Exposure, Embeddings, & Employment: Quantifying AI's Impact on Occupational Data

Richard Ren

Final Project for Graduate Econometrics I, Statistics Department at the Wharton School of Business in the University of Pennsylvania

1 Introduction

The advent of artificial intelligence (AI) and its rapid advancement presents both opportunities and challenges for the labor market. As AI technologies evolve, they increasingly intersect with various occupational fields, potentially altering employment patterns and wage structures. This intersection has given rise to a need for accessible and reliable measures of occupational exposure to AI. Such measures would not only provide insights into the impact of AI on the labor market but also facilitate the integration of this knowledge into econometric models on wage, growth, and employment – which could inform vital public policy decisions around how to address automation and minimize its harms.

I propose to study the relationship between technology exposure and labor market outcomes using text analysis of patents and occupations, inspired by the methodology used in Webb (2019).

This project's novelty lies in its method of calculating of occupation exposure scores for AI leveraging the latest advancements in natural language processing, as well as its analysis on patent data published in 2020-23. As a field, natural language processing has moved toward being more conceptual to statistical. Instead of relying on hand-written rules, which was the paradigm that dominated from the 1950s to the 1990s, modern-day NLP relies on statistical systems that leverage computational power. Modern day deep learning systems, including large language models such as ChatGPT and BERT, utilize statistical representations of language to generate or process text. Similarly, while Webb (2019) uses text-verb pairs, I aim to (somewhat poetically) use developments in AI and machine learning to study the effect that AI may have on automation and productivity.

My contributions are therefore threefold:

- (a) providing a methodology and codebase for translating patent and occupational text into a quantitative occupation exposure score to artificial intelligence through large language model embeddings,
- (b) conducting a preliminary analysis and understanding how these coefficients have empirically changed over time, and
- (c) discussing how this analysis can be used to determine employment and wages effects based on occupation.

I make publically available all of my code in this Github repository.

2 Methodology

Obtaining and Cleaning Text Data for AI-Related Patents and Occupational Task Descriptions

We begin by collecting patent and occupational text data of interest.

Patent data: Following the datasets from Webb (2019), I begin by collecting raw patent data from Google Patents Public Data provided by IFI CLAIMS Patent Services, which included comprehensive details like titles, abstracts, and publication dates. I queried the patent data for U.S. patents with terms like "machine learning", "neural network", and "deep learning" in the abstract or title. Afterwards, I remove any duplicates from the query, seperate the data by year to allow for observation of trends over time, and combine patent titles and abstracts in preparation for textual analysis. I have patent data for 1978 as well as 1987-2023 (up to July). In total, we have 89621 AI-related patents to process across 38 years.

Occupational data: For job descriptions, I use the O*NET database (describing a total of 964 occupations) and take a subset of fifteen significant occupations of interest. For each occupation, there are many possible descriptions (of tasks, skills needed, etc.) – I just use occupation task descriptions for textual analysis, which consist of approximately 15-20 bullet points I merge together into a single textual data point.

Explanation of Large Language Model Embeddings

Text embeddings transform text into numerical vectors that capture semantic meaning in interesting ways. Embeddings are used to translate some input text T (which can vary in size – e.g. on a token, word, sentence, paragraph, or document level) into a high-dimensional vector space H that encodes significant semantic data. The ability to embed words into meaningful vectors has been key in the implementation of natural language processing into large language model (LLM) deep learning systems such as transformers or residual neural networks.

LLMs are trained on extensive corpora to generate embeddings (taken at intermediate model layers) as a byproduct of their task. For example, autoregressive LLMs tend to maps an input text T into an embedding space H before performing operations aimed at decoding the next token T'. One can also have LLMs predict the context (e.g. surrounding words) given a text, known as the "skip-gram model" as seen in models like word2vec. The objective function for these models can be formally represented as:

- For Autoregressive Language Models: Maximize $\sum_{t=1}^{T} \log p(w_t | w_{t-c}, ..., w_{t+c})$
- For Skip-Gram Model: Maximize $\sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$

where w_t represents the target word, c is the context size, and T is the total number of training words. In these training objectives, the embedding spaces become trained to recognize and classify various semantically significant features and properties.

