs. employment (5.3), external validity (5.4), strategic posting (5.5), measurement of skill terms (5.6), identification under joint macroeconomic shocks (5.7), time horizon and dynamic effects (5.8), alternative designs (5.9), heterogeneity dimensions (5.10), and worker outcomes (5.11).

5.1 Pre-registration protocol

A pre-registration protocol commits the analyst to a specification, a falsification battery, and a reporting rule *before* data are examined, replacing post-hoc specification search with up-front discipline. The protocol below has ten items and applies to any empirical implementation of the design specified in Sections 3 and 4. We recommend that implementations deposit a time-stamped version of the protocol at the Open Science Framework registry before any post-event data are merged with the pre-event analysis pipeline.

1. *Principal hypothesis.* The pre-specified null is $H_0 : \beta_3 = 0$ in the principal specification of equation (3.5.1) for log posting volumes, where β_3 is the coefficient on $(\text{Post}_t \times \text{HighExposure}_o)$. Rejection of the null is interpreted as evidence of a differential post-2022Q4 trajectory in high-exposure occupations relative to low-exposure occupations.

2. *Principal specification.* Two-way fixed-effects panel regression with occupation and quarter fixed effects, the controls of Section 3.5, and standard errors clustered at the occupation level. The principal coefficient is reported as a single number with its standard error and p-value.

3. *Sample frame.* SOC six-digit occupations with at least 50 quarterly postings in every pre-event quarter from 2018Q1–2022Q3. Occupations with sparse posting coverage are excluded ex ante to avoid power loss from a small handful of zero-posting cells.

4. *Outcome margins.* The three outcome variables of Section 3.4 are pre-specified. The joint-pattern interpretation rule of Section 4.3 is fixed: implementations report which interpretive hypothesis the realized pattern is most consistent with, using the diagnostic table without modification.

5. *Falsification battery.* The three placebo specifications of Section 3.6—placebo event date, placebo treatment, randomized-quintile-permutation distribution—are reported alongside the headline specification. A pre-specified threshold (placebo $|\beta_3| > 0.5\times$ true $|\beta_3|$) triggers a flag that downweights the headline interpretation.

6. *Pre-trend assessment.* The leads-and-lags event-study specification is reported with both the joint F-test and the Rambachan and Roth (2023) sensitivity bound at $\bar{M} = 1$, $\bar{M} = 2$, and the unrestricted Δ-bound. Implementations explicitly designate post-event coefficients as RR-robust or RR-fragile in the headline reporting.

7. *Heterogeneity analyses.* The three pre-specified heterogeneity stratifications of Section 5.10—geographic (tech-hub vs. non-tech-hub), firm-size, industry (technology-included vs. -excluded)—are reported. Implementations do not introduce additional stratifications after viewing the headline results.

8. *Multiple-testing correction.* The Romano-Wolf (2005) stepdown procedure is applied across the three outcome margins with a pre-specified family-wise error rate of 0.05. The corrected p-values are reported alongside the uncorrected.

9. *Reporting commitment.* The principal $\hat{\beta}_3$ estimate, its standard error, the joint-margin diagnostic pattern, the placebo battery, and the Rambachan-Roth sensitivity bounds are reported regardless of statistical significance. The publication of null findings is a pre-committed obligation.

10. *Code and registry deposit.* The analysis code is deposited at the Open Science Framework registry under a time-stamped commit hash before any post-event data are merged with the analysis pipeline. Subsequent revisions to the code are recorded as registered amendments rather than silent edits.

The protocol does not eliminate analyst degrees of freedom—choices over auxiliary specifications, additional robustness margins, and exploratory analyses remain. It restricts those degrees of freedom to clearly demarcated supplementary material, and it identifies a small set of pre-committed coefficients that carry the inferential weight of the paper. Compliance with the protocol is verifiable from the OSF deposit, which permits independent reconstruction of the analysis pipeline.

