Labor Space : A High-Dimensional Representation of the Labor Market via Large Language Models

By

KIM, Seongwoon

THESIS

Submitted to

KDI School of Public Policy and Management

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF PUBLIC POLICY

2023

Labor Space : A High-Dimensional Representation of the Labor Market via Large Language Models

By

KIM, Seongwoon

THESIS

Submitted to

KDI School of Public Policy and Management

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF PUBLIC POLICY

2023

Professor Park, Jaehyuk

Labor Space : A High-Dimensional Representation of the Labor Market via Large Language Models

By

KIM, Seongwoon

THESIS

Submitted to

KDI School of Public Policy and Management

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF PUBLIC POLICY

Committee in charge:

Professor Park, Jaehyuk, Supervisor

Professor Yoon, Chungeun

Professor Park, Jinseong

Approval as of December, 2023

Abstract

Labor Space : A High-Dimensional Representation of the Labor Market via Large Language Models

Kim, Seongwoon

The labor market is a complex ecosystem comprising multiple economic units such as skills, jobs, industries, and firms. Hence, a true understanding of the labor market requires a holistic perspective that considers the interrelationships between these entities. However, existing studies have often focused on single or bipartite units; therefore, they don't capture the reciprocal effect of heterogeneous units of the labor market. Here, we introduce *Labor Space*, a high-dimensional space created by a large language model. Labor Space maps industry, firm, occupation, and skill to a unified embedding space representing the conceptual similarity of the labor market entities. Alignment of conceptual dimensions, such as the production-healthcare axis, reveals our numerical representation portrays the industrial structure of the labor market. Moreover, the calculation of the embedding vector catches the latent relationship of the labor market entities and their interactions with external factors, such as the impact of AI on the labor market. Labor Space offers a comprehensive and innovative approach to understanding the interconnectedness of entities within the labor market, providing a pragmatic tool for researchers, policymakers, and business leaders.

Contents

Abstract

1. Introduction

2. Related Works

2.1 Analysis of the labor market and its entities

2.2 Application of language models in social science.

3. Data and Methods

3.1 Descriptions

3.1.1 NAICS

3.1.2 O*NET

3.1.3 ESCO

3.1.4 Cruchbase

3.2 BERT model

3.3 Fine-tuning for context learning

3.4 Fine-tuning for relation learning

3.5 Obtaining vectors for labor market entities

4. Landscape of Labor Space

5. Mapping heterogeneous units on conceptual axis

6. Vector calculation for economic analogy

7. Estimating the impact of AI

8. Discussion

9. Appendix

10. Reference.

1. Introduction

The labor market is a complex economic system where different economic units such as skills, jobs, firms, and industries are interconnected. For instance, the advancement in information technology influences industry, occupation, skill, and firm all together but separately. On the industry level, innovation in IT impacts the information industry, while on the occupational dimension, it affects hiring more programmers. On the skill level, it drives workers to learn a programming language such as C or Python. Firms can also move in response to changes in market conditions, such as mergers and acquisitions of competitive IT ventures. Hence, we can imagine the labor market as a high-dimensional space of multiple types of entities, where each entity is categorized into multiple units — industry, occupation, skill, and firm in the current example.

However, existing studies have focused on a single unit (industry, occupation, skill, or firm) separately or a relationship between two units. They give a unique framework of the single unit in the labor market but don't capture the global structure of the labor market. Here, we create a high-dimensional embedding space of heterogeneous units in the labor market, called *Labor Space*, using the word embedding approach.

In our analysis of the Labor Space, we can quantify the semantic relationships within the labor market system. These methods allow us to determine the proximity or distance between different entities. Interestingly, we can identify specific conceptual dimensions, such as the production-healthcare and tradable-nontradable axes. Also, these axes are particularly informative when understanding the topology of heterogeneous economic units in the Labor Space. We identify spectra of all the labor market entities along the production-healthcare axis, in which spatial movement in the Labor Space indicates a semantic shift in a given dimension. These findings highlight the potential of the Labor Space to reveal complex relationships between various entities in the labor market. Additionally, the vector operation

in the Labor Space quantifies the relationship between target and object, clearly understanding their interconnectedness within the labor market ecosystem. It allows us to discern how closely linked various labor market entities are to AI. For instance, it can reveal whether certain jobs or industries are highly susceptible to AI's influence or, conversely, if they are less affected. In summary, Labor Space allows for measuring conceptual similarity between different entities, giving clues for interaction within the labor market's intricate ecosystem or external factors.

2. Related Works

2.1 Analysis of the labor market and its entities

To gain a comprehensive understanding of the labor market, it is imperative to adopt an ecological and holistic perspective that considers the intricate interplay between these entities. However, since there has been a lack of systematic methods to map and integrate these units into a unified space, many existing studies have focused on examining singular or bipartite aspects of the labor market. For instance, Neffke and Henning (2008) delved into the economy's structural transformation, shedding light on the evolution of the industry space. Neffke and Henning (2013) introduced an index of skill-relatedness among industries, enabling predictions regarding corporate diversification based on labor flows across industries. Alabdulkareem et al. (2018) provided valuable insights by visualizing a skill space, revealing the polarization of cognitive and physical skills as an explanation for wage inequality. Meanwhile, Anderson (2017) offered evidence suggesting that individuals with a broader range of skills tend to receive higher wages than those with narrower, specialized skill sets. Conversely, Anders et al. (2013) highlighted that firm-level factors, such as productivity and trade participation influence within-sector and within-occupation wage inequalities. Bana et al. (2020) constructed an occupation space to characterize how

occupations change over time. While these studies have provided unique insights into the individual components of the labor market, their focus on singular units has limited our ability to comprehend the intricate and multifaceted features of the labor market system, which is inherently driven by interactions among multiple units. A more comprehensive approach is needed to unravel the complexities of this dynamic ecosystem and shed light on its holistic structure.

2.2 Application of language models in social science.

The word embedding has demonstrated its ability to capture semantic or syntactic relations between words [Mikorov et al., 2013a, Pennington et al., 2014, Devlin et al., 2018, Vaswani et al., 2017] and high performance in text generation tasks [OpenAI, 2023]. Numerous studies have employed word embedding techniques within the field of social science. In social science, research using word embeddings delves into several key areas. Some studies have investigated word-to-word relationships within large corpora, or the transformation of it, providing evidence of how language and terminology evolve over time. Kozlowski et al. (2019) demonstrate the use of word embedding models to track semantic transformations related to social class in a large corpus of books, revealing stable cultural dimensions with evolving class markers, notably the growing association of education with wealth. Grand et al. (2022) examine word meaning representation in the mental lexicon using computational models, extracting context-dependent relationships from word embeddings. Additionally, other studies evaluate document similarity, helping to uncover connections and patterns within textual datasets. Chau et al. (2023) connect the university's syllabus and occupational tasks using word embedding to understand the specific skills students learn in higher education and how these skills relate to job opportunities and earnings. Also, Autor et al.

