LABOR-LLM: Language-Based Occupational Representations with Large Language Models

Tianyu Du
tianyudu@stanford.edu
Institute for Computational and Mathematical Engineering, Stanford

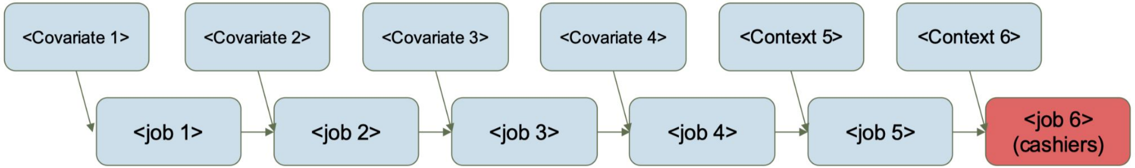
arXiv:2406.17972v2

Collaborators & Acknowledgement

- Susan Athey (Stanford GSB).
- Herman Brunborg (Stanford ICME).
- **Tianyu Du (Stanford ICME).**
- Ayush Kanodia (Stanford Computer Science)
- Keyon Vafa (Harvard Data Science Initiative, Harvard University)

Introduction & Motivation

- **Objective**: predicting workers' next job transitions.



$$\hat{P}\left(\text{next occupation} \,\middle|\, \begin{array}{c}\text{past occupations,}\\ \text{covariates (e.g., demographics, education)}\end{array}\right)$$

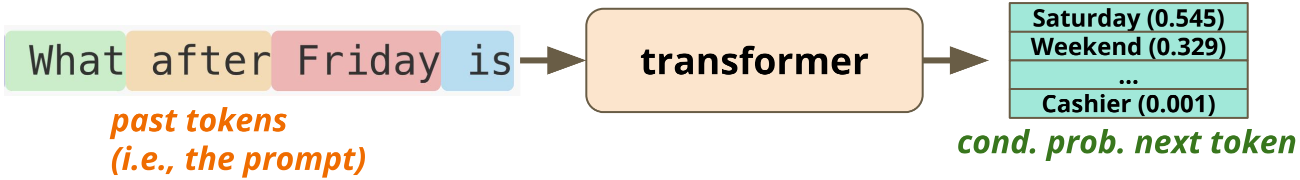
- **Challenge**: Predicting workers' next job transitions in a **<u>high-dimensional and sparse</u>** occupation space
  - OCC: ~ 300 occupations, SOC: ~ 1,000 occupations.
  - Different workers have different numbers of past jobs.
  - High-quality survey data: ~10k individuals.
- **Motivation**: Traditional models oversimplify complex career trajectories
- **Opportunity**: Use pre-trained LLMs to capture nuanced career paths via rich textual data

Traditional Econometric Models & Their Limitations

- **Traditional Approaches**: empirical transition frequencies, multinomial logistic regression, etc.
- Key Characteristics:
  - Limited to fixed-length covariates (imagine running a regression).
  - Reduce complex occupational data to summary statistics or aggregate categories.
  - Use heuristics (e.g., relying on most recent job data) to simplify analysis.
- **Limitations**:
  - Loss of nuance and detailed career history.
  - Reduced predictive power due to oversimplification.
  - Semantic meaning in occupational data.

Machine Learning Approach: CAREER Transformer

- What's a transformer?
  - Simplified view: ***P(next token | past tokens)***
  - with a predefined set of tokens (i.e., the **vocabularies**), e.g., English words/sub-words.
  - Chain rule: ***P(v1, v2 |past tokens) = P(v1 |past tokens) P(v2 |past tokens, v1)***



- **CAREER model**: transformer-based, resume pre-training approach
  - Use occupations as tokens (335 occupations, including special codes une -education/out-of-labor-force).
  - *P(n*
  - 5.6 ds massive data for training.
  - Pre-trained on job sequences from 24 million resumes, fine-tuned and evaluated on survey data.

CAREER:
Big Model + Large Noisy Data + Small High-Quality Data
⇒ Good Performance

Reference: 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. https://arxiv.org/abs/2202.08370

Research Question & Objectives

- **Problem:** Only limited amount of high-quality data.
- **Traditional Models**: oversimplified and unsatisfactory performance.
- **CAREER**: great performance but requires massive data to train.
- **Question**: Can we use LLMs to achieve great performance with limited amount of data.
    - The pre-training + fine-tuning paradigm!

