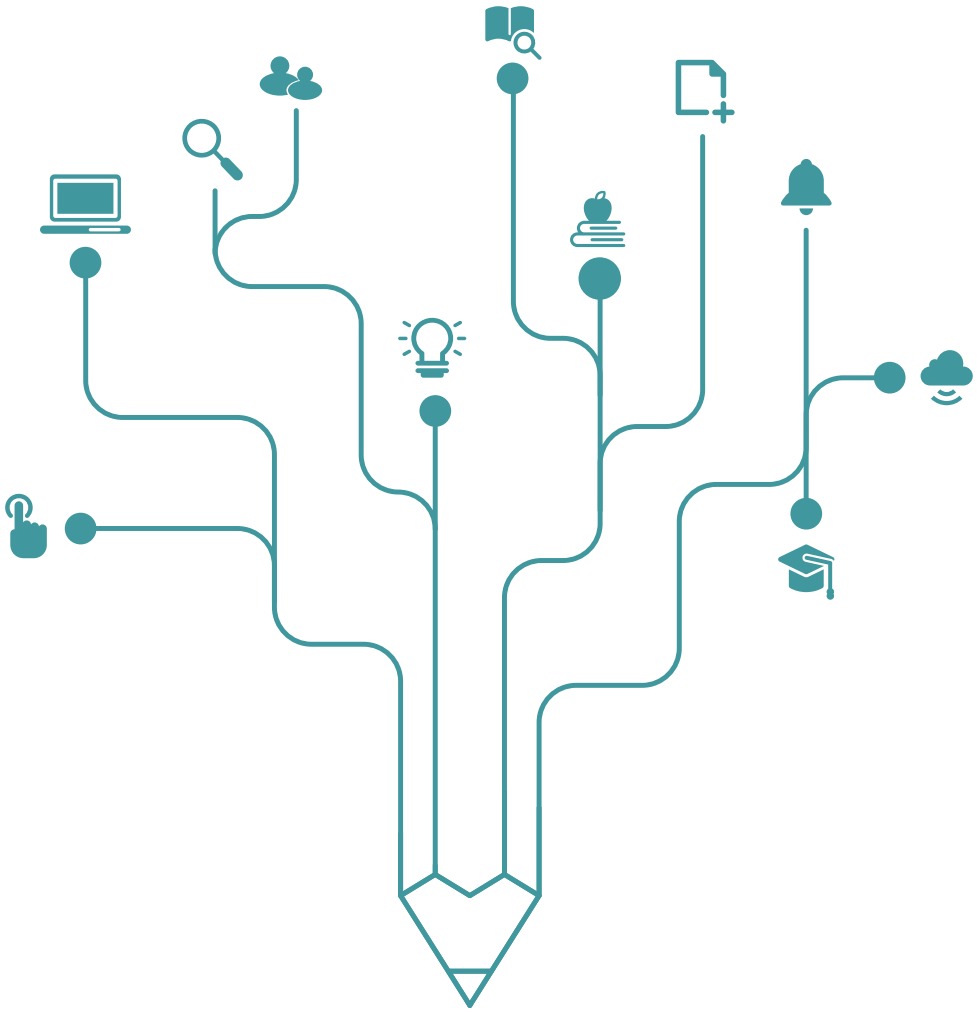
KDIS Working Paper 23-11

Labor Space: Unifying Representation of the Labor Market via Large Language Models

Seongwoon Kim (KDI School of Public Policy and Management)
Yong-Yeol Ahn (Indiana University)
Jaehyuk Park (KDI School of Public Policy and Management)



KDI SCHOOL
KDI School of Public Policy and Management

Labor Space: Unifying Representation of the Labor Market via Large Language Models

Seongwoon Kim
School of Public Policy and Management
Korea Development Institute
Sejong, Republic of Korea
tjddns0032@gmail.com

Yong-Yeol Ahn
Luddy School of Informatics, Computing, and Engineering
Indiana University
Bloomington, IN, USA
yyahn@iu.edu

Jaehyuk Park*
School of Public Policy and Management
Korea Development Institute
Sejong, Republic of Korea
jpark@kdis.ac.kr

ABSTRACT

The labor market is a complex ecosystem comprising diverse, interconnected entities, such as industries, occupations, skills, and firms. Due to the lack of a systematic method to map these heterogeneous entities together, each entity has been analyzed in isolation or only through pairwise relationships, inhibiting comprehensive understanding of the whole ecosystem. Here, we introduce *Labor Space*, a vector-space embedding of heterogeneous labor market entities, derived through applying a large language model with fine-tuning. Labor Space exposes the complex relational fabric of various labor market constituents, facilitating coherent integrative analysis of industries, occupations, skills, and firms, while retaining type-specific clustering. We demonstrate its unprecedented analytical capacities, including positioning heterogeneous entities on an economic axes, such as 'Manufacturing–Healthcare and Social Assistance'. Furthermore, by allowing vector arithmetic of these entities, Labor Space enables the exploration of complex inter-unit relations, and subsequently the estimation of the ramifications of economic shocks on individual units and their ripple effect across the labor market. We posit that Labor Space provides policymakers and business leaders with a comprehensive unifying framework for labor market analysis and simulation, fostering more nuanced and effective strategic decision-making.

CCS CONCEPTS

• **Applied computing** → **Economics**; • **Computing methodologies** → *Information extraction*; • **Information systems** → *Environment-specific retrieval*.

KEYWORDS

Labor Market, Word Embedding, Skill, Job, Industry, Firm, Large Language Model

ACM Reference Format:
Seongwoon Kim, Yong-Yeol Ahn, and Jaehyuk Park. 2024. Labor Space: Unifying Representation of the Labor Market via Large Language Models. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3589334.3645464

*Corresponding author

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
WWW' 24, May 13–17, 2024, Singapore
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0171-9/24/05.
https://doi.org/10.1145/3589334.3645464

1 INTRODUCTION

Understanding the labor market is essential to comprehend the entire economy. A strong understanding of the labor market can not only provide macroeconomic status, such as unemployment rate and household income, but also reveal where our entire economic system moves toward by grasping the complex inter-relationship between various economic units in the economy.

Human capital, encompassing the diverse skills and jobs populating the labor market, is the basis upon which industries and firms are built [10, 40, 42, 54]. It represents the nuanced interplay of individual capabilities and the broader economic forces. For instance, the rise of a new technological skill could dictate the trajectory of entire industries, and subsequently the businesses within them [1, 3, 27]. Hence, understanding this connection between human capital and larger economic entities is crucial, not only for academic comprehension but also for pragmatic decision-making in industries and policymaking.

Yet, despite its significance, a holistic framework that encompasses these interconnections has been absent [25]. Most existing methods either focus only on a specific entity [10, 29, 31, 33] or pairwise relationships [47, 48, 51], without considering the entire structure where different types of entities are entangled. This gap in analysis often results in oversimplified representations and conclusions, limiting the depth of insights that can be derived about the labor market's ecosystem.