(2022) match patent abstracts to occupational tasks to distinguish whether innovation is labor-augmenting or labor-automation.

Moreover, other applications of word embedding in social science assess the influence of generative language models on various linguistic and semantic aspects, shedding light on how language models impact language understanding and generation in the prospect of social sciences. Eloundou et al. (2023) explore the impact of large language models on the US labor market, finding that various jobs across income levels could see task changes due to these models, potentially leading to increased efficiency and broader implications.

3. Data and Methods

3.1 Descriptions

Table 1 : Data descriptions and sources

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. Beneficiating may be performed at mills operated in conjunction with the mines served or at mills, such as custom mills, operated separately.
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.~

3.1.1 NAICS

We use the North American Industry Classification System (NAICS) as a framework to incorporate industry entities into our Labor Space analysis. An industry description example can be found in Table 1. NAICS is the universally recognized system employed by various federal statistical agencies in the United States to categorize business establishments systematically. This classification system effectively organizes business activities, offering a hierarchical structure that involves 2-digit to 6-digit levels, each signifying the scope and range of specific industry activities. The embedding target is on the 4-digit industry classification, encompassing 308 distinct titles and descriptions.

3.1.2 O*NET

Our occupation data is sourced from the Occupational Information Network [O*NET, 2022], a robust database that offers comprehensive insights into a diverse range of contemporary professions within the American workforce. This data repository is prominent in academic research due to its extensive coverage and reliability. Table 1 presents an illustrative occupation description, explicitly highlighting the role of data scientists. Occupational embedding vectors are 1,016 unique occupation titles and corresponding descriptions drawn from the O*NET 27.3 database.

3.1.3 ESCO

We extract skill components from the European Skills, Competences, Qualifications, and Occupations (ESCO) system, which serves as a multilingual classification system designed for the European labor force. ESCO offers an extensive collection of approximately 15,000 skill units organized within a hierarchical structure from level 0 to level 3. For our analytical

purposes, we specifically focus on skill hierarchy level 3, which comprises 307 unique skill names and associated descriptions. Table 1 displays an example skill from this category.

3.1.4 Cruchbase

We collect firm entities from Crunchbase.com, a comprehensive platform renowned for offering detailed insights into companies, startups, investors, and industry dynamics. We specifically chose representative subset companies listed in the S&P 500 on the U.S. stock market and extracted their descriptive data. Among them, the social media company 'Meta' is available in Table 1.

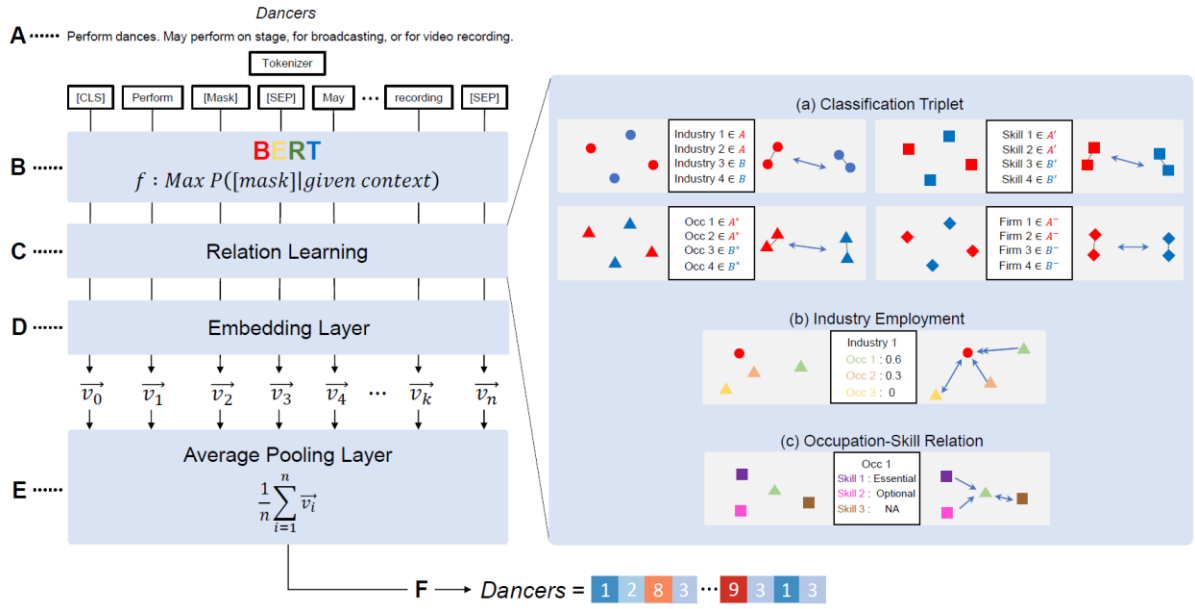

Figure 1 : Constructing the Labor Space

Note : (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) We craft 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. This results in a vectorized representation of diverse labor market units, illustrating their inter-relations and trajectories in the vector space.

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 [Devlin et al., 2018]. Studies have shown that embedding models are capable of representing rich semantic relationships between words through spatial relationships in a vector space [Mikolov et al., 2013a, Mikolov et al., 2013b, Mnih and Kavukcuoglu, 2013, Dong et al., 2017, An et al., 2018]. We choose the BERT, based on the Transformer architecture, which complements the shortcomings of existing word embedding models and achieves breakthrough performance in various natural language processing tasks [Devlin et al., 2018].

3.3 Fine-tuning for context learning

Although the base BERT model excels in general language tasks, it struggles with domain-specific nuances, particularly in scientific or medical texts [Lee et al., 2020, Beltagy et al., 2019, Chalkidis et al., 2020]. We fine-tuned the original BERT model using HuggingFace's 'fill mask' pipeline. We created a domain-specific textual dataset, merging (1) 308 NAICS 4-digit descriptions, (2) O*NET descriptions for 36 skills, 25 knowledge domains, 46 abilities, and 1,016 occupations, (3) ESCO's descriptions about 15,000 skills and 3,000 occupations, and (4) 489 Crunchbase S&P 500 firm descriptions. Our fine-tuning setup comprised a maximum token length of 512, hyperparameters configured for three epochs, a batch size of 8, and a learning rate 2e-5. All training took place on an RTX 3080 Ti GPU.