Next-Word Prediction and Next-Job Prediction

What we need:

“**career history**”: past occupations ***y*** and covariates ***X*** (e.g., gender, ethnicity, YoB)

transformer

Cashiers (0.545)
Engineers (0.329)
...
OOLF (0.001)

<A worker from the PSID dataset>
The following information is available about the work history of a **female Black or African American** US worker residing in the **south** region.
The worker was born in **1963**.
The worker has the following records of work experience, one entry per line, including year, education level, and the job title:
1986 (college): Sales Representatives Services All Other
1987 (college): Wholesale and retail buyers, except farm products
1988 (graduate degree): Elementary and middle school teachers
1989 (graduate degree): Elementary and middle school teachers
1990 (graduate degree):

Text template $\mathcal{T}(\cdot)$

Job Titles

What LLMs offer:

What after Friday is

past tokens

transformer

Saturday (0.545)
Weekend (0.329)
...
Cashier (0.001)

cond. prob. next token

Example of Text Template

Prompt summarizing the individual's career history:

<A worker from the PSID dataset>
The following information is available about the work history of a **female Black or African American** US worker residing in the **south** region.
The worker was born in **1963**.
The worker has the following records of work experience, one entry per line, including year, education level, and the job title:
1986 (college): Sales Representatives Services All Other
1987 (college): Wholesale and retail buyers, except farm products
1988 (graduate degree): Elementary and middle school teachers
1989 (graduate degree): Elementary and middle school teachers
1990 (graduate degree):

Text template $\mathcal{T}(\cdot)$

LLM generated response:

<Job history prompt> Postsecondary teachers
1991 (graduate degree): Computer and information systems managers
1992 (graduate degree): Postsecondary teachers
1993 (graduate degree): First-line supervisors

LABOR-LLM without Fine-Tuning: Prediction/Inference

Text template $\mathcal{T}(\cdot)$

$$\mathcal{T}(x_i, x_{i,<t}, y_{i,<t})$$

```
<A worker from the PSID dataset>
The following information is available about the work history of a female black or americ
     US worker residing in the south region.
The worker was born in 1963.
The worker has the following records of work experience, one entry per line, including year,
     education level, and the job title:
1984 (some college): Cooks
1985 (some college): Cooks
1987 (some college): Food servers, nonrestaurant
1989 (some college):
```

<u>Prompt 1:</u> individual career history

<u>Prompt 2:</u> include list of job titles in context

$$\begin{aligned}
&\hat{P}(y_{i,t} \mid x_i, x_{i,\leq t}, y_{i,<t}) \\
&= P_{\text{LLM}}(\texttt{software}\ \texttt{engineer} \mid \mathcal{T}) \\
&= P_{\text{LLM}}(\texttt{software} \mid \mathcal{T}) \times P_{\text{LLM}}(\texttt{engineer} \mid \mathcal{T}, \texttt{software})
\end{aligned}$$

LABOR-LLM without Fine-Tuning: Evaluation

- **Evaluation datasets**: three US-based representative survey datasets.

Survey dataset	Sample size (workers)	Observations (worker-year)
PSID (81+)	31,056	313,622
NLSY79	12,479	259,778
NLSY97	8,984	148,795

We use 70/10/20 worker-level train-test-validation

- **Prediction target**: 335 job titles, including in-education, unemployed, OOLF.
- **Evaluation metric**: the **perplexity** on test set.

$$\text{Perplexity} = \exp\left\{-\frac{1}{\sum_i T_i}\sum_i\sum_{t=1}^{T_i} w_{it}\left[\log \hat{P}(\underbrace{y_{i,t}}_{\text{Next occupation}} \mid \overbrace{x_i, x_{i,\le t}, y_{i,<t}}^{\text{Covariates (e.g., education) and past occupations}})\right]\right\}$$

  - A <u>negative</u> transformation of test-set log-likelihood ⇒ **the lower the better**.
  - **Interpretation**: an uniform distribution across *K* classes has a perplexity of *K*.