Here, we present *Labor Space*, a unifying representation of the heterogeneous entities in the labor market. By leveraging the capabilities of a large language model — Google's BERT — with additional fine-tuning with representative descriptions of the labor market entities from various corpora, we derive an unifying representation of the labor market's constituents, including skills, occupations, industries, and firms. The landscape of Labor Space provides a holistic framework, by exposing the relational fabric of various labor market constituents. At the same time, it facilitates the clustering of related industries, occupations, skills, and firms, while retaining type-specific clustering.

By systematically mapping heterogeneous units onto sensible semantic axes, such as 'Manufacturing-–Healthcare and Social Assistance', Labor Space offers unprecedented analytical capacities. Furthermore, through intricate vector arithmetic of these entities, Labor Space enables the exploration of complex inter-unit relations,

Table 1: Description Data and Source

Entity	Data Source	Number of Entities	Example
Industry	NAICS	308	Metal ore mining comprises establishments primarily engaged in developing mine sites or mining metallic minerals, and establishments primarily engaged in ore dressing and beneficiating (i.e., preparing) operations, such as crushing, grinding, washing, drying, sintering, concentrating, calcining, and leaching. ~
Occupation	O*NET	1,016	Data scientists develop and implement a set of techniques or analytics applications to transform raw data into meaningful information using data-oriented programming languages and visualization software.
Skill	ESCO	307	Counseling assists others to gain access to social, legal or other services and benefits, including making referrals to other professionals and organizations.
Firm	Crunchbase	489	Meta is a social technology company that enables people to connect, find communities, and grow businesses. Previously known as Facebook, Mark Zuckerberg announced the company rebrand to Meta on October 28, 2021 at the company's annual Connect Conference. ~

thereby allowing estimation of the impact of economic shocks or adoption of new technology, such as artificial intelligence, on individual entities and their ripple effect across the labor market.

Our overarching goal with Labor Space is to bridge the existing analytical gap, connecting the dots between individual skills and jobs and the vast expanse of industries and firms. This work underscores the significance of viewing the labor market not as isolated silos but as an interconnected web, where shifts in human capital can ripple across the entire economic ecosystem. We expect that our Labor Space empowers stakeholders, from business leaders to policymakers, with a holistic view of the labor market, enabling more informed, strategic, and effective decisions.

2 RELATED WORKS

2.1 Analysis of the labor market and its entities

The labor market, a multifaceted and intricate ecosystem, is woven together by interconnection between entities like industries, occupations, skills, and firms. Historically, research into this domain has predominantly fallen under two analytical frameworks, which have evolved as established paradigms through cumulative research: (1) understanding the influence of one entity over another, and (2) elucidating the structure of one entity in juxtaposition with another.

The first paradigm largely revolves around the human capital perspective. Human capital in economic theory, denotes the qualitative attributes of labor, extends traditional production factors, such as land, labor, and tangible capital, incorporating a worker's skills and expertise. This perspective delves deep into the impact of investments made in human resources, scaling from individual workers to overarching industries and firms [11, 41]. Existing human capital studies have ventured into areas like knowledge spillovers [5, 52, 53, 55], in addition to the private and societal returns linked to industry- and occupation-specific human capital [44, 48, 51].

Network analysis in economics, meanwhile, has illuminated intricate relationships between different entities. For instance, the skill network, inferred from the co-occurrence of skills among job seekers and providers, suggests that individuals with a diverse skill set command higher wages; more remarkably, those synergizing their varied skills earn top-tier wages [7]. Further exploring this, a skill network pinpoints the dichotomy between physical and cognitive skills based on job requirements [4]. Additionally, the ties binding educational paths, as determined by collaboration within organizations, have underscored the wage benefits of working alongside those with complementary qualifications [46]. Beyond skills, the labor market research also delves into the broader associations between firms and industries, helping demystify the macro-structures of the global economy [45, 49].

However, despite the valuable insights these studies offer, there remains an important oversight in their approach. Much of the existing research tends to examine entities in isolation or restricts itself to binary relationships, whose narrowed focus hinders a holistic grasp of the labor market's nuances. A significant limitation has been the absence of a cohesive methodology capable of simultaneously mapping the varied units of the labor market. This gap in research underscores the urgent need for a more integrated analytical framework that bridges the existing divides.

2.2 Application of language models in social science

Language models, such as word embedding, transform high-dimensional or unstructured text data into lower-dimensional representations [38]. In these models, embedding vectors are derived from the relationships of words within sentences, and each vector captures the semantic essence or underlying meaning of its corresponding word. With the introduction of word or document embedding models into the social sciences [39], a new avenue has emerged for analyzing and understanding cultural, social, and historical trends [36].

Through the use of word embeddings in social science, deeper insights into historical, cultural nuances, and societal biases have been achieved. An exploration of the vast corpus of English-language Google Books revealed the evolution and persistence of social group representations over two centuries, using word embeddings from 850 billion words [19].

A deep dive into semantic biases, trained on standard web text, showed that our historical biases, ranging from the benign to the problematic, are deeply embedded in everyday human language [17]. This finding aligns with studies that highlighted the changing dynamics of stereotypes, especially in terms of gender and ethnicity in the U.S. during the 20th and 21st centuries. Notably, these models identified significant societal events, such as the 1960s women's movement and immigration trends [28].

Beyond social stereotypes, the flexibility of word embeddings has been showcased in other domains. Insights into socio-economic

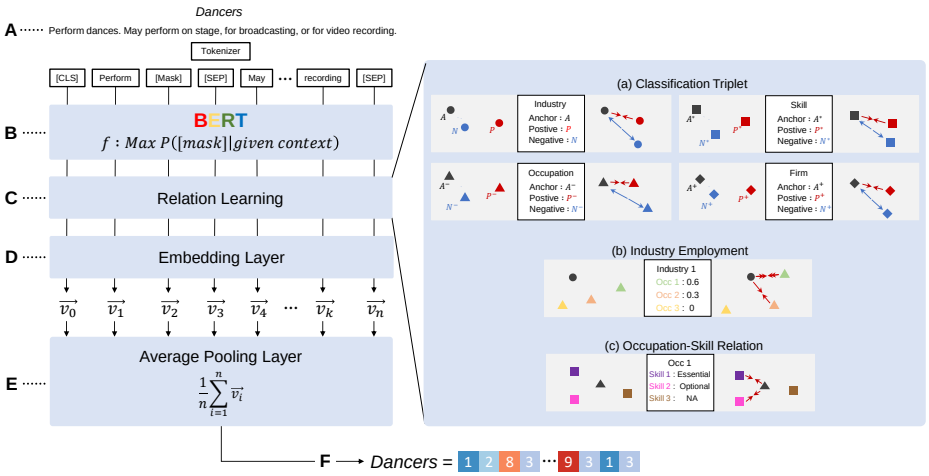


Figure 1: Constructing the Labor Space. (A) Sample entity description from the 2,120 available. (B) Google's BERT, fine-tuned with descriptions from NAICS, O*NET, ESCO, and Crunchbase, predicts the [Mask] token using its context, learning labor market nuances. (C) training inter-relations between Labor Space entities using paired datasets, as magnified in the right-side figure. (D) Both contextual and relational information is captured in BERT's final hidden layer, from which we extract word vectors. (E) A full description vector is represented by averaging its word vectors. (F) Each vector is then labeled with its corresponding title.

classes and their transformations over a century were derived using word embeddings, linking cultural theories with semantic relations in high-dimensional spaces [32]. Using embeddings from Google News articles, the persistent gender biases in occupational terms evident even in modern media were unveiled [15]. In the realm of management science, the importance of language models in business analytics was emphasized. A 'culture dictionary' developed using a word embedding model trained on earnings call transcripts connected corporate culture with tangible business outcomes, expanding our perspective on corporate innovation beyond just factors like R&D spending [35].

