AI as a research tool

SUSAN ATHEY

PROFESSOR, GSB & ICME, STANFORD UNIVERSITY

FACULTY DIRECTOR, GOLUB CAPITAL SOCIAL IMPACT LAB

STANFORD INSTITUTE FOR HUMAN-CENTERED AI

Artificial Intelligence as a Research Tool

EMPIRICAL ANALYSIS

Text/images/video embeddings/clusters as:

- X's: Controls/Predictive features
  - Controlling for confounders/adjustments
  - Predictions as an input to estimation
  - Heterogeneous treatment effects
- W's: "Treatment"
  - Ex: Reviews, style of profiles/resume, topics of news articles
- Y's: Outcomes

Generated output as data w/ structured/varied prompts

Tool for estimation or evaluating empirical methods

- HTE, Policy Learning
- Differentially private learning
- Semi-synthetic simulations for replication or methods comparison
- Federated Learning

EXPERIMENTS/ONLINE

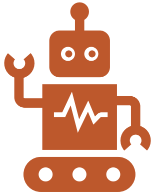
AI to Create Interventions/Stimuli

- Treatment assignment algorithms
- Controlled alt. versions of images/text
  - To expose to experimental subjects
  - To interpret differences in predictions
- Chatbot/robot interaction as interventions

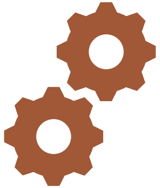
Adaptive Experiments/Reinforcement Learning

- Online algorithms
- Policy learning

Figure 8: Variation in *smile*



Off-the-shelf AI as a Tool for Social Science



Modify/Customize AI Tools for Social Science Use Cases

Improve Performance and Understanding of AI Tools

Science of AI ⟷ Social Science Methods & Applications

Heterogeneity in Treatment Effects and Policy Analysis

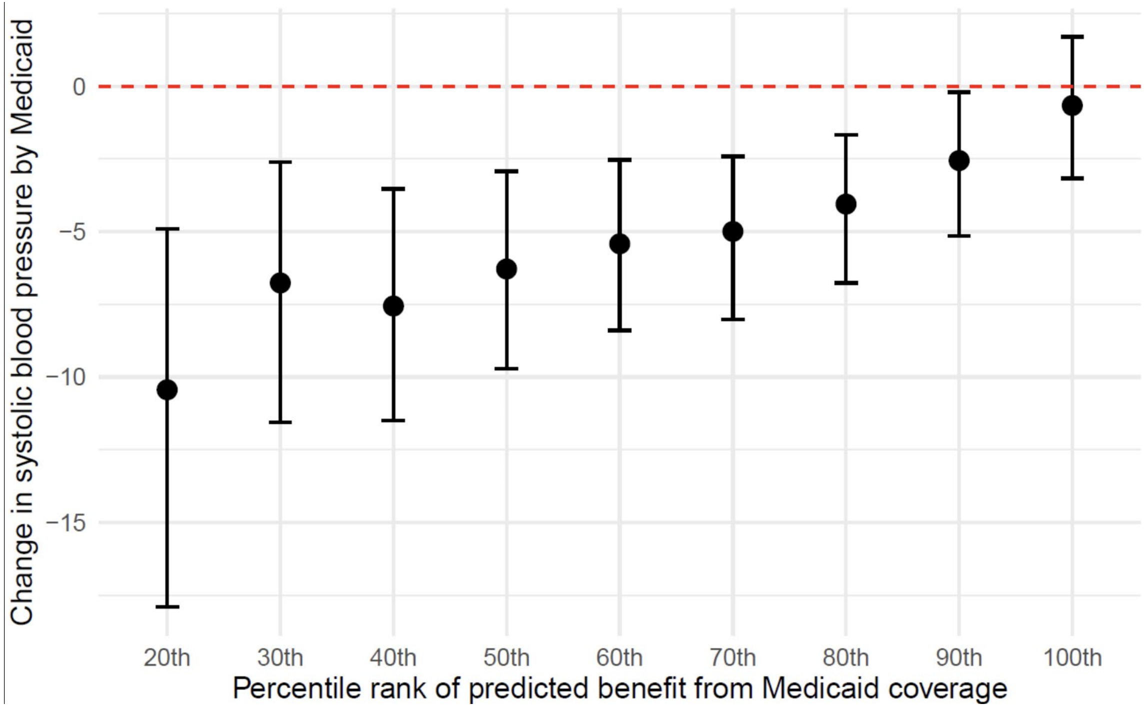
Medicaid coverage does in fact improve health outcomes, for a subgroup of individuals

Low pre-access health spending

Research » Special paper

Heterogeneous effects of Medicaid coverage on cardiovascular risk factors: secondary analysis of randomized controlled trial

BMJ 2024 ; 386 doi: https://doi.org/10.1136/bmj-2024-079377 (Published 23 September 2024)
Cite this as: *BMJ* 2024;386:e079377



Heterogeneity in Treatment Effects and Policy Analysis

Strong Heterogeneity in Response to Displacement

Heterogeneity WITHIN commonly identified groupings, e.g. age x education; firm

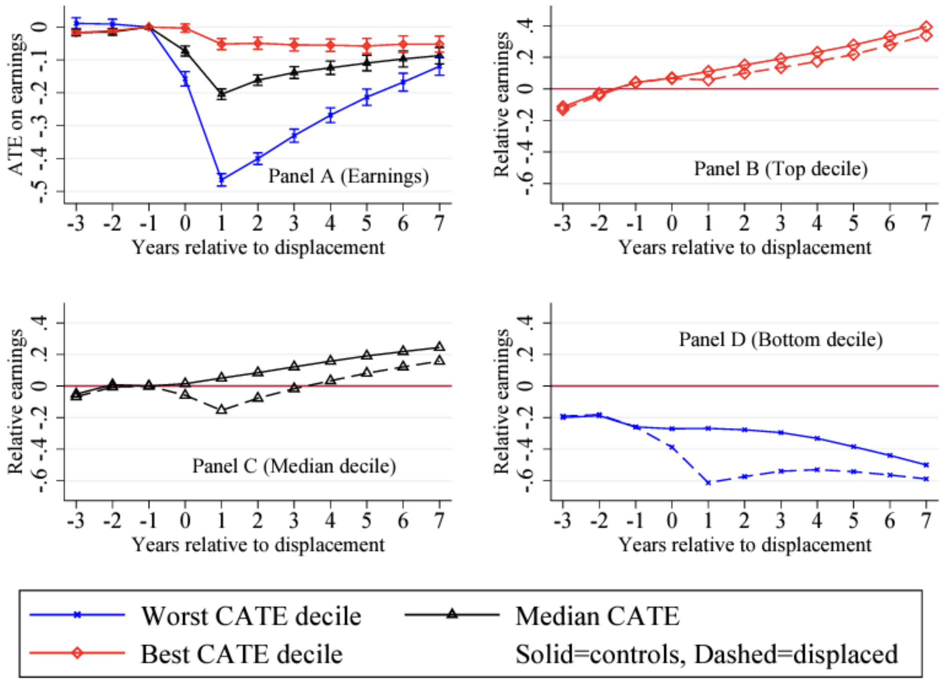
Location very important

Old workers in routine occupations hurt the most by displacement

The Heterogeneous Earnings Impact of Job Loss Across Workers, Establishments, and Markets*