LABOR-LLM without Fine-Tuning: Results

Perplexity of Prompt + Model on each Dataset

Prompt format	Model	PSID (81+)	NLSY79	NLSY97
	CAREER	8.616 (0.132)	8.609 (0.157)	6.397 (0.099)
Without list of job titles	Llama-2-7B (32k)	241.044 (22.812)	182.748 (16.373)	173.942 (22.880)
	Llama-3.1-8B	127.788 (10.564)	110.871 (8.973)	99.156 (11.408)
	Llama-3.2-1B	456.088 (51.012)	371.333 (38.769)	277.735 (40.961)
	Llama-3.2-3B	165.109 (14.493)	134.391 (11.186)	122.578 (14.671)
With list of job titles	Llama-2-7B (32k)	42.005 (2.522)	45.718 (2.678)	47.952 (4.127)
	Llama-3.1-8B	30.850 (1.633)	26.984 (1.309)	21.910 (1.394)
	Llama-3.2-1B	62.234 (3.885)	53.313 (3.068)	45.252 (3.518)
	Llama-3.2-3B	39.811 (2.199)	39.236 (2.227)	35.443 (2.700)

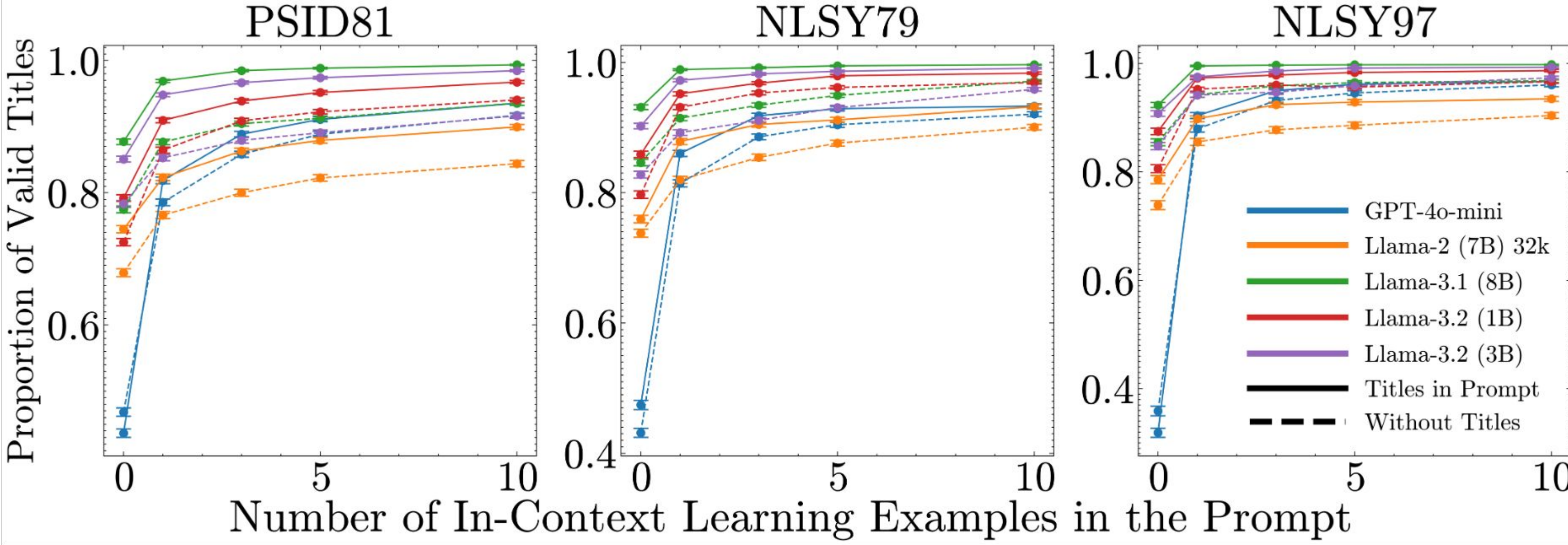
**Bootstrapped standard errors in parentheses.*

- Off-the-shelf LLMs can generate whatever it wants: ***P(any valid title | prompt)*** could be small
- Prompting helps some with valid job titles, improves ***P(any valid title | prompt).***
- Poor performance: data needed to learn conditional probabilities.

