PNAS RESEARCH ARTICLE | ECONOMIC SCIENCES STATISTICS OPEN ACCESS

Estimating wage disparities using foundation models

Keyon Vafa[a], Susan Athey[b,c,1], and David M. Blei[d,e]

Affiliations are included on p. 10.

Contributed by Susan Athey; received December 31, 2024; accepted April 23, 2025; reviewed by Elliott Ash and Whitney Newey

The rise of foundation models marks a paradigm shift in machine learning: instead of training specialized models from scratch, foundation models are trained on massive datasets before being adjusted or fine-tuned to make predictions on smaller datasets. Initially developed for text, foundation models have also excelled at making predictions about social science data. However, while many estimation problems in the social sciences use prediction as an intermediate step, they ultimately require different criteria for success. In this paper, we develop methods for fine-tuning foundation models to perform these estimation problems. We first characterize an omitted variable bias that can arise when a foundation model is fine-tuned in the standard way: to minimize predictive error. We then provide a set of conditions for fine-tuning under which estimates derived from a foundation model are $\sqrt{n}$-consistent. Based on this theory, we develop fine-tuning algorithms that empirically mitigate this omitted variable bias. To demonstrate our ideas, we study gender wage gap estimation. Classical methods for estimating the adjusted wage gap employ simple predictive models of wages, which can induce omitted variable bias because they condition on coarse summaries of career history. Instead, we use a custom-built foundation model, capturing a richer representation of career history. Using data from the Panel Study of Income Dynamics, we find that career history explains more of the gender wage gap than standard econometric models can measure, and we identify elements of career history that are omitted by standard models but are important for explaining the gap.

machine learning | foundation models | labor economics | econometrics

Foundation models have revolutionized the machine learning approach to prediction (1–3). In contrast to traditional predictive models, which are trained to make predictions on specific, individual tasks, foundation models are typically trained in two steps: they first are trained on massive, passively collected datasets and then are adapted to specific tasks. The success of these models stems from their ability to transfer information learned during the initial training period to new prediction problems through approaches like supervised fine-tuning—adjusting a model's parameters to minimize prediction error on labeled examples from a target task (1). For example, large language models (1, 2) are foundation models that were originally trained to predict the next word of Internet articles, but can be fine-tuned to make other predictions involving text, like the next word of a conversation or the sentiment of a movie review.

While foundation models have been successful at making predictions about social science data (4, 5), many core problems in social science require more than just accurate predictions. For example, social scientists often aim to estimate causal effects under the assumption of unconfoundedness (6) or decompose observed differences between groups into explained and unexplained components based on observable factors (7–9)—isomorphic problems that use prediction as an intermediate step but ultimately require different criteria for success. While fine-tuning foundation models may be useful for these analyses, optimizing for predictive accuracy alone does not guarantee valid decompositions or causal estimates.

In this paper, we develop methods for adapting foundation models to perform decomposition and causal effect estimation by modifying how they are fine-tuned. Rather than fine-tuning foundation models to minimize prediction error, we develop objectives specifically designed for these estimation problems. Our first contribution is characterizing a statistical bias that arises when a foundation model discards information that may not be important for prediction but is relevant for the estimation problem. We then provide a set of conditions for fine-tuning under which estimates derived from a foundation model are not only unbiased but also consistent at a fast asymptotic rate. These conditions motivate debiased fine-tuning methods. Our key insight is that

Significance

Understanding differences in outcomes between social groups—such as wage gaps between men and women—remains a central challenge in social science. While researchers have long studied how observable factors contribute to these differences, traditional methods oversimplify complex variables like employment trajectories. Our work adapts recent advances in artificial intelligence—specifically, foundation models that can process rich, detailed histories—to better explain group differences. We develop mathematical theory and computational methods that allow these AI models to provide more accurate and less biased estimates of how much of group differences can be explained by observable factors. Applied to real-world data, our approach reveals that detailed histories explain more of the gender wage gap than previously understood using conventional methods.

Author contributions: K.V., S.A., and D.M.B. designed research; performed research; contributed new reagents/analytic tools; analyzed data; and wrote the paper.

Reviewers: E.A., Eidgenossische Technische Hochschule Zurich; and W.N., Massachusetts Institute of Technology.

The authors declare no competing interest.

[1] To whom correspondence may be addressed. Email: athey@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2427298122/-/DCSupplemental.

Published May 30, 2025.

fine-tuning foundation models for these applications requires addressing an omitted variable bias that standard supervised fine-tuning does not address.

To demonstrate these ideas, we focus on an application that addresses a classic decomposition problem from labor economics: estimating the difference between how individuals with the same labor market experience get paid when they belong to different demographic groups (see refs. 10 and 11 for reviews). Accurately estimating this unexplained wage gap is important to help guide policy for reducing disparities. But the unexplained gap is challenging to estimate with traditional econometric models. It involves predicting an individual's wage from their labor market history, a high-dimensional and complicated variable. Our paper demonstrates that foundation models of labor market history can improve the predictions that underlie wage gap estimates.

We use CAREER, a foundation model of labor market history (4), to estimate unexplained wage gaps. CAREER is initially fit to a massive resume dataset to predict the next job an individual will have, rather than their wage. Naively, we can fine-tune CAREER to make accurate predictions of wage on the datasets used for wage gap estimation. However, using this approach to estimate the unexplained wage gap can amplify a classical problem: omitted variable bias. Instead, we develop debiased fine-tuning methods to fine-tune foundation models so they can properly estimate unexplained wage gaps. The key is to fine-tune foundation models not to minimize predictive error but rather to reduce omitted variable bias. In synthetic experiments, we show that debiased fine-tuning methods form better estimates of the unexplained wage gap than the standard fine-tuning approach.

We use our methods to estimate the explained gender wage gap on survey data from the Panel Study of Income Dynamics (PSID) (12). We first demonstrate that foundation models form accurate predictions of wage and gender; they outperform standard econometric models for predicting wage by 10 to 15%. We then use debiased fine-tuning methods to estimate the gender wage gap. We find that history consistently explains more of the gap than the variables typically included in standard econometric models. We conclude by studying which aspects of work history, captured by foundation models but omitted from prior approaches, are important for explaining the wage gap.

While this paper studies unexplained wage gaps in detail, the results and methods we develop are applicable to a broader set of problems, such as causal estimation. In particular, as observed by Fortin et al. (13) and others in the literature, the problem of estimating a decomposition of a wage gap into explained and unexplained components is isomorphic to the problem of estimating the average effect of a treatment under the assumption of unconfoundedness. Although the interpretation of the estimate is distinct for decompositions, the statistical theory that applies to estimation is the same (see ref. 14 for a review). Thus, our results also provide theory and methods for the problem of incorporating foundation models into the estimation of treatment effects.

Relative to both the causal inference and decomposition literatures, our theory is adapted to a scenario where a foundation model may bring in information from a distinct, larger dataset, and where we fine-tune the model to avoid omitted variable bias. If we solve the latter problem well enough, then the traditional semiparametric theory (e.g., ref. 15) can be applied as if the representations of high-dimensional covariates derived from the fine-tuned foundation model are sufficient statistics for the full high-dimensional covariate vector. The methods we introduce thus provide a widely applicable framework for leveraging the capabilities of foundation models while mitigating biases due to omitted variables that they may introduce.

1. Explaining Wage Gaps with Foundation Models

The unexplained wage gap is the wage gap between two groups of individuals with the same observed characteristics. We estimate an unexplained wage gap that arises when individuals in different groups have the same labor market history.

Consider the gender wage gap. In the United States, females earn roughly 80% the male hourly wage (10). Motivated by the fact that the male and female labor forces differ in observable ways, a large literature seeks to explain this wage gap through differences in these observable factors (7, 8, 10, 16). One of the most important factors for explaining the gender wage gap is differences in the number of years that males and females have spent in the labor force (10). Understanding the difference in wages between males and females with the *same career histories* can help guide policy: if the unexplained gap is large, attempts to close the gap may involve interventions to address problems in bargaining or fairness in wage setting. On the other hand, if the gap can be accounted for by gender differences in career histories, these interventions might target career pathways.