Susan Athey† Lisa K. Simon‡ Oskar N. Skans§ Johan Vikström¶
Yaroslav Yakymovych||

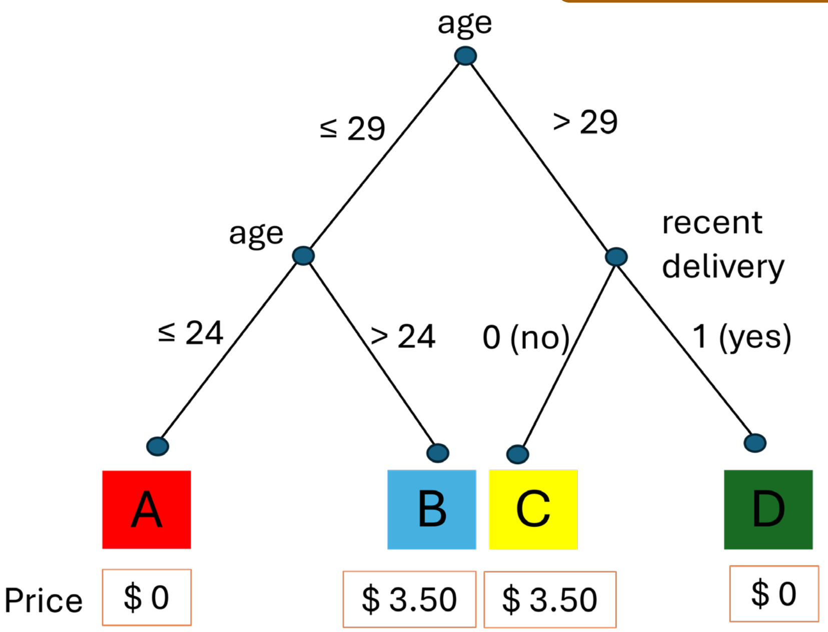
Figure 5: Effects of displacement across time and CATE deciles

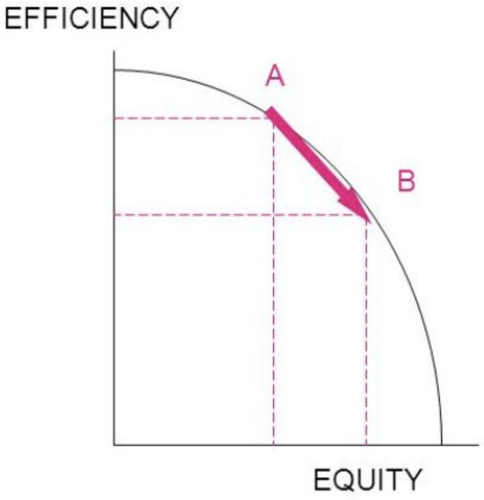


Targeted Digital Interventions

The right content → At the right time and place

With attention to equity-efficiency tradeoffs





Estimating and Evaluating Treatment Assignment Prioritization Rules

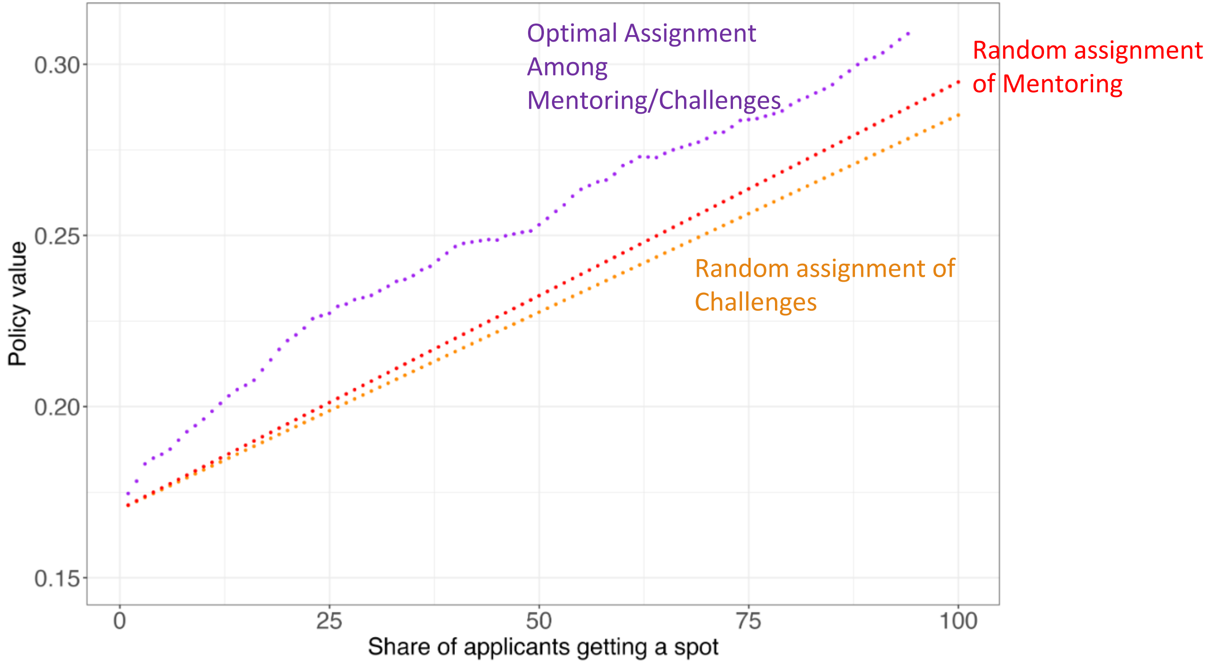
Estimate optimal policy

- For each program a and cov. x, estimate $\hat{\tau}_a(x)$
- Optimization algorithm:
  - Prioritize the program and indiv characteristics that are most effective given capacity