3.4 Fine-tuning for relation learning

Following our initial fine-tuning process to contextualize labor market information, we do a supplementary fine-tuning process to incorporate inter-entity relationships. This approach

was inspired by recent research, specifically, Cohan et al.'s (2020) work establishing connections between scientific papers through citation networks. To achieve this, we crafted three distinct datasets to facilitate relationship training across different labor market entities: We constructed a classification triplet consisting of three entity descriptions: an anchor, a positive, and a negative sample. The anchor serves as the focal entity for which our model seeks to learn relational representations. Positive samples represent related entities, while negative samples represent unrelated ones. Anchors were randomly chosen from 308 industries, 1,016 occupations, 307 skills, and 489 firm descriptions sequentially. We assigned positive and negative entities by considering the classification hierarchy system. Entities with the same classification system were designated positive samples, while those with differing systems were given negative samples. Unique classification systems were utilized for each entity type, such as 2-digit NAICS classification for industries, 2-digit SOC classes for occupations, second-level ESCO classes for skills, and 2-digit General Industry Classification System (GICS) for firms. The triplet loss function was employed in this dataset, encouraging the anchor embedding to be closer to positive and farther from negative samples, thus enhancing the model's capacity for discrimination in the embedding space.

To establish connections between industries and occupations, we conjugated data from the Occupational Employment and Wage Statistics (OEWS), which offers the number of workers across various occupations within each industry. By computing the proportion of employment for each occupation within an industry, we identified the occupations most strongly linked to specific industries. Using cosine similarity as the loss function, we trained our model to capture the relatedness between sectors and professions.

For training relationships between occupations and skills, we turned to the ESCO dataset, which categorizes skills as essential, optional, or irrelevant for each occupation. From this dataset, we made pairs, with occupation descriptions as anchor samples and skill descriptions

as positive samples, when the relationship between an occupation and a skill was deemed essential or optional. To apply connections between occupations and skills to our Labor Space, we adopted the multiple-negatives-ranking loss function, which adjusts weights based on pair data, bringing occupations and skills closer together in the vector space. This additional fine-tuning process was designed to capture intricate inter-entity relationships within the labor market. It involved three datasets focused on classification triplets, industry-occupation pairs, and occupation-skill pairs, thereby empowering the model to comprehend and represent complex connections among various labor market entities.

3.5 Obtaining vectors for labor market entities

To map entity descriptions to vector space, we utilize the BERT Wordpiece tokenizer to encode raw text into token sequences with associated token IDs. It transforms the original descriptions into sequences of tokens, the semantic units BERT comprehends. BERT then translates these token sequences into a matrix, where each row represents a 768-dimensional vector corresponding to a token ID (as depicted in Fig. 1D).

We perform a linear combination of individual word vectors to obtain a singular representation of the input description. This involves summing the embeddings of all words in the sequence and dividing by the word count (as shown in Fig. 1E). This process captures the overall semantic essence of the description. Each description embedding is associated with its respective title for clear identification.

4. Landscape of Labor Space

The Labor Space is an embedding framework where industries, occupations, skills, and firms are represented in a high-dimensional vector space. The current version encompasses 308 industries, 1,016 occupations, 307 skills, and 489 firms, with the potential for further

integration as long as representative description texts are available. In this Labor Space, the proximity of entity vectors is quantified using cosine similarity, effectively reflecting their conceptual similarity within the labor market. For instance, when considering the occupation

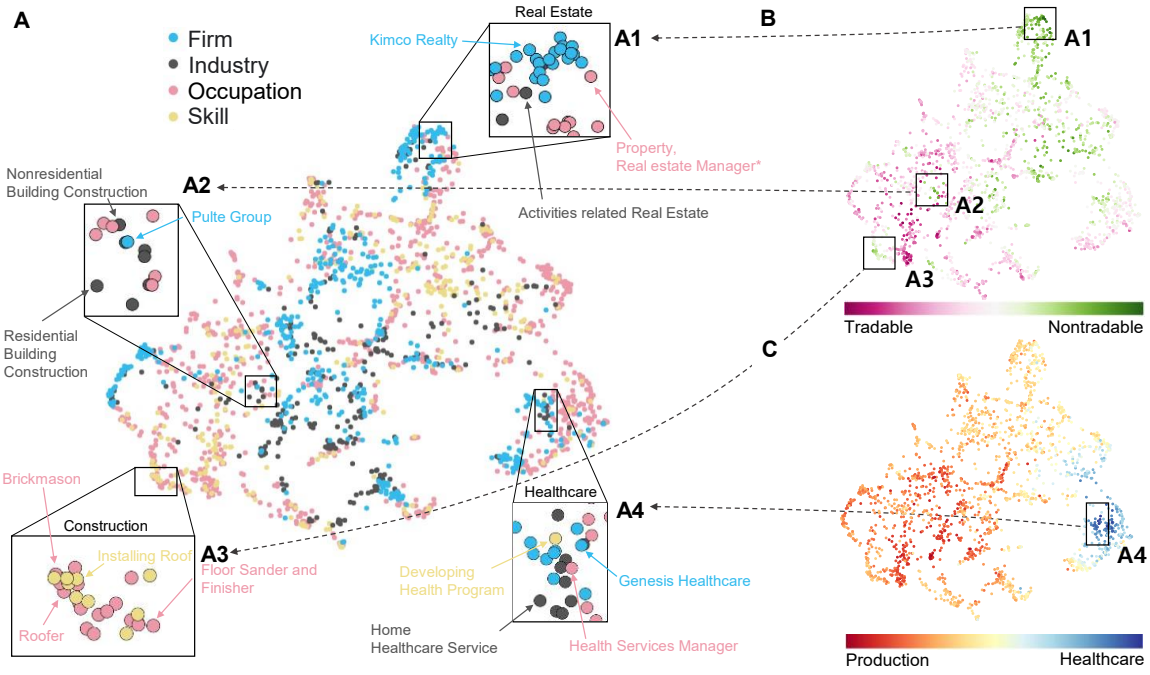

Figure 2 : Visualizing the Labor Space

Note : (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 Production-Healthcare 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, and A3. (C) Distribution of cosine similarity between V(Production → Healthcare) and labor vectors; the black rectangle pinpoints the location in A4.

'Economist,' the entities closest to it in the Labor Space are 'Statistician' (cosine similarity = 0.78) in the occupation level, 'Administration of Economic Programs' (industry level, 0.76), 'Analyzing Financial and Economic Data' (skill level, 0.66), and 'PayPal' (firm level, 0.57). Conversely, the entities farthest from 'Economist' are 'Funeral Attendants' (occupation, -

0.34), 'Death Care Services' (industry, -0.12), 'Applying Textured or Masonry Coatings' (skill, -0.32), and 'John Deere' (firm, -0.13).