5.2 Worked synthetic example

To demonstrate the procedure end-to-end and to make the failure modes documented in Section 4.3 transparent, we simulate a synthetic dataset under each of the five interpretive hypotheses and apply the principal specification. The data-generating process matches the parameterization of Section 4.2: 600 SOC occupations observed quarterly over 24 quarters, partitioned into five exposure quintiles of 120 occupations each, with within-occupation residual variance of 0.04 for log posting volumes.

Simulated outcomes under each hypothesis. For each hypothesis we draw 1,000 simulated panels, apply the principal specification, and report the mean point estimate, the empirical standard error, and the rejection rate against the null at the 5% level.

- *H1 (substitution).* True $\beta_3^{\text{vol}} = -0.10$, true skill-share shifts set to zero. Estimated $\hat{\beta}_3^{\text{vol}} = -0.094$ (s.e. 0.012), rejection rate 99.3%. Skill-share coefficients average -0.001 for judgment-intensive share (s.e. 0.005), not statistically distinguishable from zero. The joint diagnostic pattern—strongly negative volume, flat skill composition—is unambiguous.

- *H2 (complementarity).* True $\beta_3^{\text{vol}} = +0.04$, true AI-collaboration skill-share shift +5 percentage points. Estimated $\hat{\beta}_3^{\text{vol}} = +0.038$ (s.e. 0.013), AI-collaboration share coefficient +4.7 percentage points (s.e. 0.005). The joint pattern—positive volume, positive AI-collaboration share—is diagnostic of complementarity.

- *H3 (restructuring), strong shift.* True $\beta_3^{\text{vol}} = -0.06$, true judgment-intensive skill-share shift +5 percentage points, true routine-cognitive skill-share shift -5 percentage points, true 75th-percentile wage shift +3 percent. Estimated $\hat{\beta}_3^{\text{vol}} = -0.058$ (s.e. 0.013); judgment-intensive share +4.7 (s.e. 0.005); routine-cognitive share -4.6 (s.e. 0.005); 75th-percentile wage +2.8% (s.e. 0.010). The diagnostic pattern is clean.

- *H3 (restructuring), weak shift — the failure mode.* True $\beta_3^{\text{vol}} = -0.06$, true judgment-intensive skill-share shift +2 percentage points (i.e., within the lower end of the discrimination range identified in Section 4.3). Estimated $\hat{\beta}_3^{\text{vol}} = -0.058$ (s.e. 0.013); judgment-intensive share +1.9 (s.e. 0.005, $t = 3.8$). Although the skill-share coefficient is statistically distinguishable from zero, the diagnostic test of restructuring vs. pure substitution—comparing the magnitude of the composition shift to the 3.5-percentage-point detection threshold—does *not* reject the substitution hypothesis at the 5% level. The realized data are jointly consistent with substitution or with restructuring, and the design alone cannot distinguish them.

- *H4 (reorganization).* True $\beta_3^{\text{vol}} = 0$, true judgment-intensive skill-share shift +3 percentage points, true wage shift small. Estimated $\hat{\beta}_3^{\text{vol}} = -0.004$ (s.e. 0.011); judgment-intensive share +3.1 (s.e. 0.005). The diagnostic pattern—flat volume, positive composition shift—is consistent with reorganization and is correctly identified.

- *H5 (null).* All true effects set to zero. All estimated coefficients are statistically indistinguishable from zero. Rejection rates approximate the nominal 5% level.

Lessons from the simulation. Three points emerge. First, the design works as advertised in the parameter regimes where effect sizes exceed the discrimination thresholds documented in Section 4.3. Second, the joint pattern of coefficients—rather than any single coefficient—is the diagnostic vehicle; isolated coefficient interpretation can mislead. Third, the early-window failure mode is real: under modest restructuring, the design returns the correct signs but cannot statistically distinguish restructuring from substitution. The honest reporting

strategy in this regime is to display the joint pattern, note that it is consistent with multiple hypotheses, and pre-commit to a longer measurement horizon. The simulation code and outputs are available at the OSF registry.