LLMs are not magic!

In-Context Learning for Off-the-shelf LLMs

- Include “example data” (from other workers) to the prompt.



In-Context Learning of Off-the-Shelf LLMs

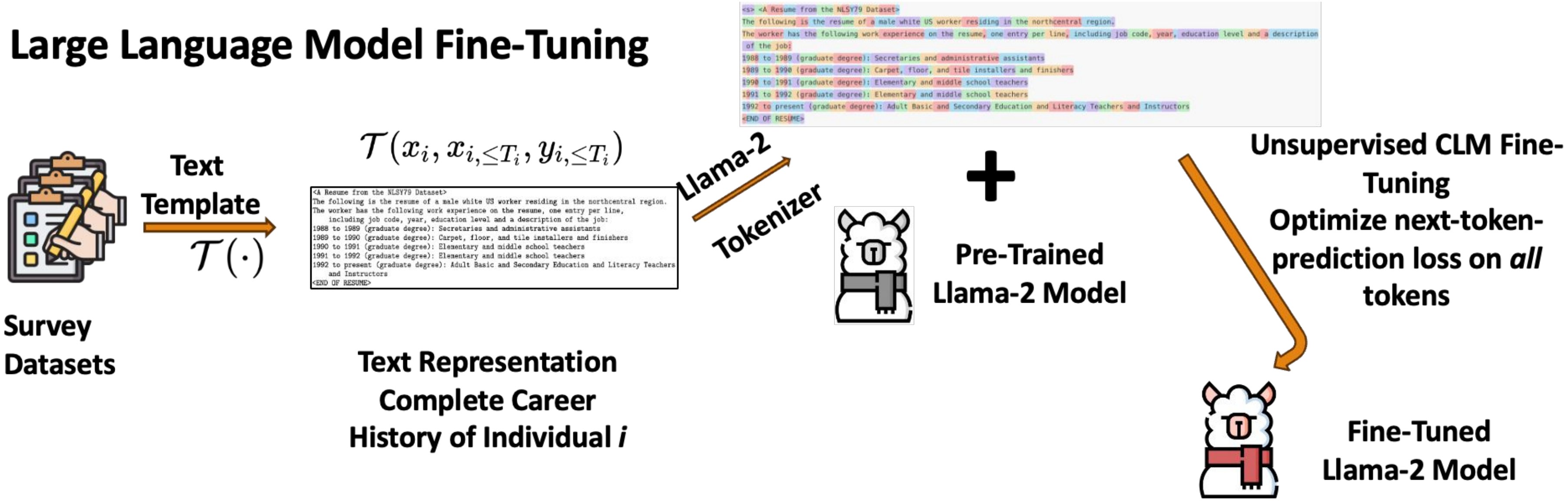
Dataset		PSID81	NLSY79	NLSY97
Number of Transitions ($\sum_{i \in \textbf{test}} T_i$)		6,177	5,159	2,995
Models Without Job Titles in Prompt	**# Resumes**			
OTS Llama-2-7b-32k	0	241.04 (22.812)	182.75 (16.373)	173.94 (22.880)
OTS Llama-2-7b-32k	10	36.53 (2.131)	26.20 (1.495)	17.52 (1.510)
OTS Llama-3.1-8B	0	127.79 (10.564)	110.87 (8.973)	99.16 (11.408)
OTS Llama-3.1-8B	10	25.08 (1.385)	19.41 (1.009)	13.68 (1.034)
OTS Llama-3.2-1B	0	456.09 (51.012)	371.33 (38.769)	277.73 (40.961)
OTS Llama-3.2-1B	10	52.90 (3.740)	36.04 (2.409)	24.99 (2.631)
OTS Llama-3.2-3B	0	165.11 (14.493)	134.39 (11.186)	122.58 (14.671)
OTS Llama-3.2-3B	10	29.92 (1.726)	22.95 (1.306)	16.21 (1.334)
Models With Job Titles in Prompt	**# Resumes**			
OTS Llama-2-7b-32k	0	42.01 (2.522)	45.72 (2.678)	47.95 (4.127)
OTS Llama-2-7b-32k	10	20.73 (0.918)	18.04 (0.732)	11.74 (0.736)
OTS Llama-3.1-8B	0	30.85 (1.633)	26.98 (1.309)	21.91 (1.394)
OTS Llama-3.1-8B	10	16.45 (0.763)	15.20 (0.631)	10.49 (0.672)
OTS Llama-3.2-1B	0	62.23 (3.885)	53.31 (3.068)	45.25 (3.518)
OTS Llama-3.2-1B	10	22.95 (1.130)	20.25 (0.913)	14.02 (0.990)
OTS Llama-3.2-3B	0	39.81 (2.199)	39.24 (2.227)	35.44 (2.700)
OTS Llama-3.2-3B	10	17.81 (0.824)	16.39 (0.683)	11.52 (0.749)

