Human Capital in Context: Measuring Firm-level Human Capital with Embeddings

Jung Ho Choi
Stanford Graduate School of Business

Mario Milone
University of California, San Diego

April 2026

Preliminary and Incomplete — Please Do Not Circulate

Abstract

Measuring human capital and its alignment with firms remains a central challenge in economics, finance, and official labor market statistics. Traditional measures rely on observable worker characteristics such as education and experience, which capture only a limited subset of the skills, knowledge, and capabilities that individuals accumulate over time, and offer little purchase on the quality of the match between workers and their employers. In this paper, we develop a new measurement framework that leverages large-scale resume data to quantify human capital and worker–firm match quality at a high granularity. We introduce the *Human Capital Transformer* (HCT), a two-tower neural network that learns latent representations of both workers and firms from millions of career histories. The worker tower processes sequential employer histories through a transformer encoder augmented with static individual covariates, while the firm tower embeds companies using textual descriptions of their business activities. Projecting workers and firms into a shared latent space yields a scalable, embedding-based measure of match quality, defined as the cosine similarity between worker and firm vectors, that can be computed for any realized or counterfactual worker–firm pair. Combining resume data from Revelio Labs with public firm characteristics from Compustat, we propose to validate the measure against individual wages and firm-level outcomes such as profitability, revenue, and stock returns. By turning unstructured career histories into structured, comparable indicators of human capital and match quality, our approach offers a new measurement tool to help bridge the literatures on labor matching, human capital management, and firm performance.

Conflict of Interest Disclosure

The authors have no relevant or material financial interests that relate to the research described in this paper. The resume data used in this study were obtained from Revelio Labs under a standard academic data license; Revelio Labs had no role in the design of the research, the analysis or interpretation of the results, or the preparation of the manuscript, and no right of review prior to dissemination. Jung Ho Choi gratefully acknowledges financial support from the Stanford Graduate School of Business. Mario Milone gratefully acknowledges financial support from the Rady School of Management at the University of California, San Diego. The views expressed in this paper are solely those of the authors and do not necessarily reflect the views of their affiliated institutions, Revelio Labs, or the National Bureau of Economic Research.

1. Introduction

Human capital is widely recognized as a primary driver of firm performance, yet its measurement remains one of the most persistent challenges in economics and finance. As Batish et al. (2021) observe, the fundamental difficulty lies in measuring the contribution of human capital to corporate strategy and performance. While some studies have linked human capital management (HCM) proxies to future outcomes—for example, Edmans (2011) documents a positive association between employee satisfaction and long-term stock returns—these measures capture at most one dimension of the underlying construct. Education and experience, the workhorses of the human capital literature since Becker (1994) and Mincer (1974), reflect important but narrow components of the skills, knowledge, and capabilities that individuals accumulate over their careers.

In this paper, we take a fundamentally different approach. Rather than relying on predefined observable attributes, we learn a latent representation of human capital directly from individuals' career histories adapting the CAREER model by Vafa et al. (2024). Leveraging millions of resumes from Revelio Labs, we develop the *Human Capital Transformer* (HCT), a machine learning framework that encodes both workers and firms as vectors in a shared latent space. The core insight is that an individual's sequence of employers reveals information about their accumulated human capital that extends far beyond what standard covariates can capture. A worker who moves from a semiconductor design firm to a cloud computing company to a fintech startup reveals a pattern of skills and adaptability that no combination of education and tenure variables can fully represent.

Our framework builds on the two-tower architecture widely used in modern recommendation systems and representation learning (Vaswani et al., 2017; van den Oord et al., 2019; He et al., 2020; Chen et al., 2020). The worker tower employs a transformer encoder—the same class of models underlying recent advances in natural language processing and semantic embeddings (Devlin et al., 2019; Reimers and Gurevych, 2019; Vafa et al., 2024)—to process the sequential history of a worker's employers, combined with static individual characteristics such as education, occupation, and location. The firm tower encodes companies using textual descriptions of their business activities through a sentence-transformer model. Both towers produce fixed-length embedding vectors in the same d-dimensional space, and the relevance between any worker-firm pair is scored by their dot product. The model is trained using an InfoNCE contrastive objective on observed employer

matches: intuitively, it learns to place workers close to the firms they actually work for, while pushing them away from firms they do not.