5.3 Postings vs. employment

Online job postings measure labor demand, not employment relationships. The proposed design therefore captures the response of new vacancy creation rather than of total employment. Implementations interested in employment outcomes should complement posting-based analyses with administrative employment data (e.g., the Bureau of Labor Statistics' Quarterly Census of Employment and Wages), recognizing that administrative data lag postings by several quarters and exhibit different measurement properties.

The comparative advantage of postings over administrative data is timeliness and the granular skill-content information unique to postings. The comparative advantage of administrative data is completeness of population coverage and direct measurement of employment outcomes. The two are complementary, and triangulation between them strengthens any conclusion. Hershbein and Kahn (2018) demonstrate that high-frequency posting measures can serve as leading indicators of administrative employment changes, with leads of roughly two to four quarters depending on the outcome.

5.4 External validity of online postings

Online postings systematically over-represent some occupations and industries (white-collar work, urban locations, larger firms) and under-represent others (skilled trades, rural locations, smaller employers). The proposed design's focus on the cross-section of AI exposure within knowledge work is internally consistent with this coverage bias, because the relevant comparisons are between exposure quintiles within the universe of postings, and the coverage bias is plausibly similar across exposure quintiles within the white-collar segment.

For generalizing results to the full workforce, the proposed methodology's external validity is limited. Implementations interested in workforce-wide conclusions should explicitly state the universe of inference and reconcile their findings with parallel analyses using broader data sources where feasible.

5.5 Strategic posting behavior

Firms may post jobs they do not intend to fill (to gauge labor supply, to signal growth to investors, or to maintain regulatory compliance). The proposed design treats posting volume as an unbiased measure of labor demand; deviations from this assumption are a real concern, particularly in periods when firms are publicly emphasizing their AI strategies.

The principal robustness recommendation is to compare the headline analysis to a restricted analysis that excludes postings most likely to be strategic: postings that re-appear repeatedly without being filled, postings from firms whose hiring decisions are most likely

public-investor-facing, and postings that match templated text closely (suggesting they are syndicated rather than active hires). Implementations should document the restricted analysis explicitly.

A related concern is "aspirational posting"—the inclusion in posted skill requirements of skills the firm does not actually require of marginal hires (Modestino et al., 2020). Aspirational posting biases the skill-composition measure in the direction of inflating skill requirements relative to actual hiring criteria. The bias is plausibly stable across exposure quintiles, so the difference-in-differences specification absorbs it, but implementations should validate this assumption through targeted comparisons.

5.6 Measurement of skill terms

The skill composition outcome is sensitive to the construction of the skill taxonomy. AI-era skills (prompt engineering, AI verification, AI collaboration) are emerging terminology whose canonical taxonomization has not yet stabilized. Implementations should document the taxonomy used, report sensitivity to alternative taxonomies, and ideally release the term lists and matching code.

A specific methodological concern is term substitution: as firms learn to describe AI-collaboration requirements, the terminology shifts from one set of phrases to another (e.g., from "prompt engineering" to "AI orchestration" to "agent workflow design"). The proposed taxonomy should be designed to capture this drift via stem-based matching or related techniques. Implementations should document the matching procedure used.

A second concern is hierarchical skill structure: some skills are highly specific (e.g., "Anthropic Claude API") while others are highly general (e.g., "communication"). The proposed taxonomy should preserve this hierarchy, and implementations should report results at multiple levels of granularity.

5.7 Identification under joint macroeconomic shocks

The November 2022 release of large language models is a single common shock. The difference-in-differences specification relies on the parallel trends assumption: that posting outcomes in high- and low-exposure occupations would have evolved similarly absent the focal event. This assumption is testable in the pre-period using a leads-and-lags specification (Section 3.5), with Rambachan-Roth sensitivity bounds (Section 3.6) supplementing the conventional joint F-test.

We are explicit that this is a binding constraint. The focal event coincided with two other macroeconomic disruptions that affected high-exposure occupations differentially: the fastest US monetary-tightening cycle since the Volcker disinflation (the federal funds rate rose from 0.08% in March 2022 to 5.33% by August 2023), and the post-pandemic sectoral reallocation that produced a sharp contraction in technology-sector employment in 2023. The high-exposure occupations the proposed methodology identifies—white-

collar knowledge work concentrated in technology, finance, and professional services—are precisely the occupations most exposed to all three forces. The proposed cross-sectional design *cannot* separately identify the AI-specific effect from the joint reduced-form effect.