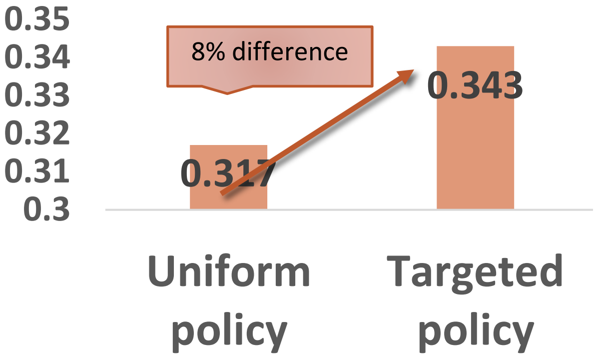
Evaluate with held-out data

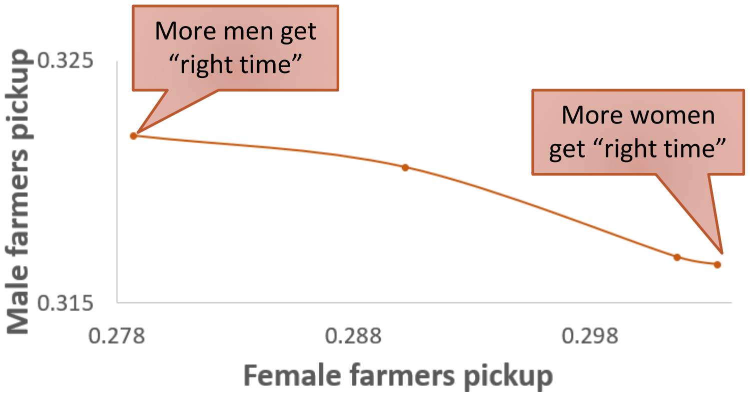
See Athey, Cole, Nath, and Zhu (2023), Sverdrup et al (2023) for more on multi-arm targeting with budget constraints

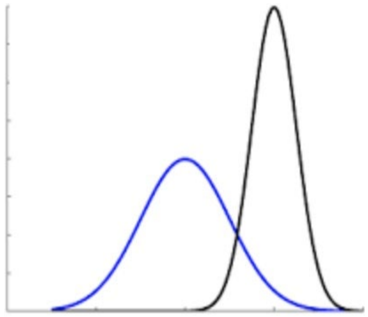
The value of targeting as a function of program capacity



Automated Calling with Agricultural Advice in India: Impact of Personalization in Call Times







Value of Targeting

- Personalizing call time increases prob. of picking up **8%.**
- <u>Impact</u>: potential to reach **26k-33k additional farmers** with ed. content.

Equity-Efficiency Tradeoff

- Capacity constraints: not everyone gets their "right time."
- Women farmers lower average engagement.
- Can improve engagement from women by 9% if we reduce men's engagement by 1.7%.

Shocks/external validity

- Targeted policy underperforms in practice.
- A farmer's "right time" shifts from week to week through season.
- Weight more recent data for better performance.

Digital Education for Students/Learners

Educ. Apps: personalized content, habit formation	• Agrawal, Athey, Kanodia, Palikot (2023a,b) https://arxiv.org/abs/2208.13940 https://arxiv.org/abs/2310.10850
MMS messages teach about misinfo	• Athey, Cersosimo, Koutout, & Li (2022) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4489759
Chatbots teach about misinfo	• Appel, Athey, Karlan, Koutout, Luca, Manjeer, Sacher, & Wernerfelt (WIP 2024)
Text message reminder for financial aid forms	• Athey, Keleher, and Spiess (2023) https://arxiv.org/abs/2310.08672
Automated advice for farmers	• Athey, Cole, Nath, and Zhu (2023) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4536641

Digital Assistants for Teachers/Providers

Teacher dashboards	• Agrawal, Athey, Kanodia (WIP)
Tablet app assists nurses counsel patients	• Athey, Bergstrom, Hadad, Jamison, Ozler, Parisoto, Sama (2023) https://www.science.org/doi/10.1126/sciadv.adg4420

Digital Interventions to Support Donors/Charities

Charity impact matters for Give at Checkout	• Athey, Cersosimo, Karlan, Koutout, & Steimer (2023) https://ssrn.com/abstract=4711399
Tradeoffs in default donation amounts	• Athey, Byambadalai, Cersosimo, Koutout,& Nath (2024) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4785704

Digital Interventions Matching Workers to Employers/Funders

Online work on portfolios to help women transition to IT	• Athey & Palikot (2023) https://arxiv.org/abs/2211.09968
Encourage workers to post credentials for online learning	• Athey & Palikot (2024) https://arxiv.org/abs/2405.00247
Improve attractiveness of online profiles	• Athey, Karlan, Palikot & Yuan (2023) https://arxiv.org/abs/2209.01235

Insights from Causal Methodology/Stats Improve Prediction Methodology

nature machine intelligence

Perspective | Published: 23 February 2022

Stable learning establishes some common ground between causal inference and machine learning