More generally, consider N individuals, indexed by $i = 1, \ldots, N$. Each individual belongs to a binary group $A_i \in \{0, 1\}$ (e.g. $A_i = 0$ denotes males and $A_i = 1$ denotes females). Each individual also has a career history X_i, a sequence of T discrete occupations and years, $X_i = ((J_{i1}, D_{i1}), \ldots, (J_{iT}, D_{iT})) \in \mathcal{X}$, and where each occupation label $J_{it} \in \{1, \ldots, N_J\}$ encodes the occupation an individual worked in during year D_t (or their labor status if they are not working, e.g. "unemployed" or "student"). Finally, denote an individual's log-wage by $Y_i \in \mathbb{R}$. Each individual is sampled i.i.d. from a joint distribution $P(X, A, Y)$. Define the conditional expectation function, $\mu_a(x) = \mathbb{E}_P[Y|A = a, X = x]$, and the propensity function, $e(x) = P(A = 1|X = x)$.

The raw **wage gap** is

$$\mathrm{WG} = \mathbb{E}_{p(x|a=1)}[\mu_1(X)] - \mathbb{E}_{p(x|a=0)}[\mu_0(X)].$$

This is the difference in the average wage between the two groups. Our goal is to estimate the wage gap that is **unexplained** by history:

$$\mathrm{UWG} = \mathbb{E}_{p(x|a=1)}[\mu_1(X) - \mu_0(X)].$$

This is the average difference in the expected wage between individuals in the two groups who have the same career histories. The unexplained and raw wage gaps are linked by a classic decomposition (7, 8, 16):

$$\begin{aligned}\mathrm{WG} = &\overbrace{\mathbb{E}_{p(x|a=1)}\left[\mu_1(X) - \mu_0(X)\right]}^{\text{unexplained wage gap}} \\ &+ \underbrace{\mathbb{E}_{p(x|a=1)}\left[\mu_0(X)\right] - \mathbb{E}_{p(x|a=0)}\left[\mu_0(X)\right]}_{\text{explained wage gap}}.\end{aligned} \quad [1]$$

The unexplained wage gap is the portion of the raw wage gap that cannot be attributed to gender differences in career histories.

The explained wage gap could be conditioned on factors in addition to labor market history, e.g., an individual's educational background. For simplicity our notation only includes history, but we incorporate additional observed characteristics of individuals in our empirical analyses. We also note that the unexplained wage gap is only nonparametrically identifiable under an overlap condition (see, e.g., ref. 6): $P(A = 1|X) < 1$. We assume overlap throughout this paper and we limit the sample of workers we analyze empirically to those with histories where the condition is satisfied.

1.1. Foundation Models Can Improve Predictions. A common approach for estimating the unexplained wage gap involves constructing an estimator $\hat{\mu}_g(x)$ for $\mu_g(x)$, which in turn can be used to form an estimate of the unexplained wage gap:

$$\frac{1}{N_1} \sum_i A_i * (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)), \qquad [2]$$

where $N_1 = \sum_i A_i$.

However, estimating the relationship between history and wage is challenging with realistic data sizes because career histories are high-dimensional: the number of possible career histories grows exponentially in the number of years someone has worked. This challenge is compounded by the fact that unexplained wage gaps are commonly estimated using small survey-based datasets, particularly in the United States where administrative data about worker histories are not generally available to researchers. For this reason, traditional econometric approaches have used a small number of hand-constructed summaries of labor market experience, which keeps the number of covariates in the predictive model small relative to the dataset size. But these summary statistics do not capture the full complexity of labor market history, and in particular they may omit factors of history that are important for explaining the wage gap.

For example, a large body of literature has focused on decomposing the gender wage gap in the United States by applying Eq. **1** to small survey datasets (see refs. 10 and 11 for surveys). Rather than including an individual's career history, most analyses include summary statistics about an individual's career history, such as the years of experience or tenure in the current job (10, 17–19). Even though many occupational taxonomies contain hundreds of fine-grained categories, it is most common to include coarse-grained occupational categories containing 20 to 30 categories (10, 20–22). Because these models rely on a relatively small number of covariates, $\hat{\mu}_0$ and $\hat{\mu}_1$ are typically constructed using relatively simple models, such as linear regressions (8–10, 20) or Lasso models (21, 23).

However, these incomplete measures of experience discard factors that help explain the wage gap. For example, Regan and Oaxaca (24) and Blau and Kahn (19) find that potential experience (an inexact measure of experience that does not measure workforce interruptions) explains less of the wage gap than years of actual work experience. Moreover, Light and Ureta (25) estimate a wage model with detailed measures of year-by-year experience, finding that the timing of work experience explains a substantial portion of the wage gap. While incorporating full histories into gender wage gap analyses could ensure these factors are not discarded, the predictive models used for gender wage gap analyses are too simple to include them.

Foundation models (3) offer an alternative approach. Foundation models are machine learning models that *learn* low-dimensional representations of high-dimensional variables from data. These representations are initially learned on massive, passively collected data after which they can be adapted on specific datasets of interest. For example, in natural language processing, foundation models that are trained to predict words using terabytes of Internet text can be adapted to generate responses to human questions (26, 27). While initially developed for text, foundation models have successfully addressed seemingly intractable prediction problems in domains such as computer vision (28), music (29), and protein generation (30)

A foundation model of labor market history can help estimate the unexplained wage gap by providing a low-dimensional representation of history that is predictive of wage. Because representations are learned from data, these estimates are not limited to the features a researcher knows to include.

Formally define a representation to be a function $\lambda(X) : \mathcal{X} \to \mathbb{R}^D$. Given a representation λ, the wage gap unexplained by the representation of history is

$$\text{UWG}(\lambda) = \mathbb{E}_{p(x|a=1)}[\mu_1(\lambda(X)) - \mu_0(\lambda(X))],$$

where $\mu_a(\lambda(x)) = \mathbb{E}[Y|A = a, \lambda(X) = \lambda(x)]$ is the expected wage as a function of the representation $\lambda(x)$.

For the rest of the paper, we consider estimating the unexplained wage gap using CAREER, a foundation model of labor market history (4). CAREER is trained to learn representations that can predict the next occupation a worker will have from a dataset of 24 million resumes posted online. When these representations are adapted to small survey datasets, CAREER makes more accurate predictions of an individual's next occupation than standard econometric approaches. We consider using these representations to predict an individual's wage.

1.2. Foundation Models Can Introduce Omitted Variable Bias. Foundation models are effective because they compress high-dimensional information into low-dimensional representations. However, we demonstrate that replacing an individual's history with a representation can introduce an omitted variable bias (31).

We say the wage gap unexplained by a representation λ is biased if it differs from the wage gap unexplained by the full history. Define this bias as $\text{OVB}(\lambda) = \text{UWG}(\lambda) - \text{UWG}$. It has a closed-form expression:

$$\text{OVB}(\lambda) = \mathbb{E}_{p(x,a)}\big[(\mu_A(\lambda(X)) - \mu_A(X)) * (\alpha_A(\lambda(X)) - \alpha_A(X))\big], \qquad [3]$$

where

$$\alpha_a(x) = -\frac{1-a}{P(A=1)} \left(\frac{e(x)}{1-e(x)}\right),$$
$$\alpha_a(\lambda(x)) = -\frac{1-a}{P(A=1)} \left(\frac{e(\lambda(x))}{1-e(\lambda(x))}\right),$$

for representation-based propensity function $e(\lambda(x)) = P(A = 1|\lambda(X) = \lambda(x))$. *SI Appendix*, section S1 contains a detailed derivation. Eq. **3** can also be seen as a special case of the general omitted-variable-bias formula in ref. 31, where $\alpha(\lambda(X))$ and $\alpha(X)$ correspond to the short and long Riesz representers, respectively, as defined in equation 6.1 of ref. 31.

Eq. **3** provides intuition for how a representation can induce bias. The omitted variable bias is a covariance of two differences: the first term is the difference in expected wage as a function of history and the representation of history, while the second term is the difference in the group propensity odds ratio as a function of history and the representation of history. A low-dimensional representation by definition discards information; for there to be no omitted variable bias, the discarded information that is related to wage should be unrelated to group propensity, and vice-versa. Eq. **3** is closely related to econometric results about the extent of omitted variable bias in semiparametric models (see, e.g., ref. 31); while this literature focuses on whether individual variables are included or not in models, we focus on representations, which can still omit variables despite being functions of all variables. Focusing on representations, Veitch et al. (32) provide a sufficient condition under which Eq. **3** is 0, but do not characterize the exact level of bias; meanwhile, Eq. **3** exactly characterizes the extent of omitted variable bias.