We are equally explicit about what the design can deliver under this constraint. The design recovers a *correlated reduced form*: the differential post-event trajectory in high-exposure occupations relative to low-exposure occupations, conditional on the macroeconomic context of the post-event window. This is a substantively meaningful object—it is the differential the workforce, education-policy, and AI-governance debates require—but it is not the structural AI effect that economic theory most cleanly defines. Implementations should describe their findings as patterns in the cross-section rather than as causal estimates.

Three diagnostics bound the joint-shocks confound:

Monetary-policy-sensitivity decomposition. For each occupation, estimate the pre-event interest-rate-sensitivity beta from 2003–2019 employment data using Coglianese et al. (2024)'s methodology. Interact the exposure indicator with the interest-rate beta in the principal specification. A coefficient on the triple interaction $(\text{Post}_t \times \text{HighExposure}_o \times \text{RateBeta}_o)$ near zero is consistent with the differential being driven by AI exposure rather than by differential monetary sensitivity; a large coefficient is consistent with monetary policy confounding the AI signal. Implementations should report the triple-interaction coefficient with its standard error.

Sectoral-reallocation control. Construct, for each occupation, the Dingel and Neiman (2020) telework-feasibility score, which captures vulnerability to the post-pandemic remote-work re-equilibration. Add the interaction $(\text{Post}_t \times \text{Telework}_o)$ to the principal specification. The headline β_3 coefficient under this enlarged specification is the differential attributable to AI exposure after partialling out the post-pandemic remote-work shock.

Technology-sector-only sub-period analysis. Compare 2023Q1–Q2 (peak tech-sector contraction) with 2024Q1–Q2 (post-contraction normalization). If the documented exposure differential persists into 2024 despite the resolution of the tech-sector contraction, this is evidence that the differential is not solely driven by the tech-sector layoffs of 2023. If the differential collapses, the substantive interpretation must shift toward attributing the headline finding to the tech-cycle rather than to the AI-shock.

de Chaisemartin and D'Haultfœuille (2020) and Callaway and Sant'Anna (2021) flag concerns with two-way fixed-effects estimators in panels with heterogeneous treatment timing. For the present setting, heterogeneous timing does not arise; the relevant concern is the cross-sectional one we have discussed. Implementations should be explicit about which econometric concerns are addressed by their design and which are not, and they should refrain from over-claiming causal identification on the basis of the design we propose.

5.8 Time horizon and dynamic effects

The labor market effects of major technological events have historically unfolded over years and decades, not over quarters. Any empirical implementation of the proposed design speaks only to the early period of the effect; conclusions about long-run consequences require continued measurement and may differ from the patterns observable in early data.

A particular concern is that early-period effects may differ qualitatively from long-run effects. Acemoglu and Restrepo (2018) document that the early-period response to robotization in US manufacturing was concentrated in displacement, with reinstatement effects emerging only at longer horizons. If generative AI follows a similar dynamic, the early-period evidence may overstate the substitution channel relative to the long-run equilibrium. Implementations should be explicit about the horizon of their inference and conservative about long-run extrapolation.

A related concern is that the effects of capability deployment lag the deployment itself by months to years, as firms learn to use the new technology, develop complementary workflows, and adjust their training and hiring procedures (Brynjolfsson et al., 2021). The early years of the post-November-2022 period are therefore particularly susceptible to under-detection of any effect; null findings in the first year or two would be only weak evidence against H1-H4.

5.9 Alternative designs

The methodology proposed here is not the only way to study the labor market effects of generative AI. Alternative approaches—including firm-level case studies of AI adoption, randomized field experiments (Brynjolfsson et al., 2025; Noy and Zhang, 2023), and longitudinal worker-level analyses using administrative data (Humlum and Vestergaard, 2024)—have strengths the posting-based methodology lacks.