Note: Perplexity on a 10% random sample of the test set, with test-set-bootstrap standard errors in parentheses.

- ICL reduces hallucination and improves performance.
- But still unsatisfactory comparing to CAREER.

LABOR-LLM Fine-Tuning

- Fine-tune Llama-2-7B/13B models on 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.



FT-LABOR-LLM: Prediction/Inference

Survey datasets

Text template $\mathcal{T}(\cdot)$

$$\mathcal{T}(x_i, x_{i,<t}, y_{i,<t})$$

```
<A Resume from the NLSY79 Dataset>
The following is the resume of a male white US worker residing in the northcentral region.
The worker has the following work experience on the resume, one entry per line,
    including job code, year, education level and a description of the job:
1988 to 1989 (graduate degree): Secretaries and administrative assistants
1989 to 1990 (graduate degree): Carpet, floor, and tile installers and finishers
1990 to 1991 (graduate degree):
```

<u>Prompt</u> Individual career history

(List of job titles no longer needed)

Updated

Predict the next occupation using job titles with the fine-tuned LLM and text representation as the prompt

$$\hat{P}(y_{i,t} \mid x_i, x_{i,\le t}, y_{i,<t})$$
$$= P_{\text{LLM}}(\texttt{software engineer} \mid \mathcal{T})$$
$$= P_{\text{LLM}}(\texttt{software} \mid \mathcal{T}) \times P_{\text{LLM}}(\texttt{engineer} \mid \mathcal{T}, \texttt{software})$$

FT-LABOR-LLM: Results

TABLE 6. Test-set perplexity and perplexity improvement for fine-tuned vs. baseline models.

Dataset	PSID81	NLSY79	NLSY97
Number of Transitions $(\sum_{i \in \textbf{test}} T_i)$	61,772	51,593	29,951
Perplexity			
Empirical Transition Frequency	14.65 (0.224)	14.26 (0.271)	10.05 (0.169)
CAREER (Vafa et al. (2024))	8.60 (0.132)	8.64 (0.158)	6.41 (0.101)
FT-7B-NBY	8.36 (0.129)	8.39 (0.148)	6.40 (0.102)
FT-13B-NBY	8.31 (0.127)	8.35 (0.146)	6.34 (0.100)
Perplexity Improvement			
PPL(CAREER)-PPL(FT-7B-NBY)	0.24 (0.020)	0.25 (0.023)	0.02 (0.018)
PPL(CAREER)-PPL(FT-13B-NBY)	0.29 (0.021)	0.28 (0.023)	0.07 (0.016)
PPL(FT-7B-NBY)-PPL(FT-13B-NBY)	0.05 (0.012)	0.04 (0.013)	0.05 (0.011)

Note: Test-set-bootstrap standard errors are in parentheses.

Perplexity improvement over CAREER

- Fine-tuned LLMs outperformed CAREER across datasets.
- Build state-of-the-art occupation models without access to proprietary resume dataset.

Source of LLM's Advantage: Use Numbers as "Job Titles"

- How does LLM performance depend on words in the job title?
- LLM gained an understanding of jobs from pre-training.
- Literal job titles as the bridge.
- Transform literal English job into numbered job titles (in random order).
- Make LLM forget about what it learned about jobs from pre-training.

<Preamble Omitted>
The worker has the following records of work experience, one entry per line, including year, education level, and the job title:
1984 (college): **Business Teachers Postsecondary**
1985 (college): **Postmasters and mail superintendents**
1986 (graduate degree): **Sales Representatives Services All Other**
1987 (graduate degree): **Sales Representatives Services All Other**

Transform "literal job titles" into generic "numbered job titles".