Previous research, notably with word2vec, has shown interesting linear relationships between embedding vectors, such as that $v_{\text{"king"}} - v_{\text{"man"}} + v_{\text{"woman"}} \approx v_{\text{"queen"}}$, suggesting that embeddings encode information about the relationships and analogies between words (Mikolov et al., 2013). More recently, embeddings have been used for retrieval-augmented generation techniques (document retriveval and reference in chatbots) as well as text sentiment analysis and classification (Lewis et al., 2021; Thongtan and Phienthrakul, 2019). With the recent rise in embeddings used to augment AI chatbots, there have been more deticated embedding-focused LLMs being released.

Therefore, this paper expands upon Webb (2019)'s methodology – where instead of simply using plain noun-verb pairs, we utilize LLM embeddings, a technique which represents the state-of-the-art methodology for knowledge and document retrieval.

Creating Embeddings of Patent and Occupational Text Data

This section elaborates on the methodologies involved in extracting and utilizing embeddings from language models to analyze the semantic similarity between patent data and occupational descriptions.

Many embeddings-focused models can be chosen, including both open-source (e.g. WhereIsAI/UAE-Large-V1, sentence-transformers/multi-qa-mpnet-base-dot-v1) as well as closed-source (Cohere Embed v3.0, OpenAI API) models, to obtain embeddings. There exists an embeddings model leaderboard in the HuggingFace website, a popular natural language processing-focused repository.

To quantify the semantic similarity, we employed OpenAI's GPT embedding API (most specifically, their text-embedding-ada-002 model), generating embeddings for both patent text and O*NET occupation descriptions. Each piece of text will be converted into a vector of numbers, representing its position in the semantic space. OpenAI's embedding model typically produces fixed-length embeddings for any input text up to 8192 tokens – meaning for a phrase or a long paragraph up to a certain length, the output embedding will have the same number of dimensions (e.g. 1536-dimensional vector). This feature ensures that every piece of text, regardless of its length, is represented in a uniform manner, facilitating simple semantic comparisons. Furthermore, the embedding vectors are already normalized.

With 89621 total patents to embed, calling the API one-by-one would result in a very long wait time. To improve the efficiency of the embedding process, I implemented parallel processing using a thread pool execution strategy and the backoff library – this approach calls the embeddings API on as many patents as simultaneously possible, staying right under the rate limit for OpenAI's API call by dynamically "backing off" if too many calls are made. This creates a smooth and fast data processing experience; these engineering details can be further explored in our publically-available code at my Github page for this project. The total cost of embedding all patents was less than ten dollars.

Afterward, by grouping embeddings, taking cosine similarities, or comparing the k-nearest neighbors, one can understand the conceptual and semantic similarity between two phrases.

Cosine Similarity Calculation

Our primary metric for analysis was cosine similarity, which we used to measure the closeness between patent and occupational embeddings. A higher cosine similarity indicated a stronger thematic or semantic relationship between the patent content and the occupational description. By calculating these similarities, we could comparatively analyze the exposure of different occupations to AI automation over time.

Suppose we have a set of patent texts $P = \{p_1, p_2, ..., p_n\}$ and a set of occupational texts $O = \{o_1, o_2, ..., o_m\}$. Each text is converted into an embedding vector:

- $v_{p_i} = \text{Embedding}(p_i)$ for patents
- $v_{o_j} = \text{Embedding}(o_j)$ for occupations

The semantic similarity between a patent text and an occupational text can be computed using cosine similarity, from a range of -1 to 1:

$$\text{Similarity}(p_i, o_j) = \frac{v_{p_i} \cdot v_{o_j}}{\|v_{p_i}\| \|v_{o_j}\|}$$

By calculating this similarity across all pairs of patent and occupational texts, we can derive a comprehensive view of how different occupations are exposed to the technologies described in patent documents.