Fig. 2A visually represents the Labor Space in two dimensions, with colors distinguishing entity types. One notable aspect of the Labor Space is how diverse labor market entities align with their conceptual similarity. Entities from all categories are distributed uniformly, offering a comprehensive overview of our economy. S&P 500 firms tend to cluster around specific industries, while occupations and skills bridge the spatial gaps. Additionally, entities with shared conceptual similarities tend to cluster spatially, such as those associated with real estate (Fig. 2A1), construction (Fig. 2A2 and Fig. 2A3, and healthcare (Fig. 2A4). This organized clustering within the Labor Space provides valuable insights for policymakers and business owners, enabling them to identify and prioritize essential skills and occupations relevant to specific industries or companies.

5. Mapping heterogeneous units on the conceptual axis

The Labor Space presents a unique capability to map various labor market entities across multiple economic dimensions through vector calculations. Fig. 2B and Fig. 2C illustrate the relative scores of labor market entities on two specific axes:

1. Tradable-Nontradable Axis: This axis distinguishes between nontradable industries, such as local services (e.g., restaurants, grocery stores, and salons), and tradable sectors, encompassing businesses that produce exportable or importable products [Jensen et al., 2005].
2. Production-Healthcare Axis: This axis reveals the relative similarities between the production & manufacturing and healthcare & service industries.

To construct an axis, we first identify a representative entity for each pole and then establish a conceptual vector connecting the two poles through vector subtraction [Peng et al., 2021]. By projecting labor market entities onto this axis vector using cosine similarity calculations,

we can measure their shared association in a continuous representation [Kozlowski et al., 2019]. This approach enables the visualization and quantification of the positioning of labor entities along the industrial dimension within the Labor Space. Fig. 2B depicts the Tradable-Nontradable dimension overlaid on Labor Space. Since the tradable and nontradable industry categories are not explicitly defined among our entities, we employ an auxiliary process to compute industry centroids (see Appendix 9.1). Entities associated with the nontradable sector, like real estate (Fig. 2A1). Similarly, Fig. 2C illustrates the distribution of projection values along the Production-Healthcare axis across Labor Space. Entities linked to production and manufacturing predominantly occupy the left side, whereas those associated with healthcare are situated toward the right. As we move from left to right, a clear transition is evident from the production to healthcare sectors. These projection maps, aligned with economic axes, provide a comprehensive view of the labor market structure, highlighting that our embedding space effectively captures the latent structure of the labor market. In further validating Labor Space's analytical effectiveness, Fig. 3, Fig. 4, Fig. 5, and Fig. 6 present a continuous spectrum resulting from the projection of labor market entities along the Production-Healthcare dimension, with annotated representative entity titles. We offer sub-spectrum plots for each classification system to validate entity alignment along this axis. The top 5 and bottom 5 sub-spectrum plots, sorted by mean projection value for each classification, are plotted for industries and occupations.

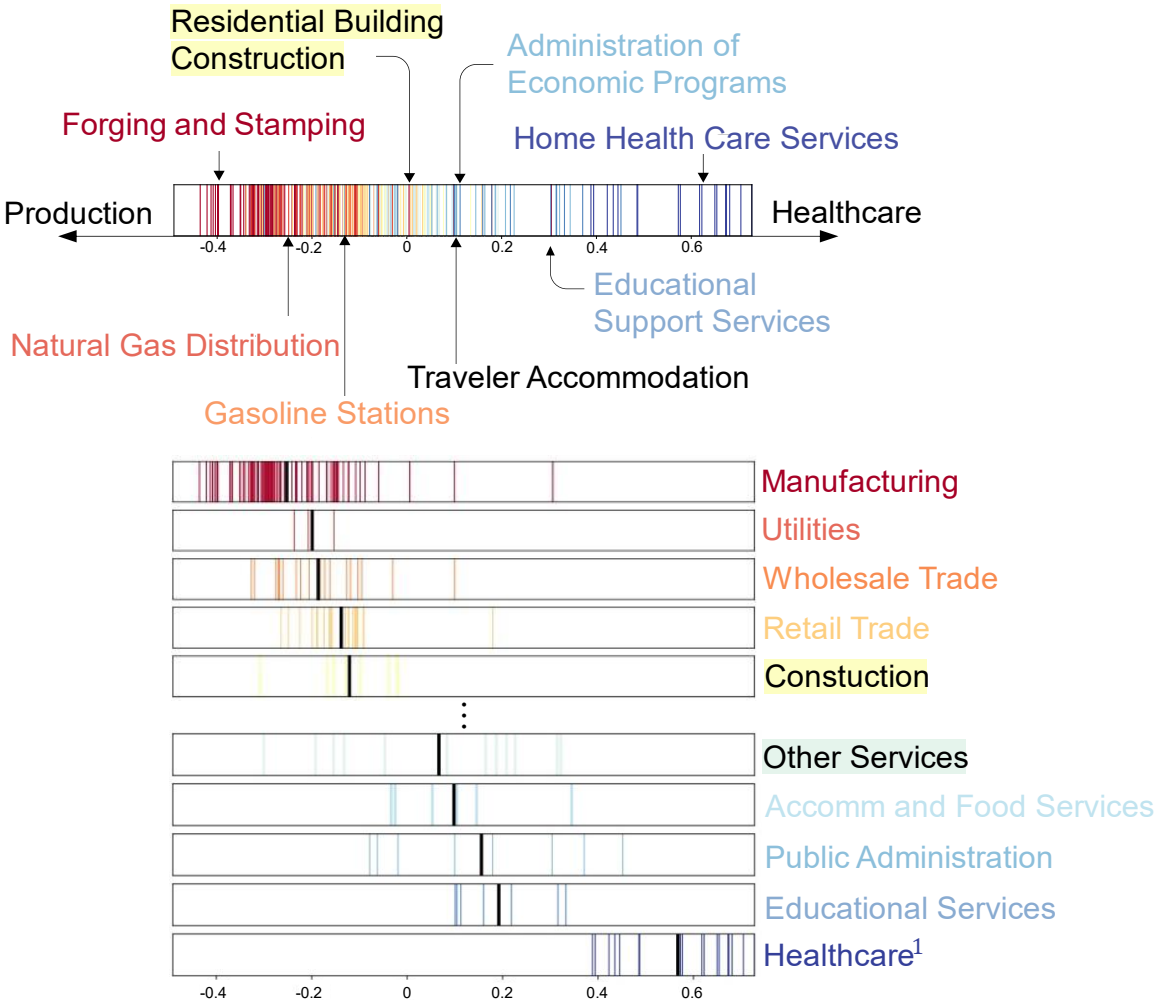

Figure 3 : Industry spectrum

Note : All industry entities are projected onto the V(Production → Healthcare) axis. Vertical lines within the spectrum box show industry embedding projections. Representative industry titles are annotated using NAICS 2-digit classification for sub-spectrum plotting.