1.3. Debiasing Foundation Models.

Although CAREER is a foundation model that is trained to learn representations from data, it is not trained to minimize omitted variable bias (Eq. **3**). One reason is that its representations are trained to optimize a single objective that does not naturally appear in Eq. **3**: the predictability of an individual's next job. Moreover, the representations are trained on a different population of individuals than those for whom we would like to estimate the unexplained wage gap.

Even biased, a foundation model can still be useful for estimating the unexplained wage gap because it can be fine-tuned. Empirically, when a foundation model's representations are adjusted to optimize a related but distinct objective from the one they were initially trained to optimize, they often outperform models trained on only the new objective (1, 33). We do not have to learn unbiased representations of career history from scratch; we can adjust the representations of a pretrained foundation model to debias it.

The standard approach for modifying foundation models is **supervised fine-tuning** (1). In our setting, supervised fine-tuning would entail modifying a foundation model's representation λ to be predictive of wage on the survey data used for wage gap estimation. But while a foundation model would likely form better wage predictions after supervised fine-tuning, it can still be biased for estimating the unexplained wage gap; unless the foundation model recovers the exact relationship between labor market history and wage, supervised fine-tuning can still introduce arbitrarily large omitted variable bias.

We now describe a set of conditions for fine-tuning under which an estimator of a wage gap that conditions on representations derived from a foundation model is not only unbiased and consistent but also converges at a rate proportional to $n^{-1/2}$:

Theorem 1. *Consider a sequence of wage models $\hat{\mu}_{n,0} : \mathbb{R}^D \to \mathbb{R}$, propensity models $\hat{e}_n : \mathbb{R}^D \to (0, 1)$, and representations $\lambda_n : \mathcal{X} \to \mathbb{R}^D$. Denote by ψ the true wage gap unexplained by history and by $\hat{\psi}_n$ the representation-based augmented inverse probability weighted (AIPW) estimator of the unexplained wage gap from n i.i.d. samples $(X_i, A_i, Y_i) \sim P$:*

$$\hat{\psi}_n = \tfrac{1}{\sum_i A_i} \sum_i \left(A_i - \tfrac{(1-A_i)\hat{e}_n(\lambda_n(X_i))}{1-\hat{e}_n(\lambda_n(X_i))} \right) (Y_i - \hat{\mu}_{n,0}(\lambda_n(X_i)). \quad \textbf{[4]}$$

Assume the following:

1. *Omitted variable bias (Eq. **3**) goes to 0 at a $\sqrt{n}$-rate:*
$$OVB(\lambda_n) = o_P(n^{-1/2}).$$
2. *Combined $\sqrt{n}$-consistency of wage/propensity models as a function of the representation:*
$$\big(\|\hat{e}_n(\lambda_n(X)) - e(\lambda_n(X))\| \\ * \|\hat{\mu}_{n,0}(\lambda_n(X)) - \mu_0(\lambda_n(X))\|\big) = o_P(n^{-1/2}).$$
3. *The representations λ_n converge to a representation λ^* in the sense that*
$$\tfrac{1}{n} \sum_i (\varphi_{\lambda_n}(X_i, A_i, Y_i; \psi_{\lambda_n}) - \varphi_{\lambda^*}(X_i, A_i, Y_i; \psi_{\lambda^*})) = o_P(n^{-1/2}),$$
with $Var(\varphi_{\lambda^*}(X, A, Y)) < \infty$, where φ_λ is the representation-based influence function:
$$\varphi_\lambda(X, A, Y; \psi_\lambda) = \tfrac{1}{P(A=1)} \Bigg[\left(A - \tfrac{(1-A)e(\lambda(X))}{1-e(\lambda(X))} \right) \\ * (Y - \mu_0)(\lambda(X)) - A\psi_\lambda \Bigg]$$
and ψ_λ is the true gap unexplained by a representation λ
$$\psi_\lambda = \mathbb{E}_P[\mu_1(\lambda(X)) - \mu_0(\lambda(X))].$$
4. *Additional assumptions in SI Appendix, section S2: cross-fitting ($\hat{\mu}_n$, $\hat{e}_n$, and λ_n are estimated on a different sample than those used to construct $\hat{\psi}$); consistency of wage and propensity models as functions of the representations; strict overlap; and boundedness of wage model errors.*

Then,

$$\sqrt{n}(\hat{\psi}_n - \psi) \to \mathcal{N}\left(0, Var(\varphi_{\lambda^*}(X, A, Y; \psi))\right).$$

The first condition is about omitted variable bias: it requires that the omitted variable bias of the representations converges to 0 at a rate proportional to $n^{-1/2}$. This will be trivial for some representations: for example, $\lambda(X) = X$ has no omitted variable bias by definition. However, the second assumption imposes restrictions about modeling wage and group membership: these models must approximate the true relationship between the representation of history and these outcomes such that error goes to zero at a combined root-n rate. Note that these modeling assumptions are with respect to the representation: the true relationships between history and the outcomes do not need to be reconstructed, only those between the representation and the outcome. Therefore, satisfying the first two assumptions involves striking a balance: representations must be detailed enough to not have omitted variable bias, but also low-dimensional enough so that outcomes can be efficiently estimated as a function of the representation. SI Appendix, section S2 contains more details and a proof.

Although Eq. **1** is stated in terms of the size n of the single dataset used for fine-tuning, note that much larger datasets are typically used to pretrain the foundation model. In practice, training a high-dimensional representation $\lambda(X)$ from scratch on moderate-sized survey dataset would be intractable; however, the fact that a much larger dataset contributes to the initial estimation of $\lambda(X)$ suggests that it may be feasible to adequately control omitted variable bias with a larger dimensional representation $\lambda(X)$ than would be possible without the foundation model. Indeed, under a repeated-sampling framework in which both the pretraining and fine-tuning samples are repeatedly resampled, we expect most of the sampling variation to arise from the smaller fine-tuning sample rather than from the larger pretraining sample used to train the foundation model. Consequently, our formal results condition on the pretrained representation and focus on the sampling variation arising from the fine-tuning sample.

Empirical evidence in domains such as computer vision, natural language processing, and protein structure suggests that such large-scale pretraining often yields robust and transferable representations for predictions (3). At the same time, we do not require the foundation model to deliver a perfect or "true" representation; we instead require that fine-tuning from the pretrained representation allows for omitted variable to go to 0 at a $\sqrt{n}$ rate. A general-purpose foundation model may still omit some information relevant for a particular downstream task, and it is precisely the role of subsequent fine-tuning to adapt the representation to the problem at hand.

1.4. Relationship to Causal Methods.

Theorem 1 relates to results from the causal inference literature (15, 34–37). Although the unexplained wage gap is not a causal quantity, it is mathematically identical to an average treatment effect on the treated (ATT). Specifically, Assumption 2 is similar to assumptions for $\sqrt{n}$-consistency in the doubly robust/double machine learning literature, which often assume $o(n^{-1/2})$ combined error of outcome and propensity models (15, 37):

$$\|\hat{e}_n(X) - e(X)\| * \|\hat{\mu}_{n,0}(X) - \mu_0(X)\| = o_P(n^{-1/2}). \quad [5]$$

This is similar to Assumption 2, with one key difference: Eq. **5** requires $o(n^{-1/2})$ combined error as a function of the *full history*, while Assumption 2 only requires $o(n^{-1/2})$ combined error *as a function of the representation* $\lambda_n(X)$. When $\lambda_n(X)$ is lower-dimensional than the full history X, Assumption 2 will be more realistic than Eq. **5**. In fact, when $\lambda_n(X) = X$, Assumptions 2 and 3 are trivially satisfied and our result reduces to the standard double-robustness result (15). Assumption 2 lifts the requirement for full combined error to go to 0 while Assumption 1 imposes restrictions on what constitutes a valid representation.