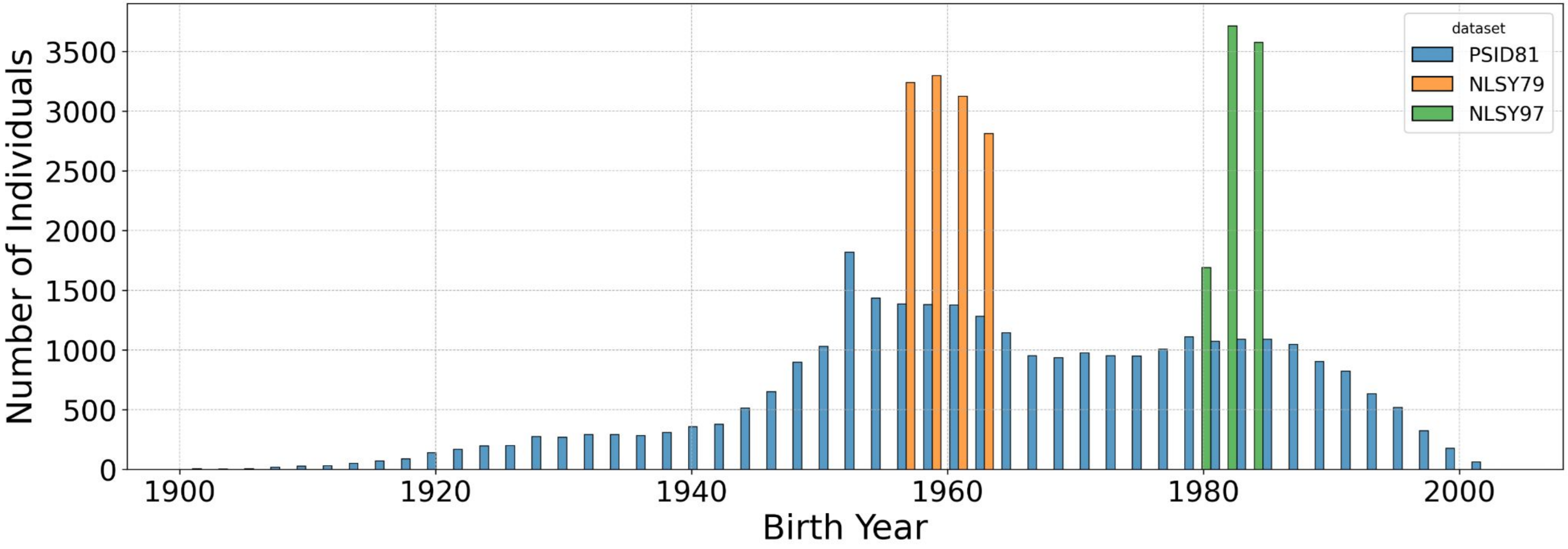
<Preamble Omitted>
The worker has the following records of work experience, one entry per line, including year, education level, and the job title:
1984 (college): **job_032**
1985 (college): **job_124**
1986 (graduate degree): **job_315**
1987 (graduate degree): **job_315**

Perplexity Improvement on Test Set	PSID (81+)	NLSY79	NLSY97
FT Llama-2-7B (Numbered Titles) – FT Llama-2-7B (Literal English Titles)	0.647 (0.027)	0.800 (0.031)	0.370 (0.021)

Information in literal job titles improves model performance significantly!

Further Boost the Performance: Value of Data

- So far, we treated three survey datasets independently.
- Learning from NLSY data should help with PSID prediction.



Fine-tuning LLMs on Mixture of Data: Setting 1

- Compare Llama-2-7B model fine-tuned on the training set of a ***single alternative*** survey dataset. (***specialist models***)

- Then compare to training on subsamples of "**mixture data**" sampled from union of NLSY79, NLSY97, PSID (81+).