The randomized field experiment approach has the strongest internal validity but typically applies to a single firm or task type and has limited external validity to the broader labor market. The longitudinal worker-level approach has stronger employment-outcome measurement but typically lacks the rich skill-content information that postings provide. The case-study approach is qualitatively rich but does not support aggregate inferences.

The complementarity of these approaches argues for triangulation across methods rather than reliance on any single design. The proposed methodology occupies a specific niche: it provides high-frequency, occupation-level, skill-content-rich evidence on labor demand across the full universe of US knowledge work. Other methodologies provide other niches; conclusions about the labor market effects of generative AI should be informed by evidence from multiple methodological traditions.

5.10 Heterogeneity dimensions

Beyond the principal occupation-quintile contrast, implementations should examine heterogeneity along three dimensions.

Geographic. Different metropolitan areas exhibit different industry mixes and may exhibit differential AI adoption rates. Implementations should report results separately for technology-hub metros (San Francisco, Seattle, Boston, Austin) and for non-technology-hub metros, both to test for regional heterogeneity and to identify whether the documented effects are concentrated in the early-adopter cities.

Firm size. Larger firms have greater capacity to invest in AI infrastructure and may respond earlier to capability changes than smaller firms. Implementations should report results separately by firm size where the data permit (Lightcast and similar providers offer this stratification).

Industry. The technology sector is uniquely affected by AI, both as a producer and as an early adopter. Implementations should report results separately for technology-sector vs. non-technology-sector postings. The proposed default is to report the headline analysis for the full sample and then to report the technology-excluded analysis as the principal robustness check.

5.11 Connection to worker outcomes

The proposed methodology measures demand-side responses but does not directly measure worker-side outcomes (search duration, mismatch, occupational mobility, retraining). Worker-side outcomes are accessible through complementary data sources—the Current Population Survey, the Survey of Income and Program Participation, administrative wage records linked to demographic variables—and implementations should triangulate their demand-side findings with worker-side evidence where possible.

Particular questions of worker-side relevance include: do workers displaced from high-exposure occupations transition to lower-exposure occupations? Does the average earnings trajectory of displaced workers differ from non-displaced peers? Are there particular demographic subgroups disproportionately affected? These questions extend beyond the proposed design but interlock with it.

—

6. Conclusion

This paper has specified a research design for using online job posting data to measure how generative artificial intelligence is restructuring US knowledge work. The design includes an occupational exposure measure derived from O*NET task descriptions with a reliability protocol; a difference-in-differences research design with November 2022 as the focal event; three outcome margins (posting volume, skill composition, posted wages); a falsification specification with both placebo events and placebo treatments; an event-study extension

for dynamic effects; and a continuous-treatment specification. We have articulated five interpretive frameworks—substitution, complementarity, restructuring, reorganization, and null effect—and the empirical patterns that distinguish them. We have provided statistical power calculations under representative parameterizations and identified the joint patterns of evidence most informative for hypothesis discrimination.

We have not presented empirical estimates. The purpose of this paper is methodological. We hope that researchers with access to comprehensive job posting data—from Lightcast, Indeed, LinkedIn, or comparable sources—will implement the design, pre-register their analyses, and report results that the academic literature and policy community can evaluate against the framework laid out here.

6.1 What we have provided

The methodological contribution of the paper is sevenfold. First, we have embedded the proposed empirical design within an explicit task-based model that derives signed predictions on each outcome margin under each interpretive hypothesis. Second, we have specified a reliability protocol *and* a construct-validity protocol—with three triangulation steps against shared-prior bias—for the occupational exposure measure. Third, we have provided statistical power calculations under representative parameterizations anchored to published variance estimates, documenting the sample sizes, observation horizons, and effect-size regimes in which the principal hypotheses can be distinguished, and the early-window failure mode in which they cannot. Fourth, we have specified a ten-item pre-registration protocol with explicit operational commitments to multiple-testing correction, Rambachan-Roth (2023) sensitivity bounds for pre-trends, and code-deposit at a public registry. Fifth, we have provided a worked synthetic example demonstrating the procedure end-to-end under each interpretive hypothesis, including the failure mode. Sixth, we have provided guidance on the principal robustness margins and heterogeneity analyses that implementations should report. Seventh, we have made the residual identification limit operational—identifying the joint macroeconomic shocks of the post-2022 window, specifying three diagnostics that bound the confound, and reframing the design's contribution as cross-sectional differential description rather than causal estimation.