Theorem 1 relates to results from the variable selection literature in causal inference (38–41). This literature is motivated by a classic result: to make valid causal inferences, it is sufficient to condition only on variables that affect both treatment assignment and outcome (42). Like Theorem 1, the results in this literature do not necessarily assume that the full outcome or propensity model can be consistently estimated as a function of the full set of covariates. While this literature proposes techniques when individual variables are shared in outcome and treatment models, these techniques do not apply when there is a more complicated shared structure; for example, the number of years spent in a blue collar job may affect both treatment and outcome, but this is a transformation rather than a single variable. In contrast, our method is based on representations, or potentially complicated functions of variables, rather than individual variables. A set of selected variables is an example of a representation; but representations can be more complex than a set of variables constructed by a researcher.

Other methods from the causal inference and econometrics literature have also proposed using representations or latent variables from machine learning models. For example, Battaglia et al. (43) demonstrate that latent variables from a machine learning model should be jointly optimized with the econometric outcome of interest rather than first estimated separately and then plugged into an econometric model. Related to our method, Veitch et al. (32) provide a sufficient condition under which a representation is unbiased for estimating a causal effect, which motivates empirical methods used by Shi et al. (44) and Chernozhukov et al. (45) (and is the basis of the multitask debiased fine-tuning objective we consider). In contrast, we provide an if-and-only-if condition under which there is no bias, and we characterize the exact level of bias with a connection to omitted variable bias (31). Additionally, we provide conditions about the level of omitted variable bias under which estimation is $\sqrt{n}$-consistent and asymptotically normal.

A strand of literature in supervised machine learning has also focused on integrating ideas from the causal inference literature into predictive methods in order to improve the properties of predictive models, such as stability (see ref. 46 for a review of this literature). Similar to the approach in our paper, these methods adjust the training of a predictive model to avoid regularization-induced omitted variable bias, but the literature on stable prediction considers reducing such bias for many covariates simultaneously in a cross-sectional prediction problem. For example, some such methods reweight data to reduce the correlation among features.

1.5. Debiased Fine-Tuning. The exact level of omitted variable bias (Eq. **3**) cannot be computed from data; it involves calculating the same high-dimensional function the representation is meant to approximate, $\mu_A(X)$. However, even if the bias cannot be computed exactly, we can still learn representations that are targeted to minimize it. Below, we develop three fine-tuning methods for minimizing this bias. Each method addresses omitted variable bias from a distinct angle. In principle, the optimal approach may vary across applications. For instance, multitask fine-tuning is straightforward to implement but can require tuning an extra hyperparameter; projection fine-tuning removes that hyperparameter but may converge more slowly; difference-based fine-tuning can capture group disparities more directly but does not deliver a direct wage or propensity predictor. To choose among them, we recommend using validation metrics such as the R-Learner metric (47) described (and subsequently used for model selection) in Section 3.

1.5.1. Multitask fine-tuning. The expression for omitted variable bias recalls a classic result from causal inference: to make valid causal inferences, it is sufficient to condition only on variables that affect both treatment assignment and outcome (42). Thus, if a representation captures all the features of history that are predictive of *both* wage and group membership, it will result in zero omitted variable bias. While a representation that perfectly captures all of the variables related to wage (or equivalently to group membership) will result in zero omitted variable bias, it might be more realistic to capture the potentially smaller set of variables that affects both wage and group membership.

We therefore fine-tune the representations to be predictive of both wage and group membership. We consider two approaches. In **Multitask Fine-Tuning**, we use the method proposed by Veitch et al. (32) and (44) to learn a representation that jointly minimizes wage and group membership predictive errors:

$$\hat{\lambda}, \hat{\mu}, \hat{e} = \arg\min_{\lambda,\mu,\lambda} \Big\{ \textstyle\sum_{i=1}^{N} \ell_Y[Y_i, \mu_{A_i}(\lambda(X_i))] \\ + \beta * \ell_A[A_i, e(\lambda(X_i))]\Big\},$$

where $\beta \in \mathbb{R}^+$ is a hyperparameter, ℓ_Y is the mean squared-error loss, and ℓ_A is the binary cross-entropy loss. We also consider a similar approach, **Projection Fine-Tuning**, that alternates between losses: λ is optimized to minimize mean-squared error loss until convergence, then λ is optimized to minimize binary cross-entropy loss until convergence, and this process repeats until the procedure converges. This procedure is based on projected gradient descent (48) and does not require choosing a hyperparameter β. See *SI Appendix*, section S4 for more details.

1.5.2. Difference-based fine-tuning. An alternative debiased fine-tuning method is motivated by the heterogeneous treatment effect literature (47, 49, 50), where the goal is to model the difference in outcome functions for each group rather than each group's function individually. If the difference in these functions is simpler than each individual function, a method that is targeted to capture this difference will be more effective than modeling each function separately.

Thus, we propose a method for fine-tuning foundation models meant to capture group differences. This method is based on the R-learner approach for estimating heterogeneous treatment effects in causal inference (47, 51). It is based on the observation that group differences can be written as the solution to an objective:

$$\arg\min_{\tau:\mathcal{X}\to\mathbb{R}} \mathbb{E}\left\{[(Y - m(X)) - (A - e(X))\tau(X)]^2\right\} \\ = \mu_1(X) - \mu_0(X), \quad [6]$$

where $m(x) = \mathbb{E}[Y|X = x]$ is the conditional wage function *averaged* over the two groups.

The **Difference-Based Fine-Tuning** procedure begins by estimating the conditional wage function $\hat{m}(\lambda_m(x))$ and the propensity function $\hat{e}(\lambda_e(x))$ using supervised fine-tuning. Starting with the wage, we estimate a function $\hat{m}$ and fine-tuned representation λ_m in order to minimize the squared loss in predicting the wage Y. We estimate $\hat{e}$ and λ_e analogously. Treating these functions and their respective representations as fixed, we then fine-tune a new representation to minimize the squared error:

$$\hat{\lambda}, \hat{\rho} = \arg\min_{\lambda,\rho} \mathbb{E}\left\{[(Y - \hat{m}(\lambda_m(X))) - (A - \hat{e}(\lambda_e(X)))\rho(\lambda(X))]^2\right\},$$

where $\rho : \mathbb{R}^D \to \mathbb{R}$ is a flexible function; in the empirical studies we use a two-layer feed-forward neural network. The unexplained wage gap is then estimated as

$$\frac{1}{N_1}\sum_i A_i * \hat{\rho}(\hat{\lambda}(X_i)), \qquad [7]$$

where $N_1 = \sum_i A_i$. We refer to this method as *difference-based fine-tuning*. This approach is based on the R-learner approach but is based on representations: separate representations are used for the wage, propensity, and wage-difference models, so that the representation used for the last model is optimized to only capture the differences in groups. See *SI Appendix*, section S4 for more details.

This optimization procedure encourages a representation that captures differences in group wages. The true relationship between wage and labor market history may be complicated for both groups. However, if the *difference* in relationships is not as complicated, it will be easier to learn a representation that captures the difference.

1.6. Implementation details of fine-tuning. In practice, we implement each fine-tuning method by optimizing the respective objective using Adam (52). Each model is initialized at the foundation model's parameters and all parameters are reoptimized with respect to the new objective. The same neural architecture from pretraining (e.g., the transformer layers of CAREER) is retained, but we typically add extra parameters on top of the representation for our downstream tasks (e.g., predicting wages or propensities) and jointly update all parameters during fine-tuning. These steps enable the representation to adapt to the new objectives while preserving its broad, pretrained knowledge. See *SI Appendix*, section S4 for more details.

2. Data

Our empirical analysis uses data from one of the leading U.S. administrative surveys, the PSID (12). PSID is a longitudinal survey that has followed a cohort of American families since 1968. It is constructed to be nationally representative and is frequently used to estimate unexplained wage gaps (10, 53). Because the same individuals are interviewed over the course of the survey, labor market histories can be constructed by tracking the trajectory of reported occupations each year an individual is in the survey.

We encode occupations into one of 330 "occ1990dd" occupational categories (54). Since the PSID includes information about individuals who are not working, we add seven categories for when an individual's occupation is not listed but their employment status is available (e.g., employed, laid off).