- Fine-tune Llama-2-7B models using sub-samples of the mixture of training datasets. (***generalist models)***

- **With full combined dataset, Llama-2-7B FT matches or beats 13B.**

Fine-tuning Data	**PSID (81+)**	**NLSY79**	**NLSY97**
PSID (81+), N=21,739	**8.184 (0.126)**	10.705 (0.198)	10.523 (0.154)
NLSY79, N=8,735	9.929 (0.154)	**8.329 (0.147)**	7.965 (0.123)
NLSY97, N=6,288	12.638 (0.213)	11.274 (0.228)	**6.350 (0.101)**
Baseline Llama-2-13B	8.140 (0.126)	8.282 (0.145)	6.326 (0.100)
Difference vs. Baseline			
20% of mix, N=7,352	-0.635 (0.024)	-0.549 (0.026)	-0.200 (0.016)
40% of mix, N=14,704	-0.246 (0.017)	-0.198 (0.016)	-0.016 (0.013)
60% of mix, N=22,057	-0.116 (0.014)	-0.053 (0.015)	**0.070 (0.012)**
80% of mix, N=29,409	-0.011 (0.013)	**0.024 (0.014)**	0.112 (0.012)
100% of mix, N=36,762	**0.057 (0.014)**	0.075 (0.015)	0.132 (0.013)

Fine-tuning Mixture of Data: Setting 2 (*if time permits*)

- Does adding more FT data to base data help, even if from distinct sources?
- Add increments of ***P%*** of workers in base dataset randomly sampled from workers in ***the other two survey datasets***.
- Additional data has different distribution of calendar years and worker experiences vs. the base data.

	PSID (81+)	NLSY79	NLSY97
Base Train # Workers	**21,739**	**8,735**	**6,288**
Base Train # Tokens	**5.5 million**	**3.8 million**	**2.2 million**
+% of Base Train Workers, Model			
+0%, 7B	**8.184 (0.126)**	**8.329 (0.147)**	**6.350 (0.101)**
+0%, 13B	**8.140 (0.126)**	**8.282 (0.145)**	**6.326 (0.100)**
Difference vs. 13B Baseline			
+10%, 7B	-0.039 (0.014)	-0.033 (0.013)	-0.007 (0.010)
+30%, 7B	**0.030 (0.014)**	-0.011 (0.012)	**0.033 (0.010)**
+50%, 7B	0.045 (0.013)	**0.001 (0.013)**	0.047 (0.010)
+70%, 7B	0.053 (0.014)	0.017 (0.013)	0.065 (0.010)

**Bootstrap standard errors in parentheses.*

Analysis: Impact of Demographic Features (*if time permits*)

Evaluation Dataset **Number of Transitions** ($\sum_{i \in \textbf{test}} T_i$)	PSID81 61,772	NLSY79 51,593	NLSY97 29,951
Perplexity			
No modification / Actual	8.18 (0.126)	8.33 (0.147)	6.35 (0.101)
Randomized ethnicity	8.45 (0.130)	8.40 (0.148)	6.39 (0.100)
Randomized gender	9.22 (0.151)	9.18 (0.167)	6.90 (0.117)
Randomized region	8.20 (0.126)	8.38 (0.148)	6.36 (0.101)
Randomized gender and ethnicity	9.37 (0.152)	9.23 (0.167)	6.94 (0.117)
Randomized gender and region	9.29 (0.152)	9.28 (0.169)	6.90 (0.117)
Randomized ethnicity and region	8.43 (0.129)	8.44 (0.149)	6.39 (0.100)
Randomized all variables	9.44 (0.153)	9.33 (0.170)	6.93 (0.117)
Perplexity Improvement			
PPL(Randomized ethnicity)-PPL(Actual)	0.27 (0.012)	0.07 (0.007)	0.04 (0.006)
PPL(Randomized gender)-PPL(Actual)	1.04 (0.033)	0.85 (0.032)	0.54 (0.027)
PPL(Randomized region)-PPL(Actual)	0.02 (0.004)	0.05 (0.005)	0.01 (0.002)
PPL(Randomized gender and ethnicity)-PPL(Actual)	1.19 (0.034)	0.90 (0.034)	0.58 (0.027)
PPL(Randomized gender and region)-PPL(Actual)	1.11 (0.034)	0.95 (0.036)	0.55 (0.027)
PPL(Randomized ethnicity and region)-PPL(Actual)	0.25 (0.011)	0.11 (0.008)	0.04 (0.006)
PPL(Randomize all)-PPL(Actual)	1.25 (0.036)	1.00 (0.037)	0.58 (0.027)

Note: The foundation model is FT-7B fine-tuned on the union of the training sets of the surveys without any modification of demographic features. Test-set-bootstrap standard errors are in parentheses.