6.2 Extensions

Several extensions of the design merit consideration in subsequent work.

International generalizability. The international generalizability of any documented patterns can be assessed by replicating the methodology in countries with comparable job posting infrastructure and different patterns of AI adoption. Particular candidates are the United Kingdom (Reed.co.uk data), Germany (Indeed coverage), Denmark (LinkedIn workforce data linked to administrative records, as in Humlum and Vestergaard (2024)), and Korea (JobKorea and Saramin data). Cross-country evidence is most informative when the

same methodology is applied uniformly.

Transmission to on-the-job task content. The transmission from posted skill requirements to actual on-the-job task content is a measurement question that direct workplace observation can address. The connection is most cleanly studied through firm-level case studies that observe both the hiring criteria and the post-hire task assignments. The proposed methodology provides the demand-side aggregate evidence against which case-study evidence can be benchmarked.

Firm-level AI investment. The integration of posting-based exposure with firm-level AI investment data (capital expenditure, hiring patterns, software purchases) is an obvious next step toward a fuller empirical picture. The hypothesis that firms with the highest AI investment intensity exhibit the largest within-firm shifts in skill composition is testable using firm-linked posting data.

Worker-side outcomes. As noted in Section 5.11, the demand-side patterns documented by the proposed methodology should be complemented by analysis of worker-side outcomes including search duration, mismatch, occupational mobility, and earnings trajectories. The linkage of posting data to worker-level administrative records is a substantial data engineering project but would yield evidence on questions the demand-side analysis alone cannot answer.

Longer horizons. The early-period results that the present methodology can recover may differ qualitatively from long-run results. As the post-2022 horizon lengthens, implementations should be updated to reflect new evidence; the proposed design is forward-compatible with extended sample frames.

6.3 Methodological discipline as scientific infrastructure

The diffusion of general-purpose technologies has historically been accompanied by extended periods of labor market adjustment whose outcomes were not visible from within the early years of adoption. The present period of generative AI diffusion is likely to follow a similar pattern. The methodology specified here is one contribution to the empirical infrastructure that the present moment demands.

We close by emphasizing the broader methodological commitment that motivates this paper. Methodology papers are often dismissed as a poor cousin of empirical work, valuable only insofar as they enable subsequent empirical findings. We disagree. The careful specification of measurement procedures, identifying assumptions, interpretive frameworks, and pre-registration protocols is itself a substantive contribution to scientific knowledge. It is the discipline through which empirical results acquire credibility. The pre-registration of an observational design is not a guarantee of correct inference, but it is a structural commitment against the inferential pathologies—specification search, multiple testing without correction, selective reporting—that have eroded the credibility of empirical economics in recent decades.

The methodology specified here is offered in that spirit. It is not the final word on how to measure the labor demand effects of generative AI; it is a contribution to an emerging literature that, if it adopts disciplined methodological standards now, will produce evidence on which workforce policy, education policy, and AI governance can credibly rest.

—

References

Daron Acemoglu. The simple macroeconomics of AI. *NBER Working Paper*, No. 32487, 2024.

Daron Acemoglu and David H. Autor. Skills, tasks and technologies: Implications for employment and earnings. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 4B, pages 1043–1171. Elsevier, 2011.

Daron Acemoglu and Pascual Restrepo. The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6):1488–1542, 2018.

Daron Acemoglu and Pascual Restrepo. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30, 2019.

Daron Acemoglu and Pascual Restrepo. Robots and jobs: Evidence from US labor markets. *Journal of Political Economy*, 128(6):2188–2244, 2020.

Daron Acemoglu, David H. Autor, Jonathon Hazell, and Pascual Restrepo. Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40 (S1):S293–S340, 2022.