In economics, language models serve to develop new economic measurements in the labor market. For instance, language models are utilized to match patent abstracts with occupational tasks, differentiating between labor-augmenting and labor-automating innovations [8], and to explore the connection between university syllabi and occupational tasks, to shed light on the specific skills imparted in higher education and their relevance to job opportunities and earnings [20]. Moreover, a recent study combines university syllabi with cutting-edge research articles and patent abstracts to gauge the gap between education and frontier knowledge [14].

However, while these aforementioned studies have been monumental in their revelations and have extensively advanced our understanding of social dynamics using word embeddings, they predominantly emphasize individual layers of units and their associated relationships. In this study, we aims to offer a more interconnected view, integrating various elements of the labor market into a cohesive space. By navigating the relationships between diverse entities, we propose a novel framework to utilize word embeddings in social science research, especially in the context of the labor market.

3 DATA AND METHODS

3.1 Entity descriptions

3.1.1 NAICS. We use the North American Industry Classification System (NAICS) to embed industry entities into the Labor Space, as the examples in Table 1. NAICS is the standard classification system used by federal statistical agencies in the United States to classify business establishments. It maps business activities onto a hierarchy of industry classifications ranging from 2-digit to 6-digit, depending on the scope and range of the activity. In this study, we focus on the 4-digit industry classification, which includes 308 distinct titles and descriptions.

*3.1.2 O*NET.* The occupation information is derived from the Occupational Information Network (O*NET). The O*NET data comprehensively describes various professions in the contemporary American workplace and is widely used in academic research. Table 1 provides an example of an occupation description for 'data scientist'. Our analysis focuses on 1,016 distinct occupation titles and descriptions from the O*NET 27.3 database.

3.1.3 ESCO. The skill components are derived from European Skills, Competences, Qualifications, and Occupations (ESCO), a multilingual classification system for the European workforce. It offers roughly 15,000 skill units and a hierarchy system ranging from level 0 to level 3. We selected a level 3 skill hierarchy encompassing 307 distinct skill names and descriptions for our analysis. An example description is presented in Table 1.

3.1.4 Crunchbase. The firm entities are sourced from *Crunchbase.com*, a platform that provides information on companies, investors, and industry trends. Here, as a representative sample of firms, we selected the firms listed as the S&P 500 companies in the U.S. stock market and extracted their descriptions. As an example,

the description of a social media company 'Meta' is presented in Table 1.

3.2 BERT model

To quantify the conceptual similarities among heterogeneous types of entities in the labor market, we use the widely-adopted pre-trained word embedding model, Bidirectional Encoder Representations from Transformers (BERT) [22]. Existing ttudies have shown that embedding models are capable of representing rich semantic relationships between words through spatial relationships in a vector space [6, 23, 38, 39, 43].

BERT, as an encoder model, is the de-facto standard for contextual representation model. It has been widely employed, especially with fine-tuning, to achieve breakthrough performance in various natural language processing tasks [22].

3.3 Fine-tuning for context learning

While the base BERT model is considered reliable for general context, existing studies have reported the relatively lower performance of the model capturing the micro-relationships of entities focusing on a specific context, such as scientific or medical text [12, 18, 34]. Hence, we fine-tune the original BERT model in two ways to capture the latent structure of the labor market. First, we use *HuggingFace*'s "fill mask" pipeline for context learning. Here, the context learning aims to adjust the pre-trained model to the context of the labor market, through additional training with a domain-specific corpus for each entity. As a domain-specific corpus, we concatenate (1) 308 NAICS 4-digit descriptions, (2) O*NET's descriptions for 36 skills, 25 knowledge domains, 46 abilities, 1,016 occupations, (3) ESCO's descriptions for 15,000 skills, 3,000 occupations, and (4) 489 Crunchbase S&P 500 firm descriptions, excluding their labels. We set the maximum token length to 512 and configured the hyperparameters for three epochs, using a batch size of 8 and a learning rate of 2e-5 on an RTX 3080 Ti GPU.

3.4 Fine-tuning for relation learning

After the initial fine-tuning to embed the context of the labor market, we conducted an additional fine-tuning process to incorporate inter-entity relatedness. Inspired by the recent work by Cohan et al. [21], which builds interconnections between scientific papers using citation networks, we constructed the following three datasets to train the connections between different types of labor market entities: (1) classification triplet examples, (2) industry-occupation pairs, and (3) occupation-skill pairs.

First, the classification triplet consists of three items — an anchor, a positive sample, and a negative sample. The anchor is the target item that our model aims to learn the relational representation for, while the positive and negative examples are items related and unrelated to the target, respectively. We randomly assign the anchor from 308 industries, 1,016 occupations, 307 skills, and 489 firm descriptions. Positive and negative items are then assigned by leveraging the classification hierarchy system. An entity description is assigned to the positive sample if they share the same parent in their classification system, and to the negative sample if they do not. For each type of entity, we use the following corresponding classification systems: 2-digit NAICS classes for industries, 2-digit SOC's classes for occupations, second-level ESCO classes for skills, and 2-digit General Industry Classification System (GICS) classes for firms. We use triplet loss as the loss function in triplet-based models. It encourages the anchor embedding to be closer to the positive and farther from the negative, improving the model's discriminative ability in the embedding space (Fig. 1C(a)).

Second, we employ the Occupational Employment and Wage Statistics (OEWS) to build connections between industries and occupations. The OEWS provides the number of workers across occupations within each industry. By calculating the proportion of employment for each occupation in an industry, we identify which occupations are most strongly associated with a given industry. With this relational data, we define the relatedness between industries and occupations using a cosine similarity loss function for training our model. This method encourages the model to produce representations that are more similar for industries with a higher proportion of shared occupations (Fig. 1C(b)).

Lastly, to train relations between occupations and skills, we utilize the ESCO dataset, which labels skills as essential, optional, or irrelevant for each occupation. From this information, we construct anchor and positive sample pairs, setting occupation descriptions as anchor samples and skill descriptions as positive samples when the relationship between an occupation and a skill is either essential or optional. We employ the multiple-negatives-ranking loss function to establish the connection between occupations and skills. This loss function uses occupation and skill relational data to adjust the weights so that they are proximal in the vector space (Fig. 1C(c)).