This design yields several advantages. First, match quality—defined as the cosine similarity between a worker's embedding and a firm's embedding—provides a continuous, interpretable measure of workforce–firm alignment that can be computed for any worker-firm pair, including counterfactual matches. Second, the two-tower structure is inherently scalable: firm embeddings can be pre-computed and indexed, enabling efficient retrieval across the entire universe of companies. Third, because both towers map to the same latent space, the resulting embeddings support a wide range of downstream analyses, from individual-level wage regressions to firm-level performance prediction and portfolio construction.

Our empirical strategy proceeds in two stages. In the first stage, we estimate firm–worker match quality by jointly modeling workers' career histories and firms' characteristics. Using data through 2020 for training, we produce worker embeddings that summarize accumulated human capital as of 2021 and firm embeddings that capture organizational skill demands. We evaluate the model using standard retrieval metrics on held-out data to assess whether the learned embeddings capture meaningful information about the worker–firm matching process.

In the second stage, we use the estimated embeddings to examine the economic content of our match quality measure. At the worker level, we study the relationship between match quality and wages, motivated by productivity-based matching theories (Lazear, 2009; Jovanovic, 1979; Taber and Vejlin, 2020). We also analyze how match quality varies across firm size, industry, and worker demographics. At the firm level, we investigate whether aggregated match quality is associated with accounting performance (such as return on assets and revenue) and with stock returns, including portfolio strategies that sort firms on their average match quality. This preliminary work is an exploratory first step: we view the HCT primarily as a new measurement tool, and our aim at this stage is to document whether the resulting embeddings carry economically meaningful signal for both individual and firm-level outcomes.

We position our contribution relative to three strands of literature. First, relative to the CAREER model of Vafa et al. (2024), which produces individual-level embeddings from job title sequences using a single transformer encoder, our framework additionally constructs firm embeddings in the same latent space, enabling direct measurement of match quality and aggregation across em-

ployees. Second, relative to the Job Assignment Quality (JAQ) measure of Coraggio et al. (2025), which estimates the probability of a worker being assigned to a particular job within a firm using random forests trained on high-performing firms, our approach is not limited to a fixed assignment probability from a benchmark sample. Instead, the learned representations can be fine-tuned for downstream tasks and applied to specialized populations such as corporate directors. In addition, our approach is context-dependent demonstrating that the same worker might be more productive for one company vs the other company. Third, relative to asset embeddings (Gabaix et al., 2025), which learn security representations from investor portfolio holdings, our framework grounds embeddings explicitly in human capital rather than financial returns, while permitting interpretable quality measures between firms and individuals as well as potentially a team formation strategy.

Bibliography

Batish, A., A. Gordon, J. D. Kepler, D. F. Larcker, B. Tayan, and C. Yu (2021, May). Human Capital Disclosure: What Do Companies Say About Their 'Most Important Asset'? Available at SSRN 3840412.

Becker, G. S. (1994, March). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Chicago, IL: University of Chicago Press.

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton (2020, July). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, Volume 119 of *ICML'20*, pp. 1597–1607. JMLR.org.

Coraggio, L., M. Pagano, A. Scognamiglio, and J. Tåg (2025, February). *JAQ* of all trades: Job mismatch, firm productivity and managerial quality. *Journal of Financial Economics 164*, 103992.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics.

Edmans, A. (2011, September). Does the stock market fully value intangibles? Employee satisfaction and equity prices. *Journal of Financial Economics 101*(3), 621–640.

Gabaix, X., R. S. J. Koijen, R. Richmond, and M. Yogo (2025, July). Asset Embeddings. Available at SSRN 4507511.

He, K., H. Fan, Y. Wu, S. Xie, and R. Girshick (2020, June). Momentum Contrast for Unsupervised Visual Representation Learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.

Jovanovic, B. (1979, October). Job Matching and the Theory of Turnover. *Journal of Political Economy 87*(5, Part 1), 972–990.

Lazear, E. P. (2009, October). Firm-Specific Human Capital: A Skill-Weights Approach. *Journal of Political Economy 117*(5), 914–940.

Mincer, J. A. (1974). *Schooling, Experience, and Earnings*. NBER.

Reimers, N. and I. Gurevych (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980–3990.

Taber, C. and R. Vejlin (2020). Estimation of a Roy/Search/Compensating Differential Model of the Labor Market. *Econometrica 88*(3), 1031–1069.

Vafa, K., E. Palikot, T. Du, A. Kanodia, S. Athey, and D. M. Blei (2024, February). CAREER: A Foundation Model for Labor Sequence Data. arXiv preprint arXiv:2202.08370.

van den Oord, A., Y. Li, and O. Vinyals (2019, January). Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017, December). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Red Hook, NY, USA, pp. 6000–6010. Curran Associates Inc.