Following ref. 10, we restrict our sample to the surveys conducted between 1990 and 2019, consisting of 91,391 observations over 19 surveys, and further restrict our sample to nonfarm and nonmilitary wage and salary workers between 25 and 64 y old who worked for at least 26 wk in nonfarm jobs. We incorporate longitudinal sample weights into our analysis, which are designed to adjust for differences in the probability of selection into the sample. *SI Appendix*, section S3 contains more details about how we construct the dataset.

3. Synthetic Experiments

We first validate our proposed methods for debiased fine-tuning. Typically, machine learning models are evaluated using measures of predictive accuracy on held-out data. However, our ultimate goal is not forming more accurate wage predictions, but rather more accurate estimates of the unexplained wage gap. To this end, we turn to synthetic experiments, a common method to assess causal estimation strategies in a controlled setting (e.g., ref. 55). Synthetic experiments allow us to evaluate the performance of different methods because the data generating process is known.

In order for the synthetic experiments to reflect real-world data, we use real labor market histories from the PSID sample (Section 2). Our synthetic experiments are based on forming a "ground-truth" representation of the history and then generating group labels and wages as a function of this history. To mimic the fact that nature is often more complicated than the model we might select, we use a more complicated representation to generate data than the one used by the foundation model to estimate wage gaps. Specifically, we use a transformer architecture that is 20 times larger than the one used for estimation to simulate a ground-truth representation $\lambda^* : \mathcal{X} \to \mathbb{R}^D$. For each setting, we simulate group labels and wages as a function of the representation λ^*. We control how much of the representation is shared between these functions by introducing binary variables $\mathbf{u}, \mathbf{v} \in \{0, 1\}^D$ that mask the dimensions of the representation used for the group and wage models, accordingly. We then simulate from the following model:

$$A_i \sim \text{Bern}\left(\sigma\left(\sum_j u_j\beta_j * \lambda(X_i)_j\right)\right)$$
$$Y_i = \tau A_i + \sum_j v_j\beta_j * \lambda(X_i)_j + \epsilon_i,$$

where $\beta \in \mathbb{R}^D$ is a random vector of regression coefficients, $\tau \in \mathbb{R}$ is the true unexplained gap, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the outcome noise, and $\sigma(\cdot)$ is the inverse-logit function. In this setup, $\mathbf{u}^\top\mathbf{v}/D$ is the proportion of the representation that is shared. For each experiment, we control the shared proportion ($\mathbf{u}^\top\mathbf{v}/D$), the true gap ($\tau$), and the level of outcome noise (σ^2). We consider 27 different settings, and perform multiple samples in each setting by resampling A_i, ϵ_i, $\mathbf{u}$, $\mathbf{v}$, and β. See *SI Appendix*, section S5 for more details.

We compare four methods for estimating the unexplained wage gap from synthetic data. Our baseline is *Supervised Fine-Tuning* (1): fine-tuning a foundation model to predict wage without an explicit debiasing objective. We compare this baseline to the three debiasing methods described above: Multitask Fine-Tuning, Projection Fine-Tuning, and Difference-Based Fine-Tuning. We perform 400 simulations, regenerating data and retraining each method for each simulation. Here, we estimate the unexplained gap using the outcome-only estimator (analogous

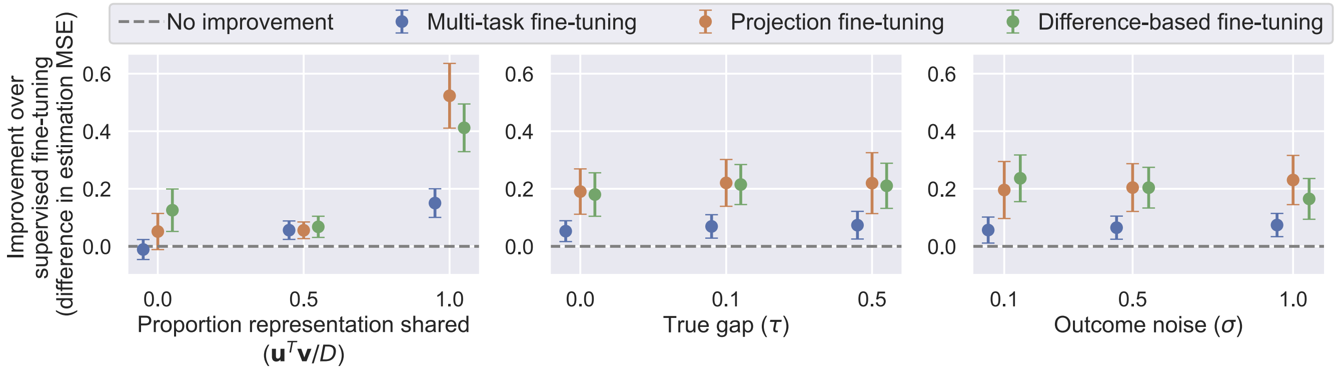

Fig. 1. Debiased fine-tuning methods are better at estimating the unexplained wage gap than standard supervised fine-tuning (1) across 270 synthetic experiments. For each synthetic experiment, the true unexplained wage gap is known, and each method provides a different estimate of this gap. This figure compares each method's average error for estimating this gap, evaluated via MSE between the true and estimated unexplained gap. Specifically, the Y-axis compares each method's estimation error to the error from estimates derived from a model using standard supervised fine-tuning (larger values on the Y-axis correspond to larger improvements). Bars represent 95% CI.

to Eq. **2**), while *SI Appendix*, section S5 shows results for the AIPW estimator (Eq. **4**), which we find to perform slightly worse in practice. Although AIPW is theoretically advantageous in large samples (34, 35), it can suffer from higher variance when propensity models are difficult to estimate, thereby degrading its finite-sample performance (56).

Fig. 1 compares the MSE of the estimate of the gender wage gap (relative to the oracle gender wage gap) derived from alternative estimation approaches. All three methods for debiasing foundation models consistently outperform the standard supervised fine-tuning approach. The advantage of debiasing is largest when more of the representation is shared across the wage model and group labels. This reinforces the motivation behind representation learning; as there is more shared structure in how group labels and wages relate to history, sharing representations can improve estimates. Projection fine-tuning and difference-based fine-tuning are both more successful than multitask fine-tuning, especially when more of the representation is shared. The full set of results is in *SI Appendix*, Table S6.

How should we validate models on real-world data? While wage and gender predictive metrics are important, they do not directly assess estimation quality. While matching-based methods (57) can also be used to validate estimation in principle, these require low-dimensional covariates. We instead consider another validation metric, inspired by the R-Learner objective in Eq. **6**. Because Eq. **6** is minimized when $\tau(X)$ is the true expected difference female and male wages, we evaluate Eq. **6** given a model's estimate of $\hat{\tau}(X) = \hat{\mu}_1(X) - \hat{\mu}_0(X)$,

$$\frac{1}{n}\sum_{i=1}^{n}[(Y_i - \hat{m}(X_i)) - (A_i - \hat{e}(X_i))\hat{\tau}(X_i)]^2, \qquad [8]$$

where $i = 1, \ldots, n$ index n held-out samples and $\hat{m}(X_i)$ and $\hat{e}(X_i)$ are models trained to predict wage and gender, respectively, using supervised fine-tuning of CAREER. We refer to Eq. **8** as the **R-Learner Metric**. We note that because this metric relies on estimates of wage and propensity models, it is sensitive to the specification of $\hat{m}$ and $\hat{e}$. However, we validate this metric on synthetic data, finding it to be a useful proxy for model performance (*SI Appendix*, Fig. S2).

4. Empirical Application

We now apply our methods to the (actual) PSID data. We begin by evaluating the quality of wage predictions derived from alternative models, including the standard econometric models from the wage gap literature (10), since predictive accuracy can be easily evaluated using held-out test data. We show that foundation-based representations substantially improve predictive performance relative to standard regression-based econometric models, suggesting that our methods can capture variables that have the potential to cause omitted variable bias when estimating unexplained wage gaps. We then directly demonstrate that representations derived from our methods capture elements of history that are predictive of *both* wage and gender, so that they indeed meet the criteria for omitted variable bias, and that these are quantitatively important for explaining wage gaps.