3.5 Obtaining vectors for labor market entities

To obtain vectors for labor market entities, we first process the textual descriptions of the entities, using the BERT's tokenizer function. The BERT tokenizer, known as Wordpiece, encodes raw text data into token sequences and maps these tokens to their respective token IDs. This conversion turns the description into token sequences, which are the smallest semantic units BERT can interpret. BERT then maps these token sequences to a matrix, where each row comprises 768-dimensional vectors representing each token ID (Fig. 1D). To derive a singular representation of the input description, we compute a linear combination of individual word vectors. This is achieved by summing the embeddings of all the words in the sequence and dividing by the word count (Fig. 1E), thus capturing the overall semantic essence of the description. Each description embedding is labeled with its corresponding title for identification.

4 LANDSCAPE OF *LABOR SPACE*

Labor Space provides an embedding environment in which industries, occupations, skills, and firms are plotted in a unifying vector space. The current version includes 308 industries, 1,016 occupations, 307 skills, and 489 firms. However, more types and entities can be integrated when their representative description texts become available. The proximity of entity vectors in our Labor Space, measured via cosine similarity, aptly mirrors their conceptual similarity in the labor market.

Fig. 2A provides a visual representation of the Labor Space in two dimensions by UMAP algorithm [37], where colors distinguish

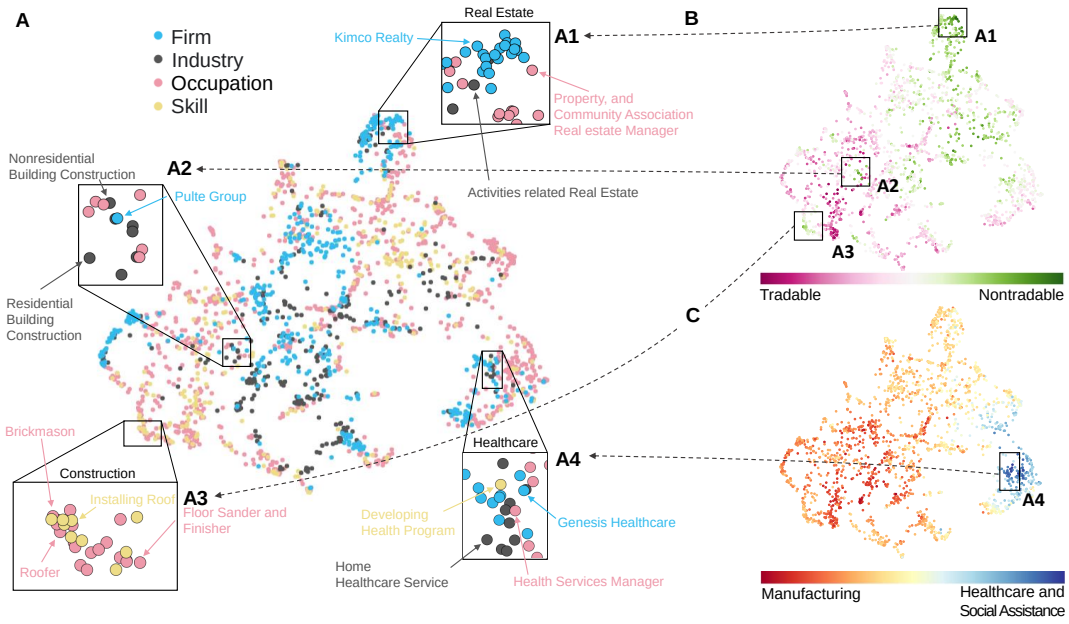


Figure 2: Visualizing Labor Space (A) Labor entities, originally 768-dimensional, are mapped to a 2D space using UMAP. (A1) Highlighted values in the Tradable–Nontradable dimension show close ties with real estate. (A2, A3) Construction-related entities cluster due to the industry's blend of manufacturing and tradability. (A4) Emphasized values on the Manufacturing–Healthcare and Social Assistance dimension show deep ties to healthcare. (B) Map colored by cosine similarity between V(Tradable → Nontradable) and labor vectors; black rectangles indicate locations from A1, A2, A3. (C) Distribution of cosine similarity between V(Manufacturing → Healthcare and Social Assistance) and labor vectors; the black rectangle pinpoints the location in A4.

entity types. Labor market entities align with their conceptual resemblances in the labor market. S&P 500 firms predominantly cluster around specific industries, while occupations and skills bridge the spatial gap, reflecting the diverse skill sets and roles required by large corporations, especially those in the S&P 500. Moreover, diverse labor market entities cluster spatially based on conceptual similarities. For example, entities associated with real estate (Fig. 2A1), construction (Fig. 2A2 and Fig. 2A3), and healthcare (Fig. 2A4) are grouped closely, while there exists a spatial separation based on specific tasks, such as the one between the entities for exterior and interior construction (Fig. 2A2 and Fig. 2A3, respectively). This organized clustering in the Labor Space empowers policymakers and business owners to pinpoint and prioritize the pivotal skills and occupations relevant to a particular industry or company.

5 MAPPING HETEROGENEOUS UNITS ON A CONCEPTUAL AXIS

Using vector arithmetic among its entities, Labor Space offers the capability to map various entities across multiple economic dimensions. Fig. 2B and C depict the relative scores of labor market entities on two distinct axes: (1) the Tradable–Nontradable axis, where nontradable industries encompass local services like restaurants, grocery stores, and salons, and tradable industries comprise businesses that produce exportable or importable products [30]; and (2) the Manufacturing–Healthcare and Social Assistance axis, which exposes the relative similarities between manufacturing and healthcare & service industries.

To structure an axis, we first identify a representative entity for each pole. Then, we determine a conceptual vector transitioning from one pole to the other through vector subtraction [50]. Projecting labor market entities onto this axis vector lets us measure the shared association between the two vectors in a continuous representation [32]. This method facilitates visualizing and quantifying how labor entities are positioned along the axis, such as the Manufacturing–Healthcare and Social Assistance dimension, within Labor Space.

Fig. 2B illustrates the Tradable–Nontradable dimension superimposed on Labor Space. As the tradable and nontradable categories are not explicitly distinguished among our entities, we introduce an auxiliary process to compute the industry centroids (see Appendix A.1). Entities aligned with the nontradable sector, such as real estate (Fig. 2A1), gravitate towards the top regions of the Labor Space. Conversely, tradable sectors like manufacturing and energy predominantly settle at the bottom.