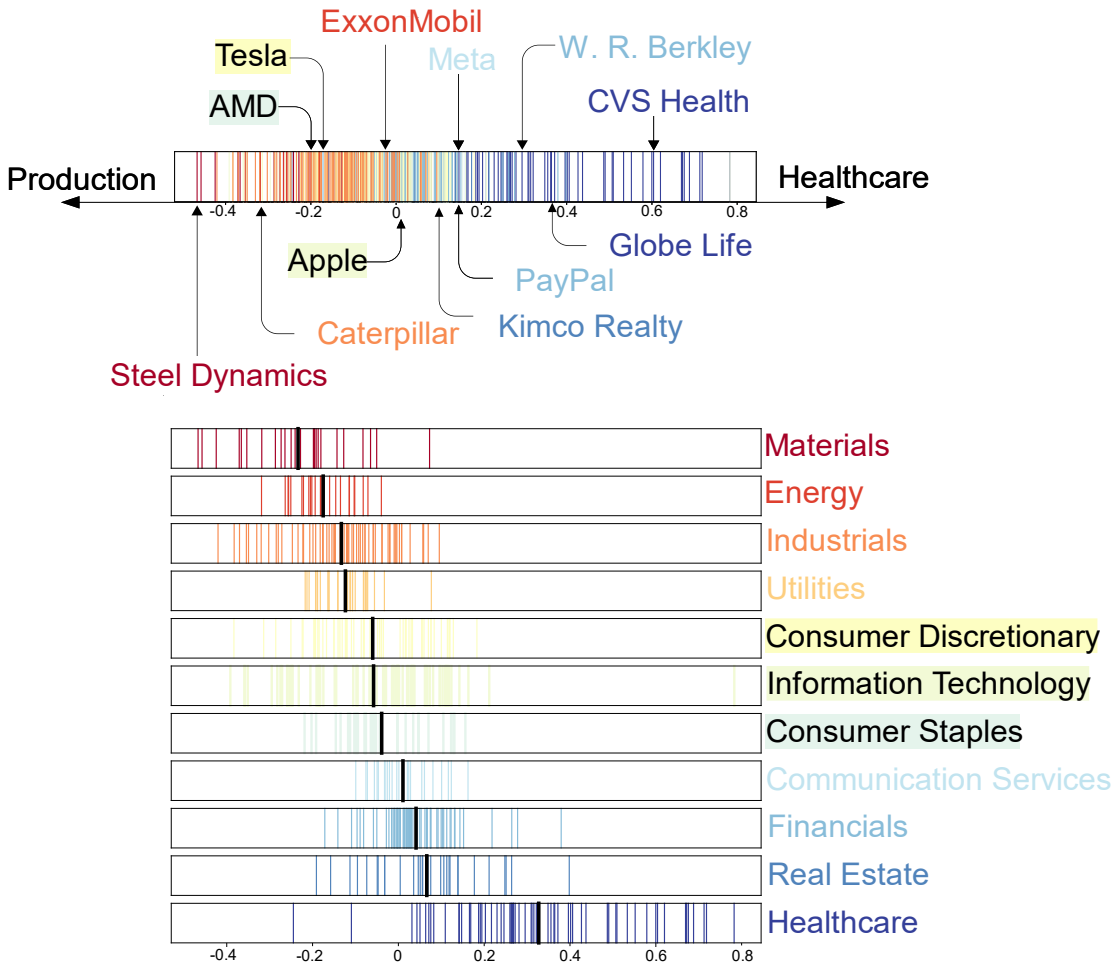

Figure 4 : Firm spectrum

Note : The same projection method applies to firms (using the General Industry Classification System)

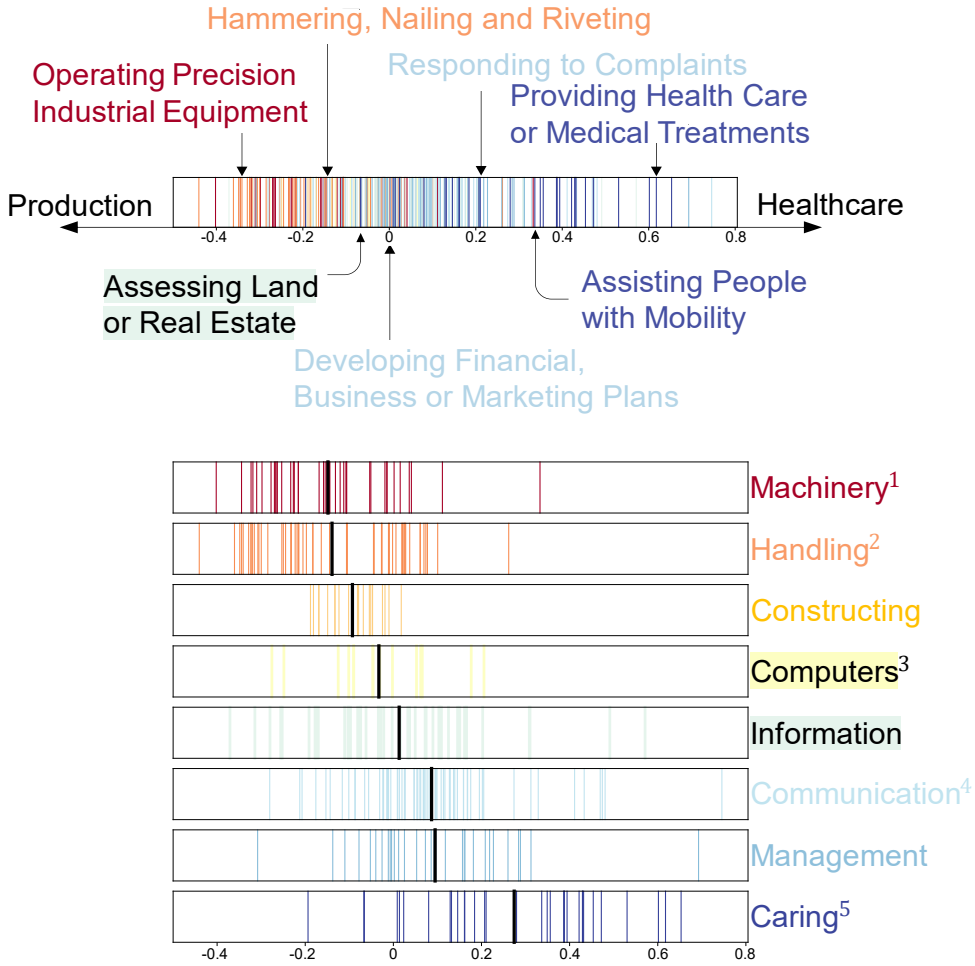

Figure 5 : Skill spectrum

Note : The same projection method applies for skills (using ESCOskill level two hierarchy).