4.1. Predictive Accuracy. As baselines, we consider econometric models that use hand-constructed summaries of an individual's career but not their full history to predict wage. Following the econometric literature, we consider two linear models: regression (fit with OLS) and LASSO. Given covariates $Z_i \in \mathbb{R}^P$, these models estimate the wage function as

$$\hat{\mu}_A(Z_i) = \theta_A + \beta_A^{\top} Z_i, \qquad [9]$$

for $A \in \{0, 1\}$, an intercept $\theta_A \in \mathbb{R}$, and regression coefficients $\beta_A \in \mathbb{R}^P$. Following ref. 10, the covariates included in Z are years of full-time and part-time experience (and their squares), years of schooling, indicators for bachelors and advanced degrees, race and ethnicity indicators, gender indicators, census and region indicators, an indicator for collective bargaining coverage, 15 industry category indicators, and 21 occupation category indicators. We also consider two different methods of encoding occupations: "**coarse-grained**," which uses the 21 coarse-grained occupational categories above, and "**fine-grained**," with an additional 330 fine-grained occupational categories.

We compare these models from the economics literature to predictions based on foundation models that use CAREER (4) to represent labor market history. CAREER is pretrained to learn representations of career trajectories on a dataset of 23.7 million resumes. We consider both supervised fine-tuning (1) and debiased fine-tuning approaches to modify CAREER's representations. When we fine-tune CAREER, we use both its representations of history and the covariates Z_i described above to predict an individual's wage; see *SI Appendix*, section S4 for more details. In order to understand how different methods of including history affect predictions, we train two additional versions of CAREER: one that uses the neural network

Table 1. The CAREER foundation model, which incorporates an individual's full labor market history, forms better predictions of wage and gender on held-out data than standard econometric methods, which summarize history with low-dimensional statistics

		Wage R^2	Gender R^2
Regression models	Coarse-grained regression	0.428 (0.004)	0.137 (0.002)
	Coarse-grained LASSO	0.428 (0.003)	0.260 (0.003)
	Fine-grained LASSO	0.455 (0.003)	0.314 (0.003)
Foundation models (supervised fine-tuning)	CAREER (no pretraining)	0.462 (0.003)	0.424 (0.004)
	CAREER (pretrained, current job only)	0.454 (0.004)	0.307 (0.003)
	CAREER (pretrained, participation only)	0.467 (0.003)	0.309 (0.004)
	CAREER (pretrained)	0.515 (0.004)	0.510 (0.004)
Foundation models (debiased fine-tuning)	CAREER (multitask fine-tuning)	0.468 (0.004)	0.514 (0.005)
	CAREER (projection fine-tuning)	0.503 (0.003)	0.513 (0.004)

Test-set bootstrapped SE are in parentheses.

to encode an individual's current job but not their history ["CAREER (current job only)"] and one that includes an individual's current job but only their workforce participation status for previous jobs (e.g., "unemployed," "out-of-labor force"), which we refer to as "CAREER (participation only)."

Table 1 shows the held-out R^2 for each model's wage and gender predictions. (Since gender is not real-valued, we use pseudo R^2 based on negative log-likelihood.) CAREER outperforms all the econometric baselines. With the standard supervised fine-tuning, CAREER has a held-out wage R^2 of 0.515 and held-out gender R^2 of 0.510. Its predictive performance is not stemming from including a better functional form for an individual's current job or capturing employment spells more fully. For the two debiased fine-tuning methods, we find that wage R^2 slightly worsens, while gender predictions might slightly improve, although this improvement is within the SE. (We do not have wage or gender predictions for difference-based fine-tuning because it predicts the difference in group wages rather than individual wages or genders.) These results extend a finding from ref. 4, which also shows that transformer-based methods can improve wage predictions relative to econometric baselines. What Table 1 additionally shows is that gender predictions are improved by a larger margin and that these gains are still present for debiased fine-tuning methods. This result demonstrates another benefit of representation learning; if group membership and wage are correlated with similar transformations of input data, then learning representations that are predictive of both can improve predictions.

4.2. Analyzing the Gender Wage Gap. We now compare gender wage gaps estimated with standard econometric techniques to those estimated with a foundation model's representations of labor market history, following the approaches described above. For the econometric models, we use a linear regression using the same covariates described in Section 4.1, encoding occupation into one of 21 coarse-grained labels. For the foundation models, we fine-tune CAREER using each of the three debiased fine-tuning methods described in Section 3. In addition to using the machine learned representations of history, these methods also incorporate the same hand-constructed covariates as the linear model. Because the unexplained wage gap is only identified when there is overlap (e.g., when there are workers with similar histories in both groups), we trim the study population. Specifically, we fine-tune CAREER with supervised fine-tuning to estimate a propensity model $\hat{e}(\lambda(X_i))$, and only include individuals i such that $0.01 < \hat{e}(\lambda(X_i)) < 0.99$. We consider other trimming strategies in *SI Appendix*, section S7, finding similar results. We compute SE with bootstrapping, where the SE reflect sampling uncertainty conditional on the fine-tuned model. We do not retrain models for each bootstrap sample. Instead, we keep models fixed, evaluating each model on the bootstrapped sample; we refer to this as *test-set bootstrapping*.

The results are summarized in Table 2. Across all fine-tuning methods, full history explains more of the gender wage gap than hand-constructed summaries of history, which are typically used to explain gender wage gaps (10, 58, 59). While the wage ratio unexplained by summaries of history is 88.6%, the ratio is above 90% for all methods that use representations of full history, ranging from 90.4% for projection fine-tuning to 93.4% to difference-based fine-tuning. Note that these numbers vary a little depending on the trimming threshold and whether the AIPW estimator (Eq. **4**) is used instead of the outcome-only estimator; additional results are presented in *SI Appendix*, section S7. Table 2 also shows the R-Learner metric in Eq. **8** for each model, which is minimized by difference-based fine-tuning. The last column of Table 2 shows that the difference-based fine-tuning method's R-Learner metric improvement over the regression and multitask fine-tuning improvement is significant at the 95% level. Our analysis for the remainder of the paper considers results from the difference-based fine-tuning approach.

To further investigate where history is explaining the gender wage gap, we consider different cuts of the survey data. *SI Appendix*, Fig. S3 shows the gender wage ratios over time. Compared to the methods that adjust the wage gap for summary statistics of history, the learned representations of history are explaining the least of the gap in 1990 to 1995 and the most of the gap in 2014 to 2019. Overall, we find that compared to methods that adjust the gap for summary statistics of history, full history explains more of the gap for later years. While this may reflect changes in the underlying wage dynamics between these two periods, it may also reflect the fact that more history is available in the later versions of the survey. To further investigate, *SI Appendix*, Fig. S4 fixes the time period to 2014 to 2019 and considers different wage ratios by age. For the youngest workers, history explains the least of the gap; the history-explained ratio is almost the same as the unadjusted ratio for 25- to 34-y-old workers. However, history consistently explains more of the gap as workers become older. Further, *SI Appendix*, Fig. S5 breaks down the adjusted wage gap by the occupational categories considered in ref. 10. Compared to the wage gap explained by summary statistics of history, the wage gap explained by

Table 2. Machine-learned representations of career history explain more of the gender wage gap than hand-constructed summary statistics typically used to estimate the gap

	Adjustment method	Wage ratio	R-Learner metric	R-Learner metric relative to difference-based fine-tuning
Unadjusted	—	0.775 (0.005)	—	—
Adjusted for summary statistics	Linear regression	0.886 (0.001)	0.1841 (0.0019)	−0.00122 (0.00029)
Adjusted for full history	Multitask fine-tuning	0.907 (0.001)	0.1835 (0.0019)	−0.00059 (0.00027)
	Projection fine-tuning	0.904 (0.001)	0.1833 (0.0019)	−0.00036 (0.00025)
	Difference-based fine-tuning	0.934 (0.000)	0.1829 (0.0018)	—

These representations are also validated by better performance on the R-Learner metric (Eq. **8**). All results are estimated with cross-fitting on fivefolds, using the outcome-only estimator. Data are for PSID from 1990 to 2019 with 1% clipping. Test-set bootstrapped SE are in parentheses.

representations of full history is smoother across occupational categories. The occupational category where history explains the most of the wage gap is in nonphysician healthcare-related occupations; the category where it explains the least of the wage gap is for computer-related occupations.