In a similar manner, Fig. 2C displays the distribution of projection values on the Manufacturing–Healthcare and Social Assistance axis throughout Labor Space. Entities linked to manufacturing are mostly situated on the left, whereas those tied to healthcare lean towards the right in the space. A clear transition is noticeable from

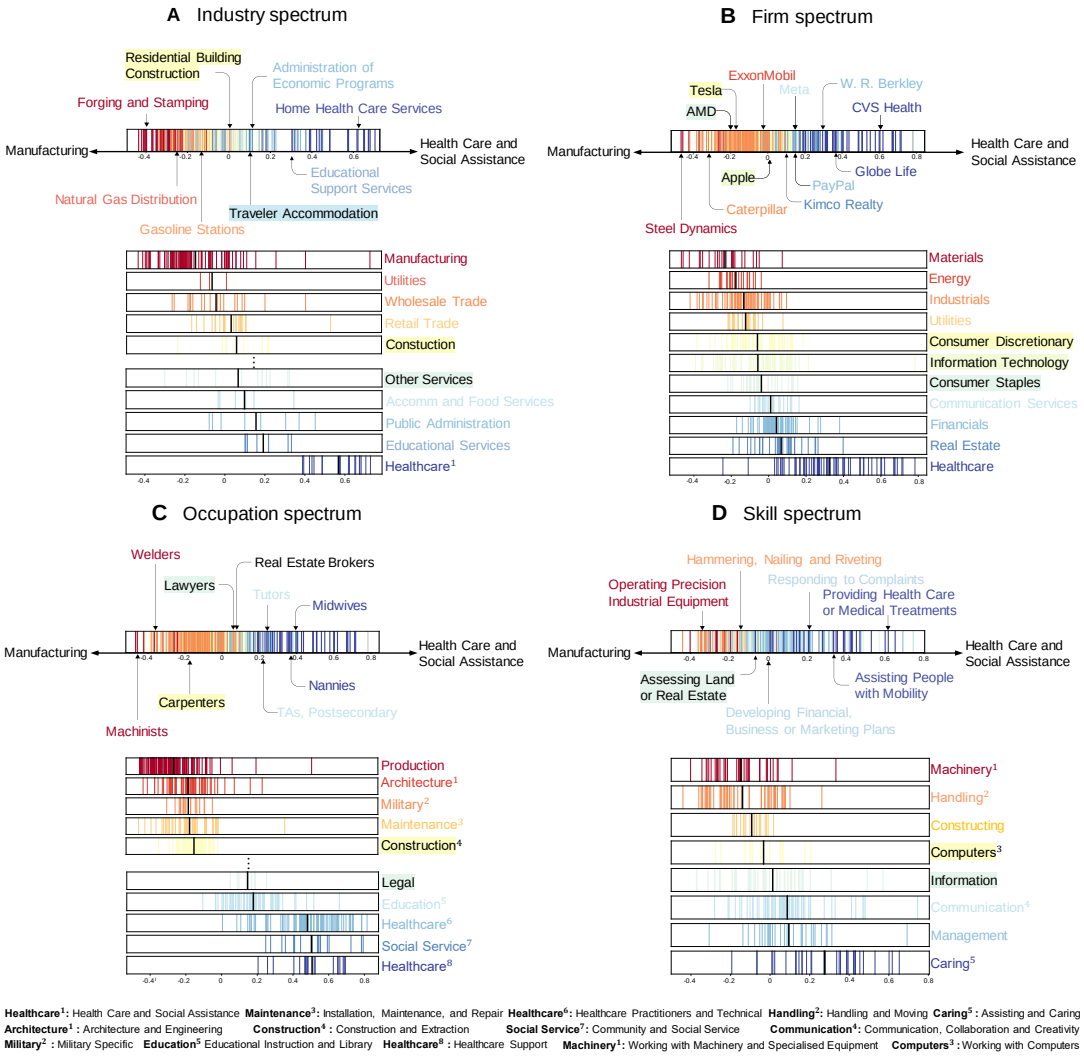


Figure 3: Spectrum Plot of Labor Market Units. All labor entities are projected onto the V(Manufacturing → Healthcare and Social Assistance) axis. (A) Vertical lines within the industry spectrum box show industry embedding projections. Representative industry titles are annotated, using NAICS 2-digit classification for sub-spectrum plotting. (B-D) The same projection method applies for firms (using General Industry Classification System), occupations (using Standard Occupation Classification), and skills (using ESCO skill level two hierarchy).

manufacturing to Healthcare and Social Assistance as we move from left to right.

These projection maps, centered on economic axes, grant a comprehensive perspective of the labor market structure, reinforcing that our embedding space authentically captures labor market dynamics. To further underscore this analytical effectiveness of Labor Space, Fig. 3 portrays the continuous spectrum produced by the projection of labor market entities along the Manufacturing–Healthcare and Social Assistance dimension. Representative entity titles are annotated on this spectrum to spotlight their positions. To validate entity alignment along this axis, we present sub-spectrum plots for each classification system. For industries and occupations, only the top 5 and bottom 5 sub-spectrum plots, sorted by mean projection value for each classification, are accentuated.

Across all labor market entity categories, projecting entities onto the conceptual axis consistently yields reliable outcomes. Firms linked to materials, energy, and industrial utilities (e.g., Steel Dynamics and Caterpillar) are proximate to the manufacturing pole, while those offering services like healthcare, real estate, and finance (e.g., CVS Health and Paypal) are nearer to the Healthcare and Social Assistance pole (Fig. 3B). Likewise, skills and occupations tied to manufacturing lean to the left (e.g., Machinists and Welders), whereas those connected to services shift to the right (e.g., Midwives and Nannies) (Fig. 3C and D). This result underscores

Figure 4: Vector analogy of firm and industry entities.

the robustness of Labor Space in mapping concepts across diverse economic categories.

6 VECTOR ARITHMETIC FOR ECONOMIC ANALOGY

Is it possible to conduct vector arithmetic across types of economic entities? For instance, consider the computation V(Firm A) - V(Skill X) + V(Skill Y). Can such an equation estimate the impact of a new entity's emergence or the absence of an existing entity from one category on an entity in another category?

Beyond an empirical validation of the vector distance in Labor Space, presented in Appendix A.2, vector analogies uncover latent connections between heterogeneous entities in Labor Space. Fig. 4 presents relationships drawn between firms and industries. In this visualization, the formula V(Firm A) - V(Industry B) + V(Industry C) ~ V(Firm D) articulates analogical relationships between firms and their corresponding industries. For instance, leading firms in the beverage and restaurant sectors are analogously seen as Nike in the footwear realm, as illustrated in Fig. 4A, B, and C. Another example from Fig. 4D posits that Amazon, if divested of its web search and IT components but equipped with physical stores, would approximate Walmart within the S&P 500 — a deduction made from the vector equation V('Amazon') - V('Web Search Portals, Libraries, Archives, and Other Information Services') + V('Department Stores') ~ V('Walmart'). Similarly, Tesla, when stripped of its electrical base but supplied with gasoline elements, aligns closely with Ford, as per the equation V('Tesla') - V('Other Electrical Equipment and Component Manufacturing') + V('Gasoline Stations') ~ V('Ford'), reinforcing our intuitive understanding (see Fig. 4E).