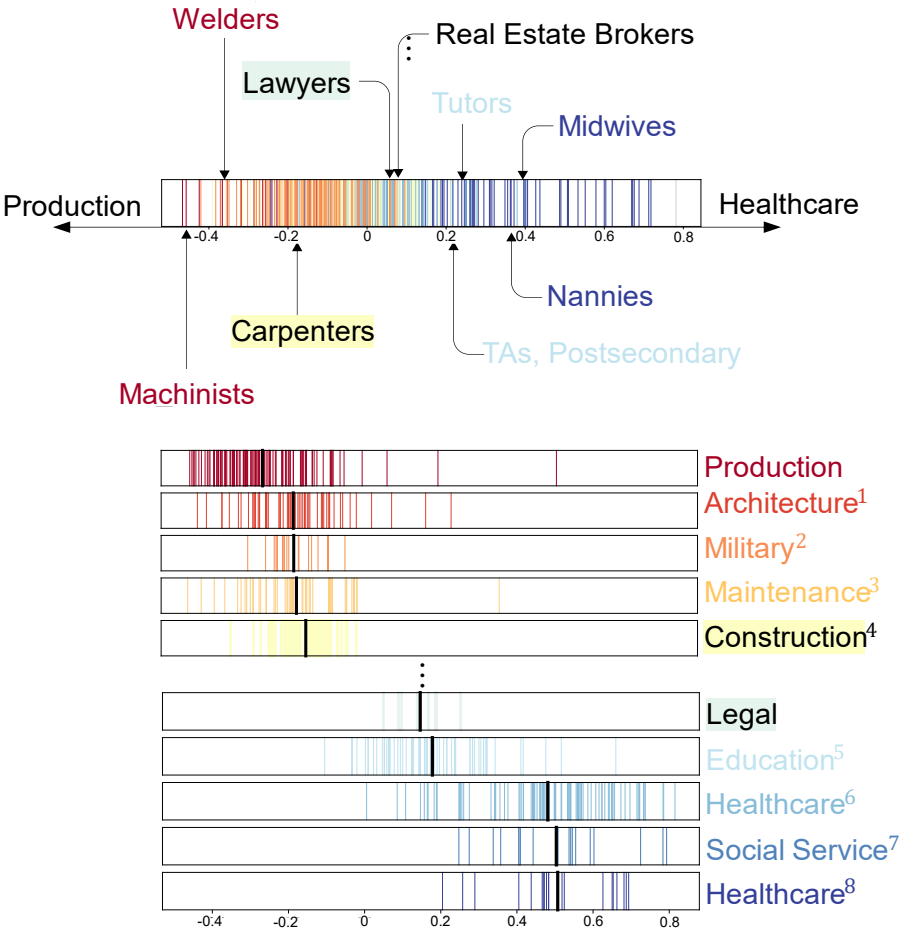

Figure 6 : Occupation spectrum

Note : The same projection method applies for skills (using Standard Occupation Classification)

Projecting entities onto the conceptual axis yields reliable results across all labor market entity categories. Firms associated with materials, energy, and industrial utilities (e.g., Steel Dynamics and Caterpillar) are proximate to the production pole. At the same time, those offering services like healthcare, real estate, and finance (e.g., CVS Health and PayPal) are closer to the healthcare pole (Fig. 4). Similarly, skills and occupations tied to manufacturing tend to align with the left side (e.g., Machinists and Welders), while those connected to services shift to the right (e.g., Midwives and Nannies) (Fig. 5 and Fig. 6). This validation underscores the robustness of Labor Space in mapping concepts across diverse economic categories.

6. Vector calculation for economic analogy

Is it possible to perform vector calculations between different types of economic entities? For example, can we compute V(Firm A) - V(Skill X) + V(Skill Y) to estimate the impact of a new entity's emergence or the absence of an existing entity from one category on an entity in another class?

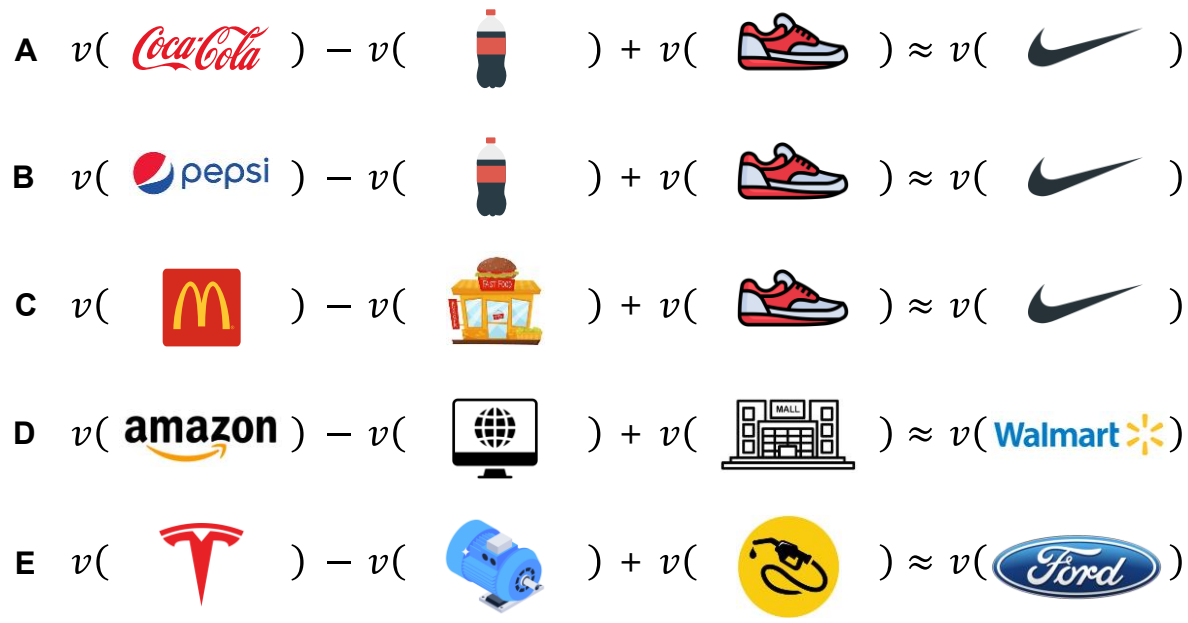

Figure 7 : Vector analogy of firm and industry entities

One of the notable applications of word embedding is vector analogies, as demonstrated by equations like V('king') - V('man') + V('woman') ~ V('queen'). This showcases how word embedding captures semantic relationships, such as the relationship between 'king' and 'queen' mirroring that between 'man' and 'woman'.