3 Results

Increase in AI Patents Over Time

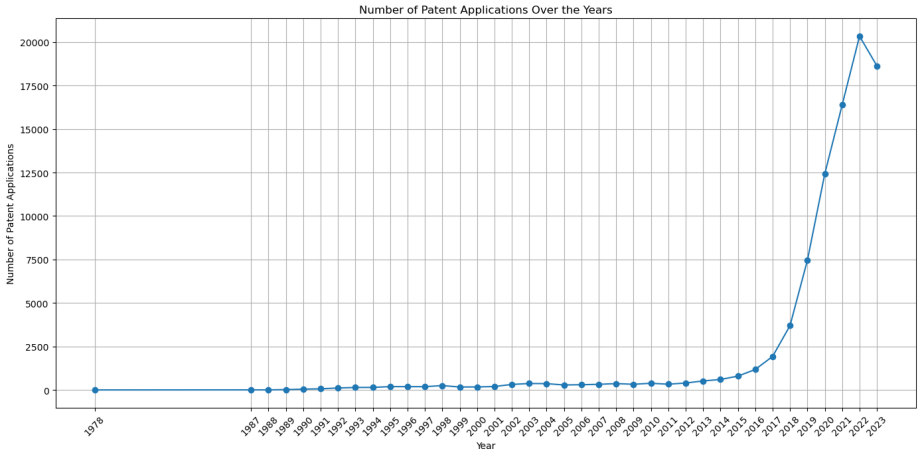

Figure 1: Yearly U.S. patent applications over data available in Google Patents Public Data with keywords "machine learning", "deep learning", "artificial intelligence", or "neural networks" in the title or abstracts.

In Fig 1, one can see that the number of AI-related patents have significantly increased, especially since the late 2010s. While it may look like 2023 has seen a decrease in the number of AI-related patent applications, it should be noted that the patent data for 2023 ends at July and therefore it is not an apples-to-apples comparison to other years.

Visualizing Cosine Similarities for All Patents

We plot various distributions of cosine similarities across ten arbitrarily selected occupations and all patents filed in a given year for five arbitrarily selected years (2022, 2020, 2018, 2017, 2016) in Fig 2. In short, this tells us the distribution of patent similarities for each occupation across different years. One can also obtain the mean and standard deviation is for each combination of occupation and yearly patent data using our publicly available code. Notably, the cosine similarities seem to be quite close and within a limited range – indicating that all patent and occupational data tends to be generally similar in the context of all human language.

Generally, for a given occupation, their mean cosine similarity with patents tends to stay quite similar from year to year (within the 0.6-0.8 range). However, in between occupation, mean cosine similarities tend to differ.

Setting Thresholds to Obtain Automation Exposure Metric

We obtain our exposure metric by setting specific thresholds to categorize patents as having "medium exposure" or "high exposure" to AI automation. These thresholds were arbitrarily set based on the distribution of cosine similarities. For each number in Fig 3 and 4, we show the proportion of patents in the entire year above the cosine similarity threshold (as a percentage) when compared with the given occupation.