Such vector operations encapsulate a myriad of interactions among labor market entities in reality through vector arithmetics across heterogeneous labor market entities. For instance, the equation 'occupation A + industry B + skill C ~ firm D', arguably the most intricate analogy in the labor market, exemplifies how vector analogies can be pragmatically harnessed for career recommendations to job aspirants. In our analysis, V('Mathematicians') + V('Other Investment Pools and Funds') + V('Providing Financial Advice') suggests Principal Financial Group, JP Morgan & Chase, and Goldman Sachs, which have been considered as the best firms for mathematicians eager to serve in investment funds, leveraging their financial abilities. Additional examples are presented in Appendix A.3.

7 ESTIMATING THE IMPACT OF AI

The labor market is currently experiencing a significant shift due to the widespread integration of artificial intelligence (AI) across numerous economic sectors. AI, broadly defined as technology that discerns patterns from data, has reignited apprehensions regarding technological unemployment [9, 13, 25]. Recent concerns about AI's influence on the labor market have spurred efforts to gauge occupation-level AI exposure [2, 16, 24, 26], aiming to guide both academic research and policy-making towards facilitating workers' adaptation to the evolving job landscape. Then, does our Labor Space offer a way to gauge AI's footprint on the labor market, spanning across types?

A standout feature of Labor Space is its inherent scalability. Being a vector space shaped by a language model, it can seamlessly admit any new entities, when there are sufficient descriptive texts at hand. To discover whether Labor Space can gauge the labor market reflecting emerging technologies, specifically AI, we compare our results with a prior study quantifying AI industry exposure (AIIE) and AI occupation exposure (AIOE) [24]. Using the top ten AI application definitions from [24] (see Appendix A.4), we derive estimates for AI's imprint on industries and occupations and then correlate these with the aforementioned AIIE and AIOE scores. We collect descriptions of the top ten AI applications, obtain their average embedding vectors, and compute its cosine similarity with both industry and occupation vectors.

Fig. 5A and B illustrate the correlation between our derived cosine similarity scores for AI applications and the established AIIE and AIOE metrics. The X-axis shows exposure scores from [24], assessing each entity's vulnerability to specific AI applications, while the Y-axis presents the cosine similarity between AI application vectors and corresponding Labor Space entries. These metrics exhibit a robust correlation, affirmed by a Pearson's correlation coefficient of 0.51 (p-value < 0.001), suggesting that such cosine similarity measures with novel technology vectors can effectively gauge AI exposure across labor market facets.

Delving deeper, can refining our definitions yield even sharper insights within Labor Space? Leveraging AIIE and AIOE metrics focused on language modeling applications from [24], we isolate descriptions pertinent to language modeling, derive corresponding vectors, and recalibrate our cosine similarity analyses. The correlation between language modeling exposure scores (X-axis) and our computed cosine similarities (Y-axis) notably strengthens for occupational exposure (rising from 0.47 to 0.59) upon this specification, even as industrial exposure maintains its robust correlation of 0.51 (Fig. 5C and D).

Labor Space, given its methodology, tends to underscore heightened AI exposure risks for entities whose tasks align with capabilities of contemporary AI models. For instance, entities like 'Software Publisher' and 'Foreign Language Teachers' are perceived as more vulnerable, while financial domains register diminished AI exposure, appeared in language modeling AI exposure visualizations as well. While determining the absolute accuracy of these estimations remains a future endeavor, analyzing AI exposure through the prism of Labor Space not only underscores its versatility but also furnishes a virtual arena for stakeholders — policymakers,

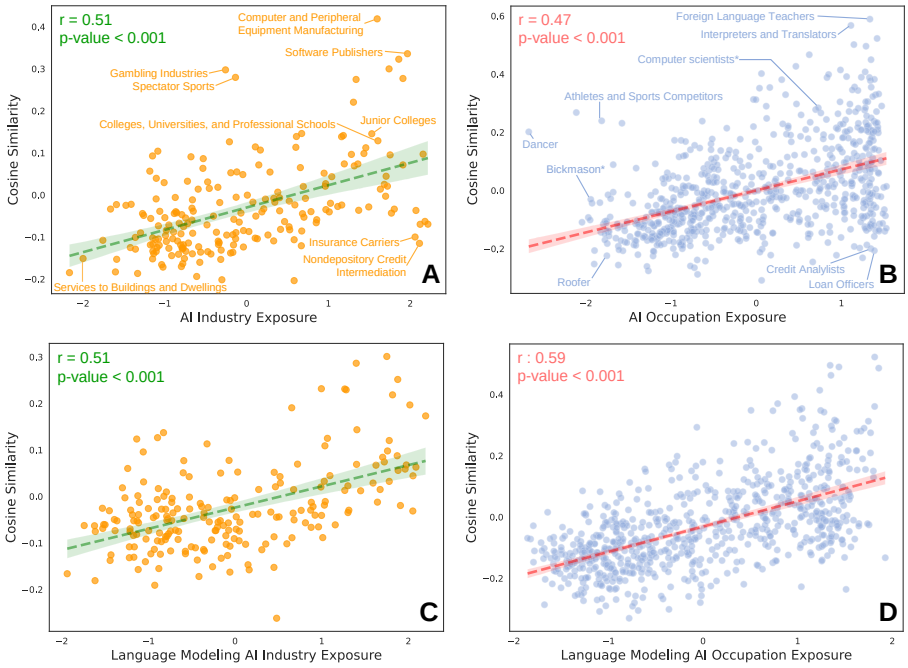


Figure 5: Correlation between the industry-wise (A and C) and occupation-wise (B and D) exposures to AI (A and B) and Language Models (C and D) by [24] (X-axis) and the cosine similarity of of AI and Language Model description vectors and industrial and occupational embedding vectors (Y-axis).

researchers, or business magnates — to conceptualize and simulate potential shifts impacting diverse labor market entities. As an exploratory application, we further apply the same approach to measure the exposure of firms and skills to AI and Language Models, which are shown in Appendix A.5

8 DISCUSSION

The intricate nature of labor markets, characterized by a vast array of interwoven elements ranging from individual skills to expansive industries, has historically posed challenges for comprehensive analysis. Despite numerous studies delving deep into specific facets like individual skills or industry trends, a holistic view connecting micro-level human capital to overarching economic entities has remained largely uncharted.

In this study, we introduce *Labor Space*, a pioneering representation that captures the multifaceted entities within the labor market. Utilizing Google's BERT, a state-of-the-art language model, complemented with further refinement, we have succeeded in crafting an unifying portrayal of the labor market's key elements, encompassing skills, occupations, industries, and firms.

The potential applications of Labor Space are vast. For academic researchers, it provides a comprehensive framework that could lead to more nuanced hypotheses and research questions, considering the interconnectedness of labor market entities. On the practical side, policymakers can leverage this to understand the ripple effects of economic or educational policies on various facets of the labor market. Similarly, industry leaders can harness its insights to make informed decisions about skill development, hiring, and industry partnerships.

Still, there are limitations to consider. Firstly, our reliance on Google's BERT, while powerful, ties our results to the biases and constraints of this model. Although BERT is trained on a vast corpus, it may not capture recent developments in the labor market or might reflect societal biases present in its training data. As the labor market evolves, there's a need for continuous fine-tuning to ensure the relevancy and accuracy of Labor Space.