In our Labor Space, we employ vector analogies to uncover latent connections between entities across categories. Fig. 7 illustrates relationships between firms and industries. The equation V(Firm A) - V(Industry B) + V(Industry C) ~ V(Firm D) signifies analogical relationships between firms and their corresponding industries. For instance, leading firms in beverages and restaurants are analogously related to Nike in the footwear sector (Fig. 7A, 7B,

and 7C). Another example from Fig. 7D suggests that if Amazon were to divest its web search and IT components but add physical stores, it would approximate Walmart within the S&P 500, as inferred from the vector equation V('Amazon') – V('Web Search Portals, Libraries, Archives, and Other Information Services') + V('Department Stores') ~ V('Walmart'). Similarly, eliminating Tesla of its electrical base but adding gasoline elements aligns it with Ford, as V('Tesla') - V('Other Electrical Equipment and Component Manufacturing') + V('Gasoline Stations') ~ V('Ford'). These vector operations encapsulate various interactions among labor market entities in reality.

Table 2 : 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
Occupation - Occupation ~ Skill	V('Teller') - V('Cashier')	1. Monitoring financial and economic resources 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

Table 2 presents multi-unit vector calculations across diverse labor market entities. For example, the equation 'occupation A - occupation B ~ occupation C' illustrates what occupation A might resemble when lacking the attributes of occupation B. In this case, V(' Data Scientists') - V(' Statisticians') highlights the distinctive occupational traits of 'Data Scientists' compared to 'Statisticians,' revealing roles linked to data science but not statistical analysis, like Data Architects or Database Administrators. The formula 'occupation A - occupation B ~ skill C' emphasizes skills intrinsic to occupation A but not occupation B. The equation V('Teller') - V('Cashier') demonstrates that financial management and budgeting

skills are vital for 'Tellers' but less for 'Cashiers.' Lastly, the equation 'occupation A + industry B + skill C ~ firm D,' perhaps the most intricate analogy, exemplifies how vector analogies can be practically employed for career recommendations to job seekers. In our analysis, the result of V(' Mathematicians') + V(' Other Investment Pools and Funds') + V(' Providing Financial Advice') suggests optimal firms for mathematicians seeking roles in investment funds, leveraging their financial expertise.

7. Estimating the impact of AI

The labor market is transforming significantly due to the widespread adoption of artificial intelligence (AI) in various economic sectors. AI, deriving patterns from data, has raised concerns about potential job displacement [Autor, 2015, Bessen, 2019, Frank et al., 2019]. Recent situations have led to efforts to measure AI's impact on different occupations [Frey et al., 2017, Brynjolfsson and Mitchell, 2017, Acemoglu et al., 2020, Felten et al., 2021] to assist workers in adapting to changing job roles.

So, can our Labor Space provide insights into AI's influence on the labor market, covering various aspects?

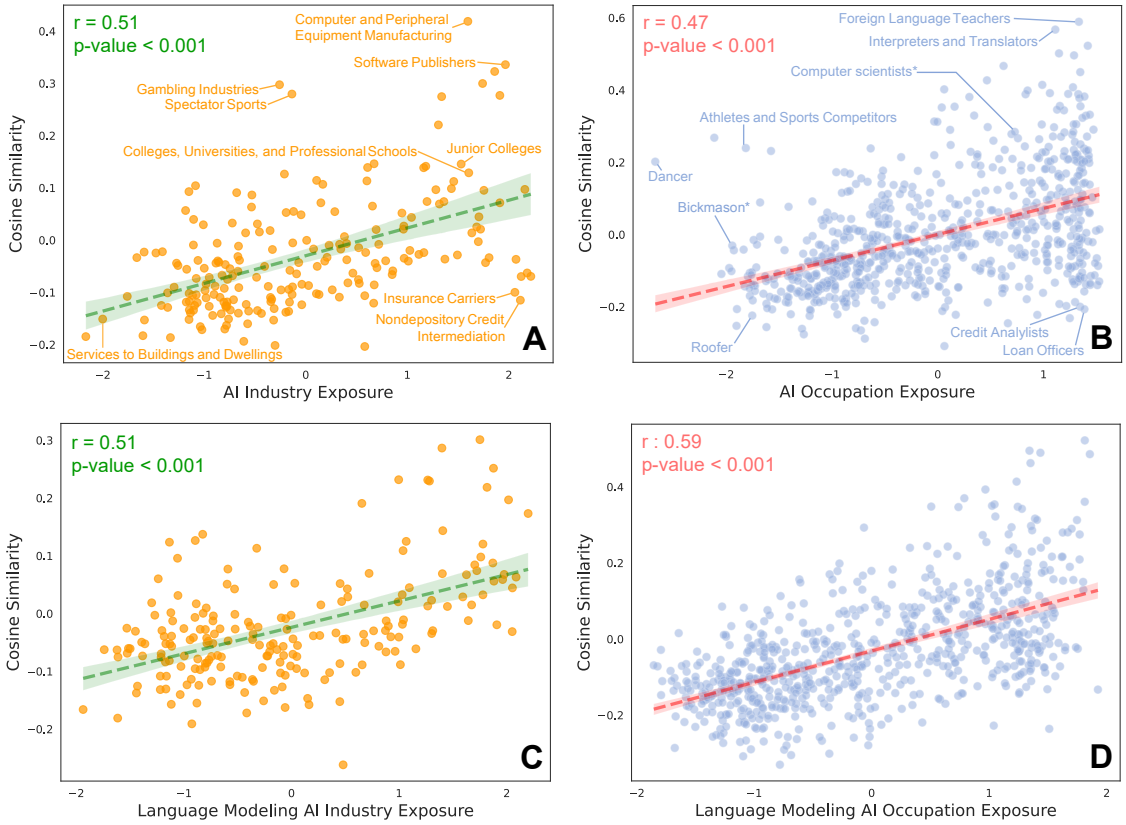

Figure 8 : Correlation between cosine similarity and exposure data

Note : (A) The X-axis is industry exposure data to the top 10 AI applications, and the Y-axis is the cosine similarity of definition embedding of the top 10 AI application list and industry embedding vectors. Definition embedding is calculated using the same method as the Labor Space description embedding. (B) This figure replaces AIIE to AIOE data. (C) We pick up language modeling among the top 10 AI applications. The X-axis is the industry exposure score of language modeling, and the Y-axis is the cosine similarity between language modeling definition embedding and industry vectors. (D) Correlation plot of language modeling exposure in the occupation level.