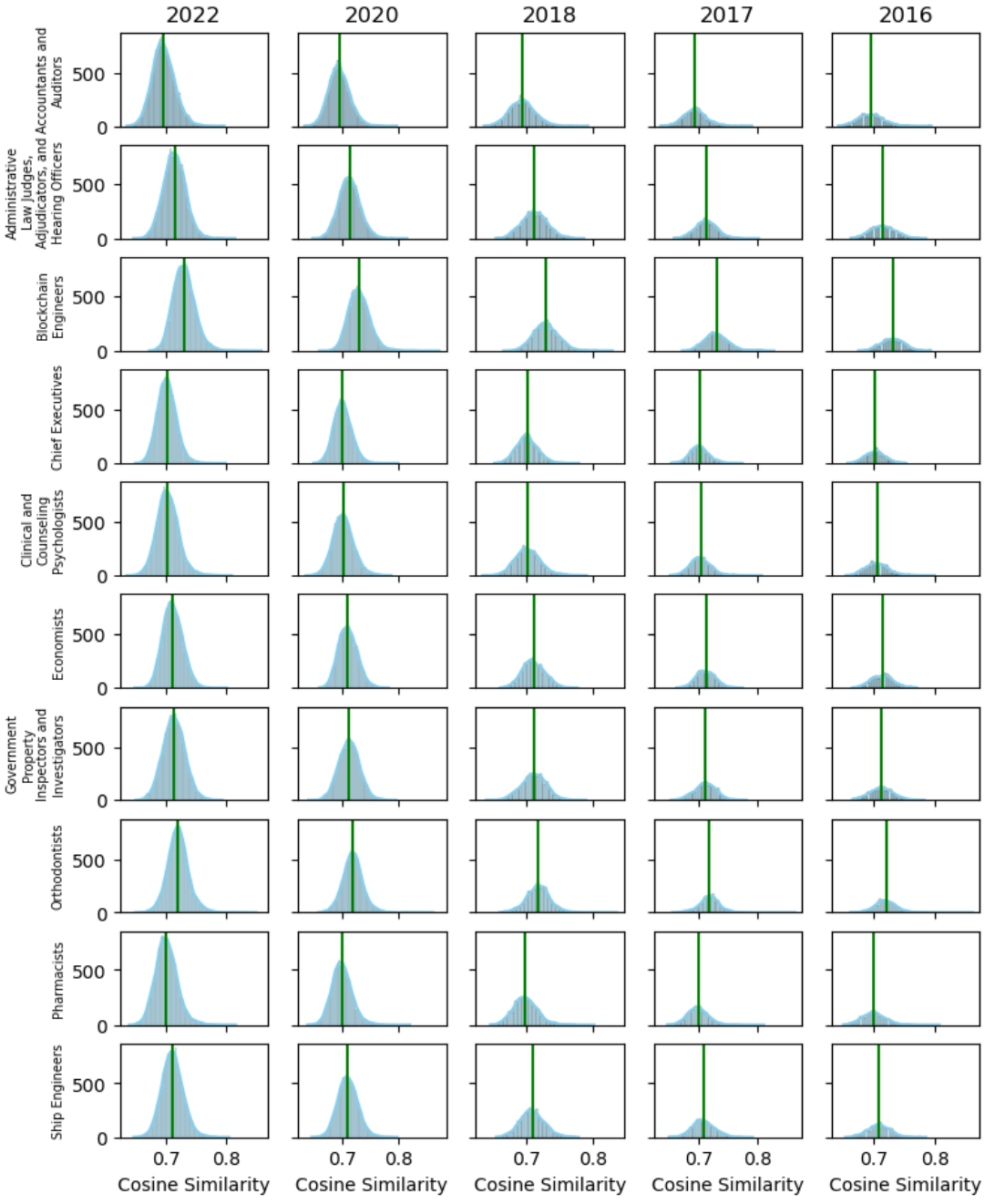

Figure 2: Distribution plot of cosine similarities with patent data from five different years, across ten occupations. Mean indicated with green line.