Another limitation lies in the granularity of data. While Labor Space can cluster related entities, the quality of these clusters largely depends on the input data. Incomplete or outdated data might lead to less accurate representations. Simplifying this for broader audiences, without losing the depth of information, is an area that requires further exploration.

In addition, while Labor Space captures the interconnectedness of various labor market entities, it may not account for external socio-economic factors or global events that can drastically influence the labor market's dynamics. Future iterations could benefit from integrating external datasets or indicators to provide a more holistic view.

In conclusion, Labor Space presents a groundbreaking approach to understanding the labor market as an ecosystem. While it has its limitations, the potential benefits it offers to both the academic and practical realms are immense. As a future direction, refining the methodology and addressing its constraints will be paramount to ensure its continued relevance and efficacy.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation of Korea (NRF) Grant, NRF-2022R1A5A7033499.

REFERENCES

[1] Daron Acemoglu. 2002. Technical change, inequality, and the labor market. *Journal of economic literature* 40, 1 (2002), 7–72.
[2] Daron Acemoglu, David Autor, Jonathon Hazell, and Pascual Restrepo. 2020. *AI and jobs: Evidence from online vacancies.* Technical Report. National Bureau of Economic Research.
[3] Daron Acemoglu and Pascual Restrepo. 2019. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives* 33, 2 (2019), 3–30.
[4] Ahmad Alabdulkareem, Morgan R Frank, Lijun Sun, Bedoor AlShebli, César Hidalgo, and Iyad Rahwan. 2018. Unpacking the polarization of workplace skills. *Science advances* 4, 7 (2018), eaao6030.
[5] Paul Almeida and Bruce Kogut. 1999. Localization of knowledge and the mobility of engineers in regional networks. *Management science* 45, 7 (1999), 905–917.
[6] Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. *arXiv preprint arXiv:1806.05521* (2018).
[7] Katharine A Anderson. 2017. Skill networks and measures of complex human capital. *Proceedings of the National Academy of Sciences* 114, 48 (2017), 12720–12724.
[8] David Autor, Caroline Chin, Anna M Salomons, and Bryan Seegmiller. 2022. *New Frontiers: The Origins and Content of New Work, 1940–2018.* Technical Report. National Bureau of Economic Research.
[9] David H Autor. 2015. Why are there still so many jobs? The history and future of workplace automation. *Journal of economic perspectives* 29, 3 (2015), 3–30.
[10] David H Autor and David Dorn. 2013. The growth of low-skill service jobs and the polarization of the US labor market. *American economic review* 103, 5 (2013), 1553–1597.
[11] Gary S Becker. 1962. Investment in human capital: A theoretical analysis. *Journal of political economy* 70, 5, Part 2 (1962), 9–49.
[12] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).
[13] James Bessen. 2019. Automation and jobs: When technology boosts employment. *Economic Policy* 34, 100 (2019), 589–626.
[14] Barbara Biasi and Song Ma. 2022. *The education-innovation gap.* Technical Report. National Bureau of Economic Research.
[15] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).
[16] Erik Brynjolfsson and Tom Mitchell. 2017. What can machine learning do? Workforce implications. *Science* 358, 6370 (2017), 1530–1534.
[17] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
[18] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559* (2020).
[19] Tessa ES Charlesworth, Aylin Caliskan, and Mahzarin R Banaji. 2022. Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences* 119, 28 (2022), e2121798119.
[20] Hung Chau, Sarah H Bana, Baptiste Bouvier, and Morgan R Frank. 2023. Connecting higher education to workplace activities and earnings. *Plos one* 18, 3 (2023), e0282323.
[21] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180* (2020).
[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[23] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 135–144.
[24] Edward Felten, Manav Raj, and Robert Seamans. 2021. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal* 42, 12 (2021), 2195–2217.
[25] Morgan R Frank, David Autor, James E Bessen, Erik Brynjolfsson, Manuel Cebrian, David J Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, et al. 2019. Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences* 116, 14 (2019), 6531–6539.
[26] Carl Benedikt Frey and Michael A Osborne. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change* 114 (2017), 254–280.
[27] Jordi Gali. 1999. Technology, employment, and the business cycle: do technology shocks explain aggregate fluctuations? *American economic review* 89, 1 (1999), 249–271.
[28] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.
[29] Christina Gathmann and Uta Schönberg. 2010. How general is human capital? A task-based approach. *Journal of Labor Economics* 28, 1 (2010), 1–49.
[30] J Bradford Jensen, Lori G Kletzer, Jared Bernstein, and Robert C Feenstra. 2005. Tradable services: Understanding the scope and impact of services offshoring [with comments and discussion]. In *Brookings trade forum*. JSTOR, 75–133.
[31] Leo Kaas and Philipp Kircher. 2015. Efficient firm dynamics in a frictional labor market. *American Economic Review* 105, 10 (2015), 3030–3060.
[32] Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review* 84, 5 (2019), 905–949.
[33] Edward P Lazear. 1995. A jobs-based analysis of labor markets. *The American economic review* 85, 2 (1995), 260–265.
[34] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.
[35] Kai Li, Feng Mai, Rui Shen, and Xinyan Yan. 2021. Measuring corporate culture using machine learning. *The Review of Financial Studies* 34, 7 (2021), 3265–3315.
[36] Akira Matsui and Emilio Ferrara. 2022. Word embedding for social sciences: An interdisciplinary survey. *arXiv preprint arXiv:2207.03086* (2022).
[37] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
[38] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).
[40] Jacob Mincer. 1958. Investment in human capital and personal income distribution. *Journal of political economy* 66, 4 (1958), 281–302.
[41] Jacob Mincer. 1974. Schooling, Experience, and Earnings. Human Behavior & Social Institutions No. 2. (1974).
[42] Jacob Mincer. 1996. Economic development, growth of human capital, and the dynamics of the wage structure. *Journal of Economic Growth* 1 (1996), 29–48.
[43] Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. *Advances in neural information processing systems* 26 (2013).
[44] Derek Neal. 1995. Industry-specific human capital: Evidence from displaced workers. *Journal of labor Economics* 13, 4 (1995), 653–677.
[45] Frank Neffke and Martin Henning. 2013. Skill relatedness and firm diversification. *Strategic Management Journal* 34, 3 (2013), 297–316.
[46] Frank MH Neffke. 2019. The value of complementary co-workers. *Science advances* 5, 12 (2019), eaax3370.
[47] Russell Ormiston. 2014. Worker displacement and occupation-specific human capital. *Work and Occupations* 41, 3 (2014), 350–384.
[48] Daniel Parent. 2000. Industry-specific capital and the wage profile: Evidence from the national longitudinal survey of youth and the panel study of income dynamics. *Journal of Labor Economics* 18, 2 (2000), 306–323.
[49] Jaehyuk Park, Ian B Wood, Elise Jing, Azadeh Nematzadeh, Souvik Ghosh, Michael D Conover, and Yong-Yeol Ahn. 2019. Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters. *Nature communications* 10, 1 (2019), 3449.
[50] Hao Peng, Qing Ke, Ceren Budak, Daniel M Romero, and Yong-Yeol Ahn. 2021. Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances* 7, 17 (2021), eabb9004.
[51] Paul Sullivan. 2010. Empirical evidence on occupation and industry specific human capital. *Labour economics* 17, 3 (2010), 567–580.
[52] Prasanna Tambe and Lorin M Hitt. 2014. Job hopping, information technology spillovers, and productivity growth. *Management science* 60, 2 (2014), 338–355.
[53] Bram Timmermans and Ron Boschma. 2014. The effect of intra-and inter-regional labour mobility on plant performance in Denmark: the significance of related labour inflows. *Journal of Economic Geography* 14, 2 (2014), 289–311.
[54] Robert Topel. 1999. Labor markets and economic growth. *Handbook of labor economics* 3 (1999), 2943–2984.
[55] Kjersten Bunker Whittington, Jason Owen-Smith, and Walter W Powell. 2009. Networks, propinquity, and innovation in knowledge-intensive industries. *Administrative science quarterly* 54, 1 (2009), 90–122.