One notable feature of Labor Space is its scalability. It can easily accommodate new entities as long as sufficient descriptive texts are available. To explore Labor Space's capability to assess emerging technologies like AI, we compared our results with a previous study that quantified AI industry exposure (AIIE) and AI occupation exposure (AIOE) [Felten et al., 2021]. We used the top ten AI application definitions (see Appendix 9.2) from

that study and estimated AI's impact on industries and occupations by calculating cosine similarities between AI application vectors and industry/occupation vectors.

Fig. 8A and Fig. 8B show the correlation between our cosine similarity scores for AI applications and established AIIE and AIOE metrics. The X-axis represents exposure scores from [Felten et al., 2021], indicating vulnerability to specific AI applications, while the Y-axis shows cosine similarity scores between AI application vectors and Labor Space entities. These metrics exhibit a strong correlation, with a Pearson's correlation coefficient of 0.51 (p-value < 0.001), suggesting that cosine similarity measures with new technology vectors can effectively assess AI exposure across different labor market aspects.

Can refining our definitions provide sharper insights within the Labor Space? By focusing on language modeling applications from Felten et al. (2023), we isolated relevant descriptions, derived vectors, and recalibrated our cosine similarity analyses. The correlation between language modeling exposure scores (X-axis) and our computed cosine similarities (Y-axis) notably strengthened for occupational exposure (increasing from 0.47 to 0.59) while maintaining a robust correlation of 0.51 for industrial exposure (Fig. 5C and Fig. 5D). Labor Space tends to highlight higher AI exposure risks for entities whose tasks align with the capabilities of contemporary AI models. For example, entities like 'Software Publisher' and 'Foreign Language Teachers' are perceived as more vulnerable, while financial domains show lower AI exposure, as reflected in language modeling AI exposure visualizations. While determining the absolute accuracy of these estimations is a future challenge, analyzing AI exposure through Labor Space underscores its versatility and provides a virtual platform for policymakers, researchers, and business leaders to conceptualize and simulate potential shifts affecting various labor market entities.

8. Discussion

One of the challenges for researchers is capturing the high-dimensional nature of the labor market. Labor Space comprehensively represents the labor market by encompassing industries, occupations, skills, and firms, offering a holistic view of the ecosystem. Also, it allows for the measurement of conceptual similarity between different labor market entities, providing insights into their relationships and shared characteristics. Its scalability makes it adaptable to new entities, making it relevant for tracking emerging trends and technologies in the labor market. Moreover, the Labor Space enables the analysis of interconnectedness between different labor market components, aiding in understanding how changes in one area can affect others. Lastly, it can assess the impact of emerging technologies like AI on the labor market, helping to identify which industries and occupations are most affected.

However, there are some potential shortcomings to consider. Firstly, the quality and availability of descriptive texts for entities are crucial, and if representative descriptions are lacking or biased, it can affect the accuracy and comprehensiveness of the Labor Space. Secondly, subjectivity can be introduced through the choice of parameters, such as the selection of descriptive texts and the configuration of the embedding model. Additionally, the Labor Space primarily relies on textual descriptions for entities. It may not fully capture other aspects of the labor market, such as quantitative data or dynamic changes over time. Lastly, it may not fully capture the dynamic nature of the labor market, as it provides a static snapshot based on available data. Despite these potential limitations, the Labor Space remains a powerful tool for gaining insights into the labor market's complexities and relationships among its various components.

9. Appendix

9.1. Tradability score

To make a tradable-nontradable dimension, we set the industry centroid with reference to the tradability score [Jensen et al., 2005]. Table 3 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 3 : 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	0	100

9.2. Top 10 applications

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 4 gives the top 10 applications list and brief definitions.

Table 4 : Top 10 applications

AI application	Definition
Abstract strategy games	The ability to play abstract games involving sometimes complex strategy and reasoning ability, such as chess, go, or checkers at a high level.
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.
Image generation	The creation of complex images
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.

References

1. Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2020). *AI and jobs: Evidence from online vacancies* (No. w28257). National Bureau of Economic Research.
2. Akerman, A., Helpman, E., Itskhoki, O., Muendler, M. A., & Redding, S. (2013). Sources of wage inequality. *American Economic Review*, *103*(3), 214-219.
3. Alabdulkareem, A., Frank, M. R., Sun, L., AlShebli, B., Hidalgo, C., & Rahwan, I. (2018). Unpacking the polarization of workplace skills. *Science advances*, *4*(7), eaao6030.
4. An, J., Kwak, H., & Ahn, Y. Y. (2018). Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. *arXiv preprint arXiv:1806.05521*.
5. Anderson, K. A. (2017). Skill networks and measures of complex human capital. *Proceedings of the National Academy of Sciences*, *114*(48), 12720-12724.
6. Autor, D., Chin, C., Salomons, A. M., & Seegmiller, B. (2022). *New Frontiers: The Origins and Content of New Work, 1940–2018* (No. w30389). National Bureau of Economic Research.
7. Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of economic perspectives*, *29*(3), 3-30.
8. Bana, S., Brynjolfsson, E., Rock, D., & Steffen, S. (2020). job2vec: Learning a representation of jobs.
9. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.
10. Bessen, J. (2019). Automation and jobs: When technology boosts employment. *Economic Policy*, *34*(100), 589-626.

11. Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530-1534.
12. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.
13. Chau, H., Bana, S. H., Bouvier, B., & Frank, M. R. (2023). Connecting higher education to workplace activities and earnings. *Plos one*, *18*(3), e0282323.
14. Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.
15. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
16. Dong, Y., Chawla, N. V., & Swami, A. (2017, August). metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 135-144).
17. Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.
18. Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., ... & Rahwan, I. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, *116*(14), 6531-6539.

19. Felten, E., Raj, M., & Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, *42*(12), 2195-2217.
20. Felten, E. W., Raj, M., & Seamans, R. (2023). Occupational heterogeneity in exposure to generative ai. Available at SSRN 4414065.
21. Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation?. *Technological forecasting and social change*, *114*, 254-280.
22. Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, *6*(7), 975-987.
23. Jensen, J. B., Kletzer, L. G., Bernstein, J., & Feenstra, R. C. (2005, January). Tradable services: Understanding the scope and impact of services offshoring [with comments and discussion]. In *Brookings trade forum* (pp. 75-133). Brookings Institution Press.
24. Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905-949.
25. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.
26. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.
27. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013b). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

28. Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. *Advances in neural information processing systems, 26*
29. Neffke, F., & Henning, M. S. (2008). Revealed relatedness: mapping industry space, Papers in Evolutionary Economic Geography (PEEG) 0819. *Utrecht University, Section of Economic Geography*.
30. Neffke, F., & Henning, M. (2013). Skill relatedness and firm diversification. *Strategic Management Journal*, *34*(3), 297-316.
31. OpenAI (2023). Gpt-4 technical report. Technical report, OpenAI.
32. O*NET (2022). O*net 27.3 database.
33. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.