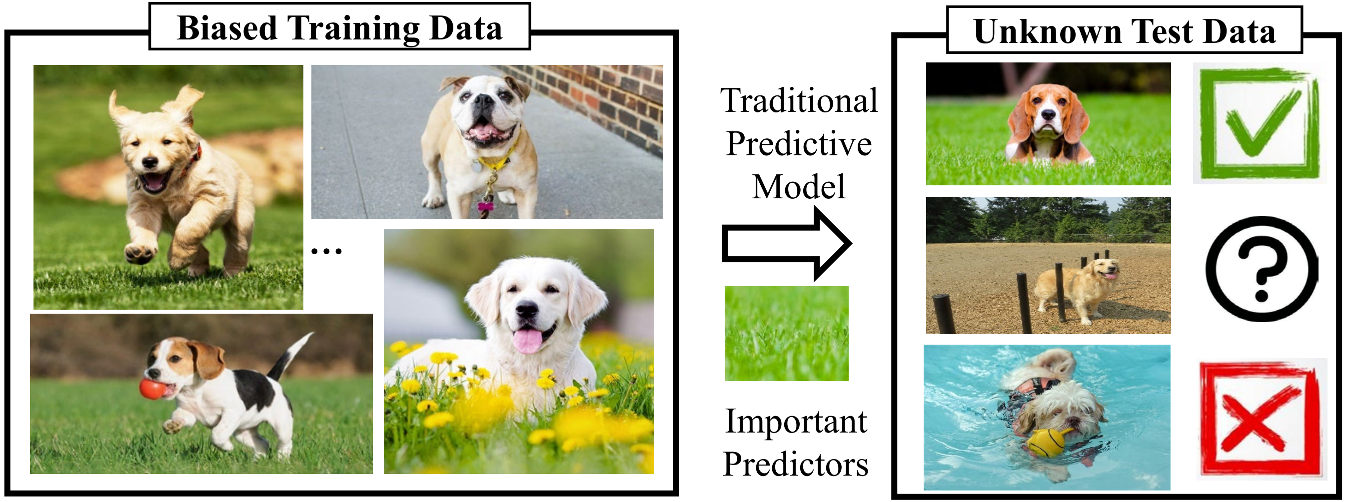
Peng Cui & Susan Athey

Stable Prediction with Model Misspecification and Agnostic Distribution Shift

Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, Bo Li, AAAI 2020

generalized random forests

A package for forest-based statistical estimation and inference. GRF provides non-parametric methods for

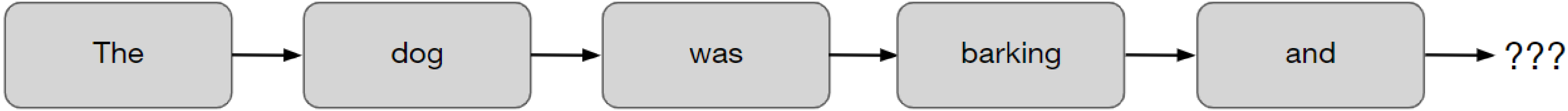


Foundation models for machine learning

ML model trained on large amounts of complicated (high-dimensional) data before being adapted to downstream tasks.

Big ideas:

1. **Self-supervised**: Treat data as a large number of next-object prediction problems without much or any structure (no cleaning/normalizing)

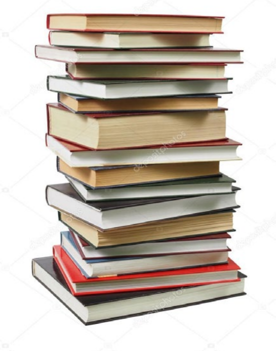


Foundation models for machine learning

ML model trained on large amounts of complicated (high-dimensional) data before being adapted to downstream tasks.

Big ideas:

2. **Embedding functions**: Transform high-dim data to low-dim vectors



`3.4, -1.1, -3.5, 2.4, …, 0.1`

low-dimensional embedding

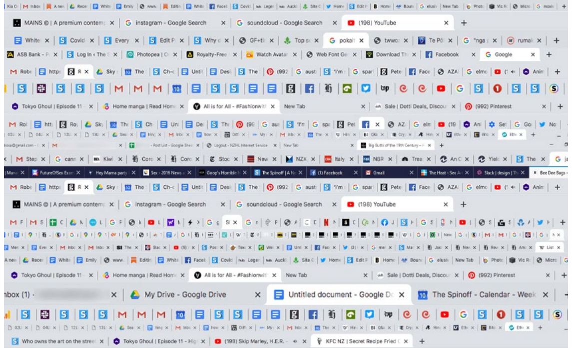
Foundation models for machine learning

ML model trained on large amounts of complicated (high-dimensional) data before being adapted to downstream tasks.

Big ideas:

3. **Fine-tuning**: Training model to fit well on smaller, specialized dataset

Main data source:



Specialized data:

Foundation Models for Economic Problems

Jobs and Careers

- Zhang et al. (2019). "Job2Vec: Job title benchmarking with collective multi-view representation learning."
- Vafa et al. (2024). "CAREER: Transfer Learning for Economic Prediction of Labor Data."

Shopping and Retail Choices

- Ruiz, Athey, and Blei (2020). "SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements."
- Athey et al. (2018). "Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time using Mobile Location Data."
- Donnelly, R., Ruiz, F.J., Blei, D. and Athey, S., 2021. Counterfactual inference for consumer choice across many product categories. Quantitative Marketing and Economics, pp.1-39.

Product reviews

- Boluki, A., Pourmostafa Roshan Sharami, J. and Shterionov, D., 2023, September. Evaluating the effectiveness of pre-trained language models in predicting the helpfulness of online product reviews. In Intelligent Systems Conference (pp. 15-35). Cham: Springer Nature Switzerland.
- Praveen, S.V., Gajjar, P., Ray, R.K. and Dutt, A., 2024. Crafting clarity: Leveraging large language models to decode consumer reviews. Journal of Retailing and Consumer Services, 81, p.103975.

Profile images

- Athey et al. (2022) "Smiles in Profiles: Improving Fairness and Efficiency Using Estimates of User Preferences in Online Marketplaces."
- Ludwig and Mullainathan (2024). "Machine learning as a tool for hypothesis generation."

Government documents and text

- Lee, et al. (2021). "Fednlp: an interpretable nlp system to decode federal reserve communications. "
- Yang, Uy, and Huang (2020). "Finbert: A pretrained language model for financial communications."
- Gentzkow, Shapiro, and Taddy (2019). "Measuring group differences in high-dimensional choices: method and application to congressional speech."
- Liu et al. (2022). "POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection."

Representations of Products: Substitutes and Complements

Ruiz, Athey and Blei, AOAS, 2019:

- Model of boundedly rational consumer choice over shopping baskets with 1000s of products
- Estimates preference parameters from data that includes 1000s of price changes
- Heuristic model of sequential decision-making to simplify computation

query items	complementarity score	
mission tortilla soft taco	2.51	ortega taco shells white corn
	2.40	mcrmck seasoning mix taco
	2.26	lawrys taco seasoning mix
private brand hot dog buns	3.02	bp franks bun size
	2.94	bp franks beef bun length
	2.86	private brand hamburger buns
private brand mustard squeeze bottle	0.53	private brand hamburger buns
	0.44	private brand cutlery full size asst
	0.29	private brand hot dog buns
private brand napkins all occasion	1.01	private brand cutlery full size forks
	0.62	dixie heavy duty plates dspbl 10 1/4 in
	0.39	private brand plate dsgnr 6 7/8 in

Ignoring all textual information and product hierarchy, we infer complementary products from observed choices

$$U_{uit} = \log\left(\sum_{i=1}^{I}\left(y_{uit}\lambda_i + y_{uit}\theta_u^\top\alpha_i + \sum_{j\neq i} y_{uit}y_{ujt}\rho_i^\top\alpha_j\right)\right) + \sum_{i=1}^{I}\left(y_{uit}(-\gamma_u^\top\eta_i \log p_{uit} + \epsilon_{uit})\right)$$

Questions for Social Science Methods & Applications

Can foundation models (FM) **improve empirical methods**?