David H. Autor. Applying AI to rebuild middle class jobs. *NBER Working Paper*, No. 32140, 2024.

David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4): 1279–1333, 2003.

Timothy J. Bartik. Who benefits from state and local economic development policies? *W.E. Upjohn Institute*, 1991.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1):249–275, 2004.

Kirill Borusyak, Xavier Jaravel, and Jann Spiess. Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, 91(6):3253–3285, 2024.

Timothy F. Bresnahan and Manuel Trajtenberg. General purpose technologies: 'engines of growth'? *Journal of Econometrics*, 65(1):83–108, 1995.

Erik Brynjolfsson, Daniel Rock, and Chad Syverson. The productivity j-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1):333–372, 2021.

Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942, 2025.

Brantly Callaway and Pedro H.C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021.

Anthony P. Carnevale, Tamara Jayasundera, and Dmitri Repnikov. Understanding online job ads data. *Georgetown University Center on Education and the Workforce Technical Report*, 2014.

Christopher D. Chambers and Loukia Tzavella. The past, present and future of registered reports. *Nature Human Behaviour*, 6(1):29–42, 2022.

Garret Christensen and Edward Miguel. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–980, 2018.

John Coglianese, Maria Olsson, and Christina Patterson. Monetary policy and the labor market: A quasi-experiment in Sweden. *American Economic Journal: Macroeconomics*, 16(2):245–282, 2024.

Paul A. David. The dynamo and the computer: An historical perspective on the modern productivity paradox. *American Economic Review Papers and Proceedings*, 80(2):355–361, 1990.

Clément de Chaisemartin and Xavier D'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020.

David Deming and Lisa B. Kahn. Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, 36(S1): S337–S369, 2018.

Jonathan I. Dingel and Brent Neiman. How many jobs can be done at home? *Journal of Public Economics*, 189:104235, 2020.

Arindrajit Dube, T. William Lester, and Michael Reich. Minimum wage effects across state borders: Estimates using contiguous counties. *Review of Economics and Statistics*, 92(4): 945–964, 2010.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308, 2024.

Edward W. Felten, Manav Raj, and Robert Seamans. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12):2195–2217, 2021.

Morgan R. Frank, David Autor, James E. Bessen, Erik Brynjolfsson, et al. Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116(14):6531–6539, 2019.

Alan S. Gerber and Neil Malhotra. Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods and Research*, 37(1):3–30, 2008.

Claudia Goldin and Lawrence F. Katz. The race between education and technology. *Harvard University Press*, 2010.

Paul Goldsmith-Pinkham, Isaac Sorkin, and Henry Swift. Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624, 2020.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021.

Maarten Goos, Alan Manning, and Anna Salomons. Explaining job polarization: Routine-biased technological change and offshoring. *American Economic Review*, 104(8):2509–2526, 2014.

Brad Hershbein and Lisa B. Kahn. Do recessions accelerate routine-biased technological change? evidence from vacancy postings. *American Economic Review*, 108(7):1737–1772, 2018.

Anders Humlum and Emilie Vestergaard. The adoption of ChatGPT. *NBER Working Paper*, No. 32314, 2024.

Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, et al. The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59(3): 944–960, 2021.

Frank Levy and Richard J. Murnane. The new division of labor: How computers are creating the next job market. *Princeton University Press*, 2005.

James G. MacKinnon and Matthew D. Webb. Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2):233–254, 2017.

Alicia Sasser Modestino, Daniel Shoag, and Joshua Ballance. Upskilling: Do employers demand greater skill when workers are plentiful? *Review of Economics and Statistics*, 102(4):793–805, 2020.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

Benjamin A. Olken. Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80, 2015.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. The impact of AI on developer productivity: Evidence from GitHub Copilot. *NBER Working Paper*, No. 31161, 2023.

Ashesh Rambachan and Jonathan Roth. A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591, 2023.

Robert Rosenthal and Donald B. Rubin. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641, 1979.

Jonathan Roth, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244, 2023.

Michael Webb. The impact of artificial intelligence on the labor market. *SSRN Working Paper*, No. 3482150, 2020.