Occupation	Accountants and Auditors	Administrative Law Judges, Adjudicators, and Hearing Officers	Blockchain Engineers	Chief Executives	Clinical and Counseling Psychologists	Computer Network Architects	Computer and Information Systems Managers	Economists	Government Property Inspectors and Investigators	Orthodontists	Pharmacists	Ship Engineers	Tax Preparers	Travel Agents	Zoologists and Wildlife Biologists
Year															
1978	0.000000	0.000000	0.000000	0.000000	100.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1987	0.000000	0.000000	0.000000	0.000000	0.000000	20.000000	20.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1988	0.000000	0.000000	0.000000	0.000000	0.000000	50.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1989	10.526316	21.052632	31.578947	5.263158	5.263158	78.947368	42.105263	10.526316	15.789474	10.526316	5.263158	21.052632	5.263158	15.789474	15.789474
1990	0.000000	2.127660	12.765957	0.000000	2.127660	68.085106	21.276596	2.127660	2.127660	2.127660	0.000000	0.000000	0.000000	2.127660	8.510638
1991	1.515152	4.545455	21.212121	0.000000	0.000000	69.696970	31.818182	0.000000	4.545455	9.090909	1.515152	6.060606	1.515152	6.060606	6.060606
1992	0.000000	5.309735	12.389381	0.000000	4.424779	76.991150	23.893805	2.654867	7.079646	12.389381	0.884956	2.654867	0.000000	3.539823	6.194690
1993	1.408451	2.816901	25.352113	0.704225	2.816901	76.056338	31.690141	4.929577	6.338028	6.338028	0.704225	3.521127	1.408451	6.338028	8.450704
1994	0.000000	3.401361	12.244898	0.000000	2.721088	70.068027	27.891156	4.081633	4.081633	14.965986	2.040816	10.204082	0.000000	6.802721	6.122449
1995	0.000000	6.770833	18.229167	2.083333	5.729167	76.041667	30.729167	5.208333	5.729167	10.937500	0.000000	9.375000	0.000000	7.291667	11.458333
1996	1.562500	5.729167	19.270833	1.562500	4.166667	76.041667	30.208333	2.083333	10.416667	13.020833	1.562500	9.895833	0.520833	9.895833	11.458333
1997	3.191489	9.574468	19.680851	1.595745	5.319149	73.404255	25.000000	4.787234	11.702128	17.553191	3.191489	10.106383	1.063830	6.914894	10.638298
1998	2.766798	9.881423	22.924901	2.766798	7.114625	69.169960	34.782609	5.533597	15.019763	22.924901	4.743083	13.043478	1.185771	10.276680	16.600791
1999	3.550296	11.242604	27.810651	2.366864	5.917160	66.863905	35.502959	2.958580	13.609467	21.893491	6.508876	9.467456	1.775148	13.609467	15.384615
2000	3.448276	24.137931	37.356322	8.620690	18.390805	73.563218	41.379310	18.390805	13.218391	25.862069	5.747126	15.517241	2.873563	18.965517	25.862069
2001	4.545455	9.595960	28.787879	3.535354	6.565657	69.696970	42.929293	9.090909	9.595960	20.707071	4.545455	20.202020	3.030303	16.666667	17.676768
2002	4.402516	12.893082	29.559748	5.345912	8.490566	70.440252	42.767296	11.320755	17.610063	19.182390	3.773585	14.150943	2.201258	17.610063	22.327044
2003	4.054054	15.405405	38.648649	6.216216	10.540541	69.729730	40.000000	14.594595	14.054054	23.513514	5.405405	15.405405	1.081081	18.918919	22.972973
2004	1.928375	9.917355	33.057851	3.581267	7.713499	66.391185	35.261708	9.641873	9.917355	21.212121	2.479339	13.498623	1.377410	17.079890	20.385675
2005	3.832753	12.543554	32.404181	5.226481	8.013937	64.459930	43.902439	13.937282	16.027875	25.087108	5.574913	12.891986	2.787456	17.421603	22.996516
2006	3.311258	15.562914	40.066225	4.966887	10.927152	67.549669	47.682119	18.543046	15.894040	24.503311	5.960265	9.933775	0.993377	14.900662	27.814570
2007	4.307692	13.230769	35.384615	4.615385	6.461538	68.923077	46.461538	13.846154	20.000000	25.846154	3.076923	17.538462	2.769231	16.000000	23.384615
2008	3.835616	14.794521	43.835616	8.767123	10.136986	73.424658	50.958904	15.616438	15.890411	24.383562	8.219178	17.808219	2.465753	24.383562	34.246575
2009	4.878049	21.036585	44.207317	5.487805	8.841463	70.731707	55.487805	16.768293	17.378049	23.475610	4.268293	11.890244	4.573171	25.609756	31.707317
2010	4.699739	17.493473	35.248042	5.221932	6.527415	63.185379	44.647520	14.099217	18.015666	20.626632	6.005222	11.488251	3.133159	21.148825	24.804178
2011	5.438066	21.752266	41.389728	3.927492	10.574018	67.975831	52.567976	16.918429	21.450151	26.283988	6.042296	14.803625	3.625378	28.700906	29.607251
2012	6.516291	25.563910	47.869674	8.521303	13.784461	65.914787	57.644110	19.298246	21.804511	27.318296	8.020050	12.781955	7.017544	27.819549	31.328321
2013	4.633205	17.374517	45.366795	5.984556	10.810811	68.725869	48.455598	16.795367	16.216216	23.938224	5.212355	12.355212	4.440154	26.447876	30.115830
2014	6.301824	22.222222	49.087894	7.296849	9.784411	70.812604	54.394693	19.237148	17.910448	28.358209	7.131012	14.262023	4.643449	26.533997	29.850746
2015	3.526448	18.513854	46.347607	4.785894	9.193955	65.743073	46.473552	18.010076	14.609572	24.937028	3.652393	11.335013	1.889169	23.047859	23.803526
2016	4.166667	22.193878	51.615646	4.336735	10.119048	71.513605	52.380952	16.581633	17.517007	26.530612	5.867347	11.139456	2.295918	23.299320	26.700680
2017	4.097510	16.804979	48.443983	4.564315	6.742739	71.058091	51.348548	13.692946	16.182573	22.717842	4.097510	13.174274	1.815353	21.836100	19.087137
2018	3.544372	17.126623	47.916667	4.139610	7.656926	67.153680	50.216450	13.988095	16.017316	22.808442	4.707792	12.662338	1.758658	21.996753	21.374459
2019	3.583412	16.105221	47.201718	3.918937	7.247349	66.890350	50.476446	11.407865	15.004697	23.567306	5.650248	12.978124	1.677627	22.426520	19.460475
2020	3.639291	16.674718	47.190016	3.880837	6.932367	65.829308	49.758454	11.908213	16.143317	23.945250	5.161031	12.624799	1.940419	22.512077	19.170692
2021	3.875732	17.352295	47.802734	3.961182	6.958008	65.045166	51.330566	11.785889	17.053223	25.701904	5.822754	13.073730	2.166748	22.894287	18.872070
2022	4.071197	17.931950	47.192448	3.668011	6.416560	63.727997	50.693284	12.518438	17.086243	25.292556	5.767529	13.054381	2.192939	22.140820	18.285967
2023	3.993130	18.221340	49.028553	3.424216	7.047016	64.958137	51.137827	12.924002	17.137183	25.579648	6.048733	13.074281	2.216617	22.627737	18.650708