- Off-the-shelf or custom-created by researcher?
- Do they add value at all?
- Interaction between model size & data size, diversity

What is the **role of fine-tuning (FT) & how should FT be done**?

- Where does it improve performance?
- What is interaction between model size and data?
- Are there tradeoffs in fine-tuning to fit multiple outcomes?

How to **tailor FM/FT methods for economic objectives**?

- E.g. causal questions & representativeness
- Econometric theory that engages with the approach and acknowledges imperfect, general-purpose representations

What **new questions and opportunities** arise?

- Making FM better generally through insights & improvements
- Incorporating richer, messier input data
- Better performance with FM + FT on small dataset
- Richer measured confounders or controls for HTE
- Richer outcomes

Using Transformer Models, LLMs, and Fine Tuning to Analyze Worker Job Transitions and Wages

Vafa, Keyon, Emil Palikot, Tianyu Du, Ayush Kanodia, Susan Athey, and David M. Blei. "CAREER: Transfer Learning for Economic Prediction of Labor Sequence Data." *Transactions of Machine Learning Research*, 2023.

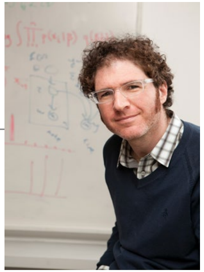
Vafa, Keyon, Susan Athey, and David M. Blei. "Estimating Wage Disparities Using Foundation Models." https://arxiv.org/abs/2409.09894, 2024.

Tianyu Du, Ayush Kanodia, Herman Brunborg, Keyon Vafa, Susan Athey. "LABOR-LLM: Language-Based Occupational Representations with Large Language Models." https://arxiv.org/abs/2406.17972, 2024.









Keyon Vafa
Harvard University

Tianyu Du
Stanford University

Ayush Kanodia
Stanford University

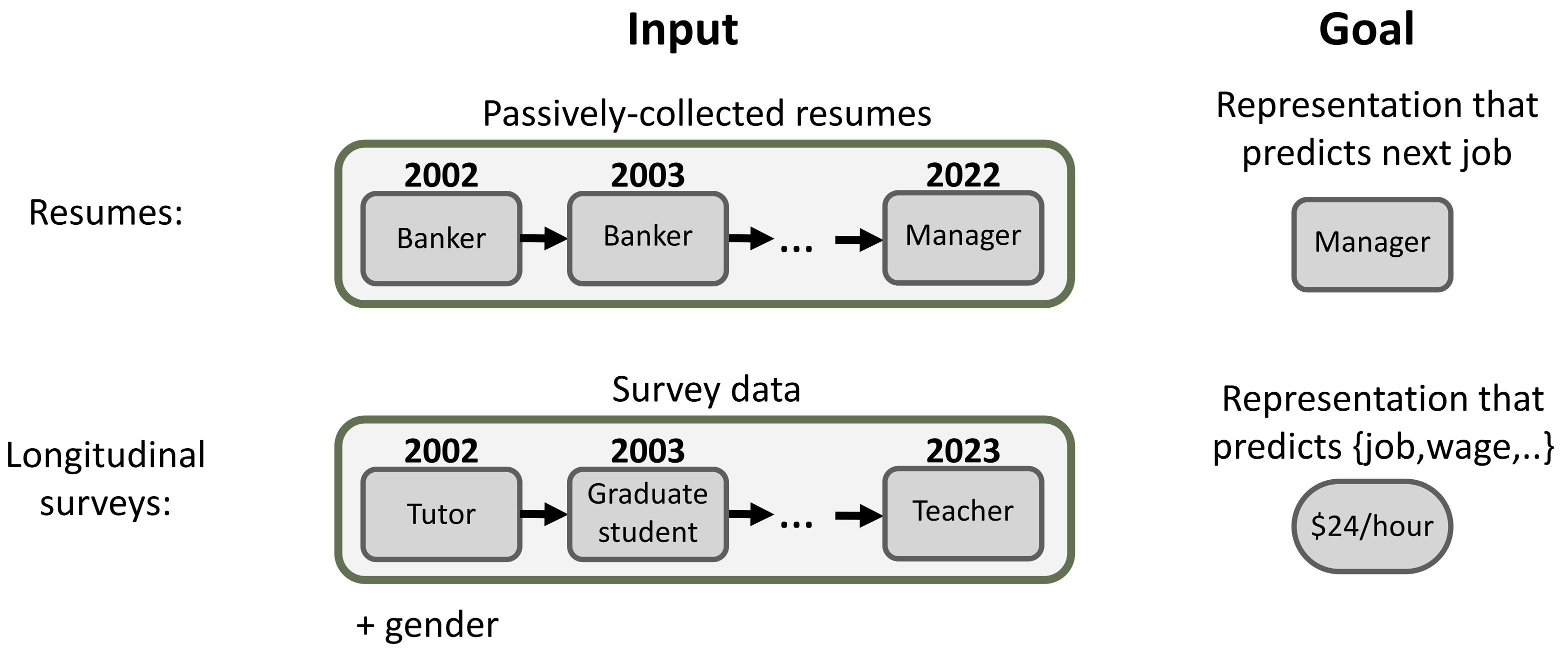
Herman Brunborg
Stanford University

Emil Palikot
Northeastern University

Susan Athey
Stanford University

David Blei
Columbia University

Fine-tuning the foundation model: example



Contextual Attention-based Representations of Employment Encoded from Resumes (**CAREER**), Vafa et al (TMLR, 2023)

Predicting Wages

All models are fit with cross-fitting; reported values are **out-of-sample**.