The promise of incorporating more complete representations of history into wage gap analyses is that they can include variables that are typically omitted (recall from Section 1 that omitted variables include variables that are correlated with both wage and gender). This is not unique to machine learning methods; for example, prior studies have found that potential experience (an inexact measure of experience that does not measure workforce interruptions) explains less of the wage gap than years of actual work experience (19, 24), as do representations of history that do not include timing of labor force participation (25). What machine learning offers is the ability to automatically learn these variables without prespecifying them.

Here, we investigate the aspects of history that are captured by the foundation model but omitted by traditional methods that summarize history with summary statistics. While our model learns representations of history for each individual, they are difficult to interpret directly because they are continuous. To better understand these representations of histories, we form clusters of histories. We constrain each history in a cluster to have the same current occupational category but allow the kinds of history in a cluster to vary. We then study which clusters are important for predicting wages by building a regression tree to predict wage from the clusters. The regression tree adds clusters one-at-a-time in the order that is most useful for minimizing wage error. This ordering allows us to identify and interpret which occupation patterns—beyond an individual's current job—are most predictive of wage.

Fig. 2 shows the six most predictive groups (*SI Appendix*, section S6 for more details and *SI Appendix*, Table S2 for the top 15 groups). These groups reveal important aspects of history that are omitted by hand-constructed summaries. For example, one of the most predictive types of histories consists of managers who were previously computer scientists and engineering technicians. This is a predominantly male group (83% male), and managers with these jobs in their histories get paid more than managers without them. There are multiple interpretations for why managers with these histories are paid more than managers without them; perhaps these managers have different skills than other managers, or perhaps they are performing different jobs that are not captured by the occupational encoding scheme (e.g., there is no occupational category for engineering manager). In addition to this group, the other top groups that are important for wage predictions are also correlated with gender. Omitting these variables—like standard econometric methods do—induces omitted variable bias. *SI Appendix*, Table S3 includes additional analyses of omitted variables found at the fine-grained occupational level.

5. Discussion

We used foundation models to study a classic problem from labor economics: estimating the difference between how individuals with the same labor market experience get paid when they belong to different groups. With foundation models, wage predictions improve over econometric baselines by up to 15%. We also showed that an omitted variable bias arises when a foundation model discards relevant information about group differences. To mitigate this problem, we proposed procedures for debiasing foundation models, which we validated on semisynthetic data. On survey data from the PSID, we found that labor market history explains more of the gender wage gap than the summary statistics of history used by standard econometric methods.

These findings are suggestive of how to use foundation models in social science research. One direct application is using the debiasing methods we propose to estimate causal effects with foundation models. While we study a foundation model of labor market history trained on resume data, these methods

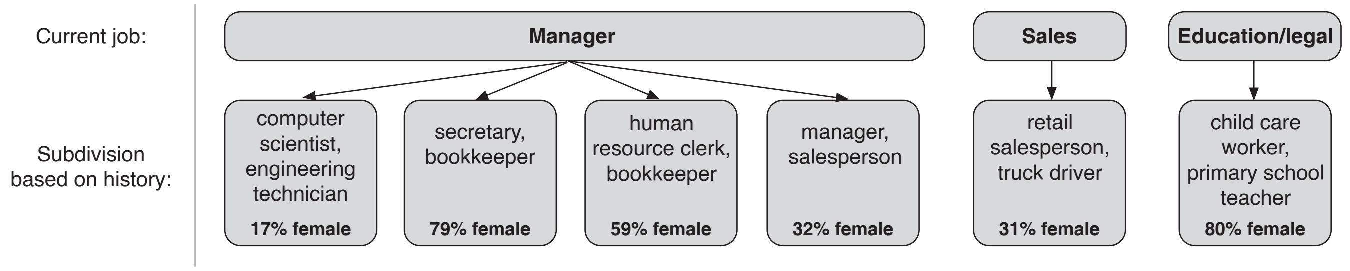
Current job:
Manager
Sales
Education/legal
Subdivision based on history:
computer scientist, engineering technician
17% female
secretary, bookkeeper
79% female
human resource clerk, bookkeeper
59% female
manager, salesperson
32% female
retail salesperson, truck driver
31% female
child care worker, primary school teacher
80% female

Fig. 2. CAREER finds omitted variables from a worker's job history that are important for explaining the gender wage gap. These omitted variables, which are identified by a regression tree as being most predictive of wage, are correlated with both wage and gender.

can extend to analyses involving other foundation models. For example, foundation models trained on rich, nationwide administrative data can help answer a variety of descriptive and causal questions (5).

Additionally, our methods can help address questions about the representativeness of large language models (LLMs), the most common type of foundation model. A recent literature has found that large language models, when queried to answer questions from surveys, do not respond in ways that are representative of the national population (60). The problem of representative predictions and debiased foundation models are closely related. Our methods could be adapted to improve the representativeness of LLMs.

Materials and Methods

The foundation for this work consists of theoretical development and empirical analysis using real-world data. The key components are as follows:

5.1. Data Source. We analyze data from the PSID, a longitudinal survey following American families since 1968. We focus on surveys from 1990 to 2019, comprising 91,391 observations across 19 surveys. Following prior literature, the sample includes nonfarm and nonmilitary wage workers aged 25 to 64 who worked at least 26 wk. Longitudinal sample weights are incorporated to adjust for selection probabilities.

5.2. Model Architecture. We employ CAREER, a foundation model pretrained on 24 million resumes, using a transformer architecture with 64-dimensional representations, 4 encoder layers, 4 attention heads, and 256 hidden units for feedforward neural networks. The model is ensembled over 16 instances.

5.3. Methodology. Our approach involves three key steps:

- Theoretical development of conditions for unbiased estimation using foundation models
- Development of debiased fine-tuning methods to meet these conditions
- Empirical validation through both semisynthetic experiments and real-world analysis

For the debiased fine-tuning, we implement three distinct approaches in addition to the supervised fine-tuning approach:

- Multitask fine-tuning
- Projection fine-tuning
- Difference-based fine-tuning

Detailed mathematical derivations, proof of theoretical results, and comprehensive experimental protocols are provided in *SI Appendix*.

Data, Materials, and Software Availability. The data and code for reproducing the tables and figures in this paper are available in the GitHub repository: https://github.com/gsbDBI/career-wage-gaps-replication (61). The Panel Study of Income Dynamics (PSID) data used for fine-tuning and evaluation are available through the University of Michigan's Institute for Social Research (University of Michigan, 2024). Due to the proprietary nature of the resume data used for pretraining the CAREER model, the data and pretrained models are not publicly accessible. However, a dataset containing all predictions from these models is included in the repository, from which all the figures and tables presented in the analysis can be reproduced.

ACKNOWLEDGMENTS. K.V. is supported by the Harvard Data Science Initiative. S.A. is supported by the Golub Capital Social Impact Lab, the Stanford Institute for Human-Centered Artificial Intelligence, and the Business, Governance, and Society Initiative at Stanford's Graduate School of Business. D.M.B. is supported by NSF IIS-2127869, NSF DMS-2311108, ONR N000142412243, and The Simons Foundation. We thank Revati Kapshikar. We also thank Zippia for generously sharing the resume dataset used in this study.