Limitations & Future Directions

- Challenges:
  - High computational cost for fine-tuning large models
  - Sensitivity to training set variations and demographic inputs
- Future Work:
  - Extend to larger, diverse datasets and additional outcome domains
  - Develop cost-efficient fine-tuning and robust uncertainty quantification
  - Explore broader applications (e.g., consumer behavior prediction)
    - P(***other things*** | career history)

Conclusion

- Performance
    - Better predictive performance using fine-tuned LLMs.
- Availability
    - Large-scale resume datasets are often proprietary/restricted.
    - LLMs are open source or available through API.
- Scope of data
    - LLM’s large training corpus deepens model’s understanding of rare jobs and transitions.
- Computation
    - Substantial computation required for pre-training using large models & large datasets
    - Fine-tuning may be more costly with larger LLMs (e.g., 70B and 405B) which contain much broader foundational knowledge and billions+ parameters.
    - Note: various methods to compress size (e.g., LoRA).
    - In our setting, fine-tuning open models was cheaper than building our own custom foundation model (in GPU-hours and $$).

Appendix: Comparison of Methods

Approach	Primary Prediction Outcome	Feature Generation	# of Estimated Params	Estimation/ Fine-tuning Data	Found. Model	Found. Model Prediction Outcome	Found. Training Data	Computation Cost
Traditional econometric models	Wage or Occupation	Manual	10s, 100s, or 1000s	N/A	N/A	N/A	N/A	CPU minutes to hours
Deep learning methods	Wage or Occupation	Automated	Millions	Survey or resume	N/A	N/A	N/A	GPU hours
CAREER	Wage or Occupation	Automated Latent embed dimension: 768	Millions	Survey datasets, target = occupations	Custom Trans-former	Occupation	Proprietary resume data	Foundation: 18 hrs of pre-training for each config. Fine-tuning cheap.
Labor LLM	Occupation (as text)	Automated Latent embed dimension: 4096-8192	7B, 13B	Text repr. of survey data. Fine-tune on all tokens in template.	Off-the-shelf LLM	Text	Public text corpus (maybe scraped resume data)	Only fine-tuning cost. 7B: $10, < 20 min to fine-tune on GL40 cards 70b: $50

Demo

- Standard Python libraries: pytorch, hugging-face transformers.
- Model Fine-Tuning:
  - We used Together AI
    - Upload the training data as text.
    - Download the “model checkpoint”.
    - Disclaimer: we are not affiliated with the together AI team.
    - There are other options (e.g., openpipe), you can also fine-tune on your local cluster.
- Model Inference:
  - **Local machine** (e.g., macbook): using 7B/13B models on a few test cases.
  - **Google Colab**: works with the free Nvidia T4 (16GiB).
  - **Cloud (e.g., AWS)**: a GPU with 32GiB vRAM would be ideal.
  - **Computing Cluster**: launch model inference on clusters.
- Please contact us if you want to play with the fine-tuned model.
  - We just need to verify that you have agreed the NLSY/PSID data usage agreement.

More Papers on Artificial Intelligence for Modeling Worker Career Transitions

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.* https://arxiv.org/abs/2202.08370

Vafa, Keyon, Susan Athey, and David M. Blei. "Decomposing Changes in the Gender Wage Gap over Worker Careers." (2023). https://conference.nber.org/conf_papers/f189605.pdf

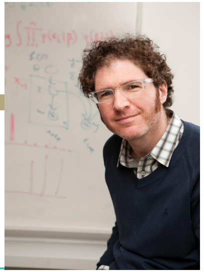
Vafa, Keyon, Susan Athey, and David M. Blei. “Estimating Wage Disparities Using Foundation Models." (2024). https://arxiv.org/abs/2409.09894

[**Presented Today**] 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









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

Thanks!

Questions?

Tianyu Du (*tianyudu@stanford.edu*)