Model	**Overall R^2**
Coarse-grained regression	0.417 (0.010)
Coarse-grained LASSO	0.430 (0.010)
Fine-grained LASSO	0.456 (0.010)
CAREER (current job only)	0.458 (0.010)
CAREER (participation only)	0.475 (0.009)
CAREER (no pretraining on resumes)	0.521 (0.004)
CAREER (pretraining on resumes)	**0.526 (0.004)**

Improvement is not only due to better functional form of current occupation or capturing workforce participation spells.

Representations and Omitted Variable Bias: The Gender Wage Gap

Vafa, Keyon, Susan Athey, and David M. Blei. "Decomposing Changes in the Gender Wage Gap over Worker Careers." (2023).

Potential for omitted variable bias (OVB)

The full-history-adjusted gender wage gap:

$$\mathbb{E}_H \left\{ \underbrace{\mathbb{E}[Y|G=f,H]}_{A1} - \underbrace{\mathbb{E}[Y|G=m,H]}_{A2} \right\}$$

The representation-adjusted gender wage gap:

$$\mathbb{E}_H \left\{ \underbrace{\mathbb{E}[Y|G=f,\lambda(H)]}_{B1} - \underbrace{\mathbb{E}[Y|G=m,\lambda(H)]}_{B2} \right\}$$

Typical fine-tuning objective: *A1* ≈ *B1* and *A2* ≈ *B2.*

But we only care that *A2* - *A1* ≈ *B2* - *B1*.

Small errors in wage predictions can result in large **omitted variable bias**.

Insight: Modify Fine-Tuning to Optimize OVB

Bias from estimating the average GWG conditional on $\lambda(H)$ vs. H is given by:

$$\mathrm{OVB}(\lambda) = \mathrm{Cov}_{P(h|G=m)}\left(\mathbb{E}[Y|H] - \mathbb{E}[Y|\lambda(H)], \frac{P(G=f|H)}{1-P(G=f|H)} - \frac{P(G=f|\lambda(H))}{1-P(G=f|\lambda(H))}\right)$$

Difference in expected wage as a function of history and representation of history

Difference in gender odds ratio as a function of history and representation of history

Omitted variables related to wage must be unrelated to gender, and vice-versa.

We **introduce methods** for fine-tuning **to optimize OVB**, e.g. R-learner (Xie & Wager)
In our setting, applying OVB-optimized methods **ALSO improves predictions**

Root-*n* consistent estimation with embeddings

Consider a sequence of wage models $\hat{\mu}_n$, propensity models $\hat{e}_n$, and embedding functions λ_n satisfying the following main assumptions:

1. OVB goes to 0 at a root-n rate:

$$\mathrm{OVB}(\lambda_n) = o_P(n^{-1/2})$$

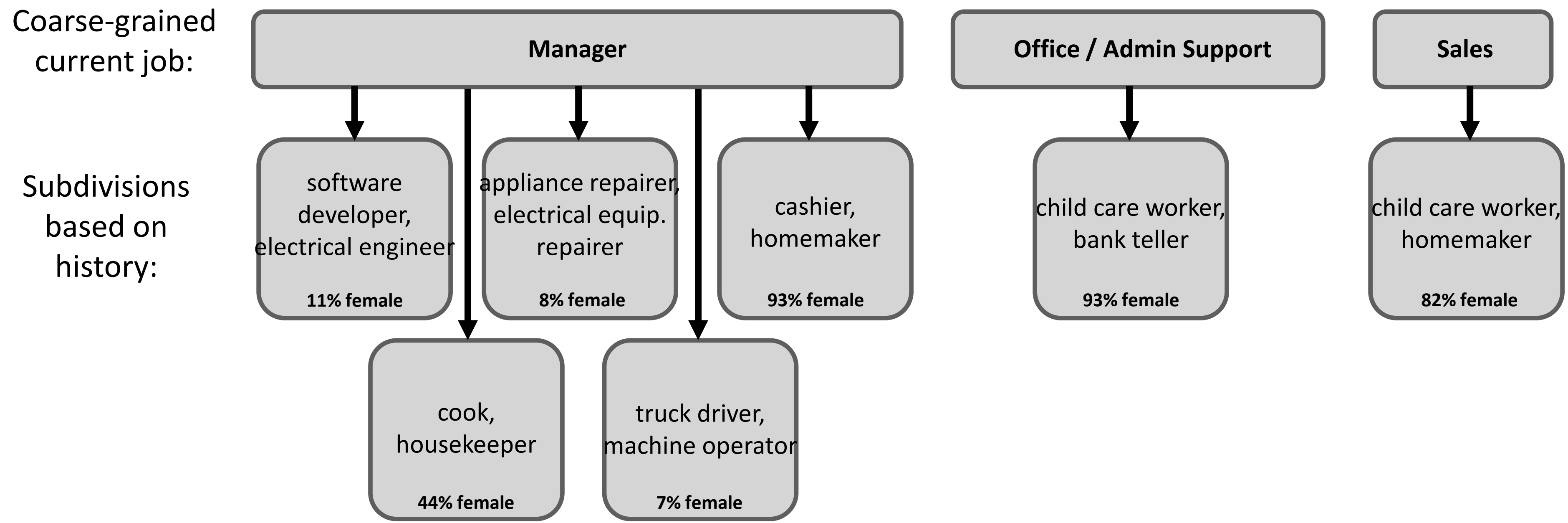
2. Combined root-*n* consistency of wage and propensity models ***as a function of the embedding***:

$$\|\hat{e}_n(\lambda_n(H)) - e(\lambda_n(H))\| \|\hat{\mu}_n(\lambda_n(H)) - \mu(\lambda_n(H))\| = o_P(n^{-1/2})$$

Then the AIPW estimator $\hat{\psi}(\hat{\mu}_n, \hat{e}_n, \lambda_n)$ is root-n consistent for the true history-adjusted gap ψ and asymptotically normal:

$$\sqrt{n}(\hat{\psi}(\hat{\mu}_n, \hat{e}_n, \lambda_n) - \psi) \rightarrow \mathcal{N}(0, \mathrm{Var}(\varphi_{P_{\lambda^*}}(H, G, Y)))$$

Which histories are improving predictions (gender wage gap)?