A APPENDIX

A.1 Tradability of industry

To make a tradable-nontradable dimension, we set industry centroid with reference to tradability score [30]. Table 2 displays the tradability score for each NAICS 2-digit classification. We designate industries with a score of 100 percent as either tradable or nontradable industry poles.

Table 2: Tradability score

Percent of Industry	Nontradable	Tradable
Accommodation and food services	100	0
Administrative and waste services	89.8	10.2
Agriculture, forestry, fishing, and hunting	0	100
Arts, entertainment, recreation	90	10
Construction	100	0
Educational services	98.89	1.11
Finance and insurance	32.05	67.95
Government	90	10
Healthcare and social assistance	97.8	2.2
Information	34.1	65.9
Manufacturing	0	100
Mining	0	100
Other services	100	0
Professional Services	39.2	60.8
Real estate and rental and leasing	100	0
Retail trade	85.185	14.815
Transportation and warehousing	0	100
Utilities	40	60
Wholesale trade	100	0

Since the "Mining" industry does not align precisely with the current NAICS 2-digit classification, we employ the "Mining, Quarrying, and Oil and Gas Extraction" to encompass it.

A.2 Validation of vector distance using *k*-nearest neighbors algorithm (*k*-NN)

We have conducted a comprehensive validation of our Labor Space model using the *k*-nearest neighbors algorithm (*k*-NN) that adds robustness to our framework. The *k*-NN algorithm is to predict the upper hierarchy class (e.g., NAICS 2-digit-level title for industries, SOC 2-digit-level title for occupations, ESCO skill hierarchy level 2 title for skills, and GICS 2-digit classification title for firms) for each entity within Labor Space based on the classes of its nearest neighbors.

Table 3: Prediction accuracy of the vector distance in Labor Space

	Accuracy			
	$k = 3$	$k = 5$	$k = 7$	$k = 9$
Industry	0.86	0.84	0.83	0.80
Occupation	0.78	0.74	0.72	0.72
Skill	0.80	0.77	0.74	0.73
Firm	0.85	0.82	0.80	0.77

The accuracy, ranging from 0.72 to 0.86, are particularly noteworthy, especially when considering the baseline for multi-class (for instance, the NAICS system has twenty 2-digit classes) classification problems with random selections. These high accuracy levels suggest that the clustering of labor market entities within Labor Space is not arbitrary but is instead based on meaningful semantic relationships. Furthermore, the observation that accuracy decreases as the value of k increases aligns with our expectations, as a larger k implies a broader neighborhood and a potential inclusion of less similar neighbors, which naturally leads to lower accuracy.

A.3 Additional examples for vector arithmetic

Table 4: Vector analogy of heterogeneous labor market units

Formula	Top 3 entities
Occupation - Occupation ~ Occupation V("Data Scientist") - V("Statistician")	1. Data Architects 2. Database Administrators 3. Data Warehousing Specialists
Occupation - Occupation ~ Skill V("Teller") - V("Cashier")	1. Monitoring financial and economic resources and activities 2. Managing budgets or finance 3. Analysing financial and economic data
Occupation + Industry + Skill ~ Firm V("Mathematicians") + V("Other Investment Pools and Funds") + V("Providing Financial Advice")	1. Principal Financial Group 2. JP Morgan Chase 3. Goldman Sachs

A.4 The top 10 AI applications

Table 5: Top 10 AI application

AI application	Definition
Abstract strategy	The ability to play abstract games involving sometimes complex strategy and reasoning ability, such as chess, go, or checkers, at a high level.
Image recognition	The determination of what objects are present in a still image.
Visual question answering	The recognition of events, relationships, and context from a still image.
Reading comprehension	The ability to answer simple reasoning questions based on an understanding of text.
Language modeling	The ability to model, predict, or mimic human language.
Translation	The translation of words or text from one language into another.
Speech recognition	The recognition of spoken language into text.
Instrumental track recognition	The recognition of instrumental musical tracks.
Real-time video games	The ability to play a variety of real-time video games of increasing complexity at a high level.

The Electronic Frontier Foundation (EFF), a respected digital rights nonprofit, has a substantial presence in the academic and research community and collects AI progress statistics from verified sources, including academic literature, blogs, and websites. The EFF selected the top 10 AI applications with recorded scientific progress since 2010, as these are deemed to be experiencing rapid growth and

have medium-term relevance. Table 5 gives the top 10 applications list and brief definitions.

Table 6: Top 10 firms exposed to AI/LM

	AI	Language Modeling
1	NetApp	Google
2	Take-Two Interactive Software	Alphabet
3	Alphabet	EPAM Systems
4	Match Group	Cadence Design Systems
5	Netflix	FactSet
6	Google	Adobe
7	Microsoft	Zebra Technologies
8	Electronic Arts	Gartner
9	Warner Bros. Discovery	NetApp
10	Adobe	Celanese

A.5 AI exposure of firms and skills

This scalability across different types of entities is a unique feature of Labor Space, enabling us to predict the economic impact of various types for which crowdsourced or representative quantified datasets may not exist. Here, the AI exposure scores of firms and skills are measure by their cosine similarities with the conceptual vectors for AI. Table 6 and Table 7 show top 10 firms and skills that are highly exposed to AI and LM, respectively.

Table 7: Top 10 skills exposed to AI/LM

	AI	Language Modeling
1	Using more than one language	Using more than one language
2	Translating and interpreting	Translating and interpreting
3	Using foreign languages	Using foreign languages
4	Technical or academic writing	Using computer aided design and drawing tools
5	Artistic and creative writing	Using digital tools for collaboration, content creation and problem-solving
6	Using digital tools for collaboration, content creation and problem-solving	Artistic and creative writing
7	Using computer aided design and and drawing tools	Designing ICT systems or applications
8	Writing and composing	Technical or academic writing
9	Communication, collaboration and creativity	Browsing, searching and filtering digital data
10	Conducting gaming activities	Accessing and analysing digital data