Author affiliations: [a]Harvard Data Science Initiative, Harvard University, Cambridge, MA 02138; [b]Graduate School of Business, Stanford University, Stanford, CA 94305; [c]Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, Stanford, CA 94305; [d]Department of Computer Science, Columbia University, New York, NY 10027; and [e]Department of Statistics, Columbia University, New York, NY 10027

1. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv [Preprint] (2018). https://arxiv.org/abs/1810.04805 (Accessed 1 December 2024).
2. A. Radford *et al.*, Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
3. R. Bommasani *et al.*, On the opportunities and risks of foundation models. arXiv [Preprint] (2021). https://arxiv.org/abs/2108.07258 (Accessed 1 December 2024).
4. K. Vafa *et al.*, CAREER: A foundation model for labor sequence data. *Trans. Mach. Learn. Res.* **2024** (2023).
5. G. Savcisens *et al.*, Using sequences of life-events to predict human lives. *Nat. Comput. Sci.* **4**, 43–56 (2024).
6. G. W. Imbens, Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86**, 4–29 (2004).
7. E. M. Kitagawa, Components of a difference between two rates. *J. Am. Stat. Assoc.* **50**, 1168–1194 (1955).
8. R. Oaxaca, Male–female wage differentials in urban labor markets. *Int. Econ. Rev.* **14**, 693–709 (1973).
9. A. S. Blinder, Wage discrimination: Reduced form and structural estimates. *J. Hum. Resour.* **8**, 436–455 (1973).
10. F. D. Blau, L. M. Kahn, The gender wage gap: Extent, trends, and explanations. *J. Econ. Lit.* **55**, 789–865 (2017).
11. J. G. Altonji, R. M. Blank, "Chapter 48: Race and gender in the labor market" in *Handbook of Labor Economics*, O. Ashenfelter, D. Card, Eds. (Elsevier, 1999), vol. 3, pp. 3143–3259.
12. Panel Study of Income Dynamics, Public use dataset, produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI (2023). https://psidonline.isr.umich.edu/. Accessed 26 November 2024.
13. N. Fortin, T. Lemieux, S. Firpo, "Decomposition methods in economics" in *Handbook of Labor Economics*, O. Ashenfelter, D. Card, Eds. (Elsevier, 2011), vol. 4, pp. 1–102.
14. G. W. Imbens, D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).
15. V. Chernozhukov *et al.*, Double/debiased machine learning for treatment and structural parameters. *Economet. J.* **21**, C1–C68 (2018).
16. A. S. Blinder, Wage discrimination: Reduced form and structural estimates. *J. Hum. Resour.* **8**, 436–455 (1973).
17. J. Mincer, S. Polachek, Family investments in human capital: Earnings of women. *J. Polit. Econ.* **82**, S76–S108 (1974).
18. A. Manning, J. Swaffield, The gender gap in early-career wage growth. *Econ. J.* **118**, 983–1024 (2008).
19. F. D. Blau, L. M. Kahn, The feasibility and importance of adding measures of actual experience to cross-sectional data collection. *J. Law Econ.* **31**, S17–S58 (2013).
20. F. D. Blau, L. M. Kahn, Analyzing the gender pay gap. *Q. Rev. Econ. Bus.* **39**, 625–646 (1999).
21. R. Böheim, P. Stöllinger, Decomposition of the gender wage gap using the lasso estimator. *Appl. Econ. Lett.* **28**, 817–828 (2021).
22. A. Hegewisch, H. Hartmann, "Occupational segregation and the gender wage gap: A job half done" (Tech. Rep. C419, Institute for Women's Policy Research, Washington, DC, 2014).
23. M. Bonaccolto-Töpfer, S. Briel, The gender pay gap revisited: Does machine learning offer new insights? *Labour Econ.* **78**, 102223 (2022).
24. T. L. Regan, R. L. Oaxaca, Work experience as a source of specification error in earnings models: Implications for gender wage decompositions. *J. Popul. Econ.* **22**, 463–499 (2009).
25. A. Light, M. Ureta, Early-career work experience and gender wage differentials. *J. Law Econ.* **13**, 121–154 (1995).
26. J. Achiam *et al.*, GPT-4 technical report. arXiv [Preprint] (2023). https://arxiv.org/abs/2303.08774 (Accessed 1 December 2024).
27. H. Touvron *et al.*, Llama 2: Open foundation and fine-tuned chat models. arXiv [Preprint] (2023). https://arxiv.org/abs/2307.09288 (Accessed 1 December 2024).
28. A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale" in *International Conference on Learning Representations* (2021).
29. C. Z. A. Huang *et al.*, "Music transformer: Generating music with long-term structure" in *International Conference on Learning Representations* (2019).
30. A. Madani *et al.*, Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
31. V. Chernozhukov, C. Cinelli, W. Newey, A. Sharma, V. Syrgkanis, "Long story short: Omitted variable bias in causal machine learning" (Tech. Rep., National Bureau of Economic Research, 2022).
32. V. Veitch, D. Sridhar, D. Blei, "Adapting text embeddings for causal inference" in *Conference on Uncertainty in Artificial Intelligence* (PMLR, 2020), pp. 919–928.
33. M. Lewis *et al.*, Denoising sequence-to-sequence pre-training for natural language generation. translation, and comprehension. arXiv [Preprint] (2019). https://arxiv.org/abs/1910.13461 (Accessed 1 December 2024).
34. J. M. Robins, Y. Ritov, Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Stat. Med.* **16**, 285–319 (1997).

35. A. A. Tsiatis, *Semiparametric Theory and Missing Data* (Springer, 2006), vol. 4.
36. S. Athey *et al.*, “Efficient inference of average treatment effects in high dimensions via approximate residual balancing” (Tech. Rep., 2016).
37. E. H. Kennedy, Semiparametric doubly robust targeted double machine learning: A review. arXiv [Preprint] (2022). https://arxiv.org/abs/2203.06469 (Accessed 1 December 2024).
38. A. Belloni, V. Chernozhukov, C. Hansen, High-dimensional methods and inference on structural and treatment effects. *J. Econ. Perspect.* **28**, 29–50 (2014).
39. S. M. Shortreed, A. Ertefaie, Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics* **73**, 1111–1122 (2017).
40. D. Tang, D. Kong, W. Pan, L. Wang, Ultra-high dimensional variable selection for doubly robust causal inference. *Biometrics* **79**, 903–914 (2023).
41. E. Cho, S. Yang, Variable selection for doubly robust causal inference. arXiv [Preprint] (2023). https://arxiv.org/abs/2301.11094 (Accessed 1 December 2024).
42. P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
43. L. Battaglia, T. Christensen, S. Hansen, S. Sacher, Inference for regression with variables generated from unstructured data. arXiv [Preprint] (2024). https://arxiv.org/abs/2402.15585 (Accessed 1 December 2024).
44. C. Shi, D. Blei, V. Veitch, “Adapting neural networks for the estimation of treatment effects” in *Neural Information Processing Systems* (2019).
45. V. Chernozhukov, W. Newey, V. M. Quintas-Martinez, V. Syrgkanis, “RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests” in *International Conference on Machine Learning* (2022).
46. P. Cui, S. Athey, Stable learning establishes some common ground between causal inference and machine learning. *Nat. Mach. Intell.* **4**, 110–115 (2022).
47. X. Nie, S. Wager, Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**, 299–319 (2021).
48. P. H. Calamai, J. J. Moré, Projected gradient methods for linearly constrained problems. *Math. Program.* **39**, 93–116 (1987).
49. S. Athey, G. Imbens, Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7353–7360 (2016).
50. S. Wager, S. Athey, Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **113**, 1228–1242 (2018).
51. P. M. Robinson, Root-n-consistent semiparametric regression. *Econometrica* **56**, 931–954 (1988).
52. D. P. Kingma, J. Ba, “Adam: A method for stochastic optimization” in *International Conference on Learning Representations* (2015).
53. F. D. Blau, L. M. Kahn, F. D. Blau, L. M. Kahn, The U.S. gender pay gap in the 1990s: Slowing convergence. *Ind. Labor Relat. Rev.* **60**, 45–66 (2006).
54. D. Autor, D. Dorn, The growth of low-skill service jobs and the polarization of the U.S. labor market. *Am. Econ. Rev.* **103**, 1553–1597 (2013).
55. S. Athey, G. W. Imbens, J. Metzger, E. Munro, Using Wasserstein generative adversarial networks for the design of Monte Carlo simulations. *J. Economet.* **240**, 105076 (2021).
56. J. D. Kang, J. L. Schafer, Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22**, 523–539 (2007).
57. E. A. Stuart, Matching methods for causal inference: A review and a look forward. *Stat. Sci.* **25**, 1 (2010).
58. F. D. Blau, P. Brummund, A. Y. H. Liu, Trends in occupational segregation by gender 1970–2009: Adjusting for the impact of changes in the occupational coding system. *Demography* **50**, 471–494 (2013).
59. C. Goldin, A grand gender convergence: Its last chapter. *Am. Econ. Rev.* **104**, 1091–1119 (2014).
60. S. Santurkar *et al.*, “Whose opinions do language models reflect?” in *International Conference on Machine Learning* (PMLR, 2023), pp. 29971–30004.
61. K. Vafa, R. Kapshikar, S. Athey, D. M. Blei, Data from “Estimating Wage Disparities Using Foundation Models.” GitHub. https://github.com/gsbDBI/career-wage-gaps-replication. Accessed 11 May 2025.