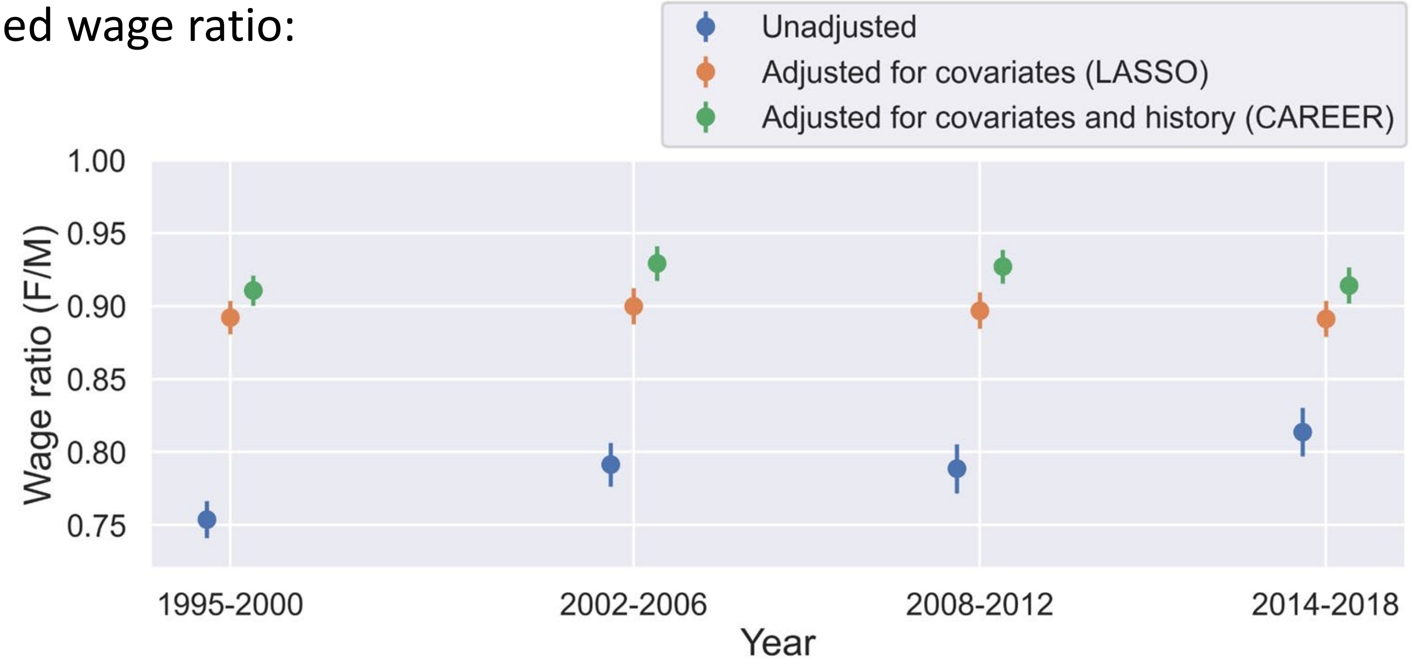


Use prediction tree to define clusters.

Most informative clusters are also predictive of gender (OVB when excluded).

Decomposing wage gap

Unexplained wage ratio:

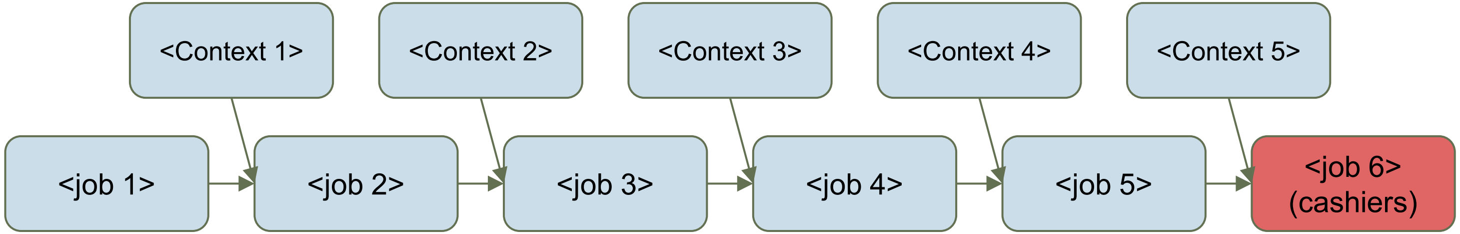


History explains ~16% of remaining wage gap when history is not included.

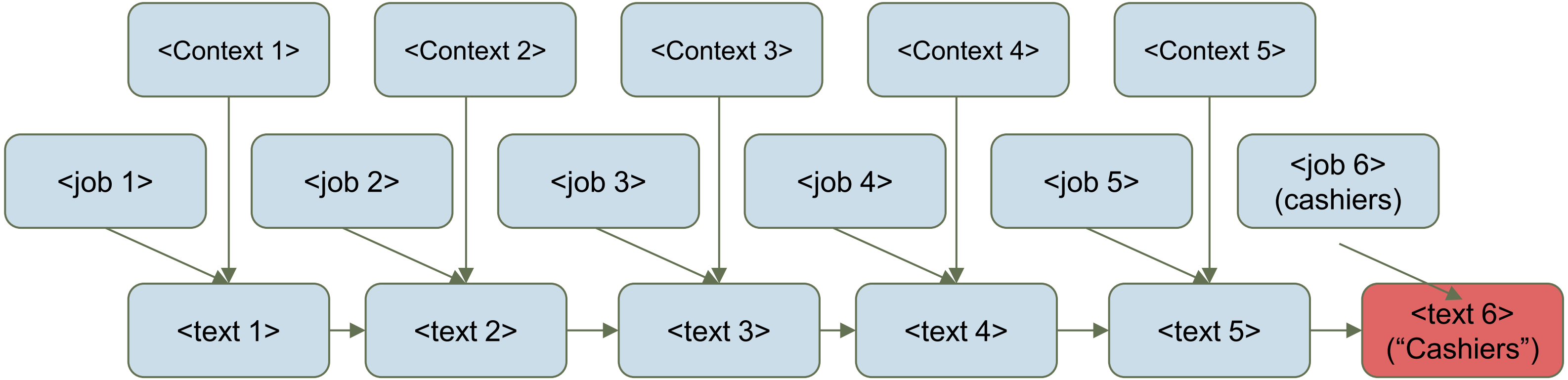
Labor-LLM: General Purpose LLM as Foundation Model for Predicting Occupations

Tianyu Du, Ayush Kanodia, Herman Brunborg, Keyon Vafa, Susan Athey. “LABOR-LLM: Language-Based Occupational Representations with Large Language Models.” https://arxiv.org/abs/2406.17972

Next Word Prediction vs Next Job Prediction



CAREER directly predicts next job, P(<job 6>| <job1><context1>...<job5><context5>)



A fine tuned LLM predicts associated text P(Cashiers | <text1>...<text5>)

Labor LLM: Overview

Labor LLM leverages LLMs as foundation models to improve predictions.

- We show **LLM-based models perform better** than state-of-art occupational choice models.
- Text-based alternative to modeling:
  - Predictions based on English words or embeddings (not discrete occupation set).
  - Enables flexible, diverse data to be used in both training and as input to predictions.
    - Varying structure, gaps in coverage, missing features

- We show **fine-tuning can substitute for larger models** to improve performance.
  - Fine tune on publicly available survey datasets (no proprietary data needed).

Survey dataset	Sample size (workers)	Observations (worker-year)
PSID (79+)	27,700	229,000
NLSY79	12,200	240,000
NLSY97	8,800	114,000

We use 70/10/20 Train-test-validation

Text representations of workers' career histories

Replace structured, categorical description of career with text.

Text template: $T\left(y_{i,<t}, x_i, x_{i,\leq t}\right)$

- Human-readable text file summarizing job history
- Text job titles from the standard occupation classification (SOC)

Prompt summarizing the individual's career history:

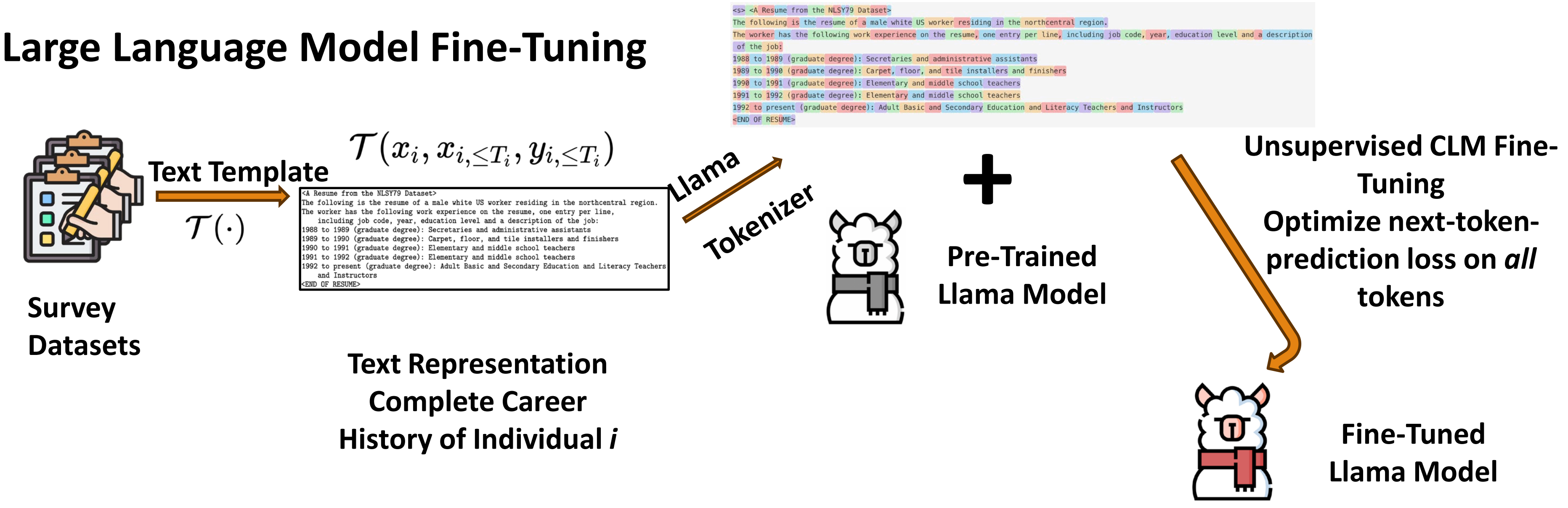
The following is the resume of a **female white** US worker residing in the **northeast region**.
The worker has the following work experience on the resume, one entry per line, including year, education level and the job title:
1979 to 1980 (high school diploma): Cashiers
1980 to 1981 (high school diploma): Not in labor force
1981 to 1982 (high school diploma): Food servers, nonrestaurant
1982 to 1983 (high school diploma): Food servers, nonrestaurant
1983 to 1984 (high school diploma): Food servers, nonrestaurant
1984 to 1985 (high school diploma):

LLM generated response:

<Job history prompt> Waiters and waitresses
1985 to 1986 (high school diploma): Cashiers
1986 to 1987 (high school diploma): Cashiers and office clerks, general
1987 to 1988 (high school diploma): Office clerks, general
1988 to 1989 (high school diploma): Food servers, nonrestaurant

Fine-Tuning LLM (Optional Step)

- Fine-tune Llama-2-7B/13B or Llama-3-8B models on resume text.
- Loss considers **all tokens**, not only occupation titles, learning:
  - Distribution of future jobs conditional on career history.
  - Our template design for representing career histories as text.



Potential Advantages of LLMs as Foundation Models for Job Prediction

Data availability

- Large-scale resume datasets are often proprietary/restricted.
- LLMs are open source or available through API ($$).

Limited scope of data

- LLM’s large training corpus deepen model’s understanding of rare jobs/transitions.

Computation

- Substantial computation required for pre-training using large models & large datasets
- Fine-tuning may be more costly with larger LLMs which contain much broader foundational knowledge and billions+ parameters.
  - Note: various methods to compress size.
- In our setting, fine-tuning open models was cheaper than building our own custom foundation model.

Extensibility

- Incorporating more/different data

Some Findings

Foundation model approach

- Potential to improve performance
  - Incorporate latent structure from larger, broader, but possibly unrepresentative and incomplete data.
- Tradeoffs when comparing methods:
  - Computation
  - Replicability
  - Data availability
  - Representativeness of training data
  - Handling diverse data sources, gaps & missingness

LABOR LLM framework:

- Match state-of-the-art occupational choice models with similar but smaller architecture
- With only open LLM + small public data
- More fine-tuning data substitutes for larger model

Performance improvements

- Derive from text understanding
- Upside: flexibility incorporate more info

High quality embeddings of high-D variables for causal inference and decompositions

- Reduce omitted variable bias (OVB)
- Modify fine-tuning to optimize for OVB