Figure 3: Medium exposure coefficient table over patent publishing years and occupation. Defined as proportion of patents with cosine similarities above 0.73 for a given occupation and expressed as percentage.

Occupation	Accountants and Auditors	Administrative Law Judges, Adjudicators, and Hearing Officers	Blockchain Engineers	Chief Executives	Clinical and Counseling Psychologists	Computer Network Architects	Computer and Information Systems Managers	Economists	Government Property Inspectors and Investigators	Orthodontists	Pharmacists	Ship Engineers	Tax Preparers	Travel Agents	Zoologists and Wildlife Biologists
Year															
1978	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1987	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1988	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1989	0.000000	0.000000	0.000000	0.000000	0.000000	15.789474	15.789474	0.000000	0.000000	5.263158	0.000000	10.526316	0.000000	0.000000	0.000000
1990	0.000000	0.000000	2.127660	0.000000	0.000000	14.893617	2.127660	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1991	0.000000	0.000000	0.000000	0.000000	0.000000	4.545455	1.515152	0.000000	0.000000	0.000000	0.000000	3.030303	0.000000	1.515152	0.000000
1992	0.000000	0.000000	0.884956	0.000000	0.000000	13.274336	1.769912	0.000000	0.000000	0.884956	0.000000	0.000000	0.000000	0.000000	0.000000
1993	0.000000	0.000000	0.000000	0.000000	0.000000	15.492958	2.112676	0.000000	0.704225	0.000000	0.000000	1.408451	0.000000	0.704225	0.704225
1994	0.000000	0.000000	0.000000	0.000000	0.000000	8.163265	1.360544	0.000000	0.000000	0.000000	0.000000	2.040816	0.000000	0.680272	0.000000
1995	0.000000	0.000000	0.520833	0.000000	0.520833	13.541667	6.250000	0.520833	0.000000	1.562500	0.000000	1.041667	0.000000	1.041667	0.000000
1996	0.000000	0.000000	0.000000	0.000000	0.000000	10.416667	3.125000	0.520833	0.000000	0.520833	0.000000	0.000000	0.000000	0.000000	0.000000
1997	0.000000	0.000000	0.531915	0.000000	1.063830	13.829787	3.723404	0.000000	0.531915	3.723404	0.000000	1.063830	0.000000	0.531915	0.000000
1998	0.395257	0.395257	0.395257	0.395257	1.185771	14.229249	4.743083	0.395257	0.790514	2.371542	0.395257	1.185771	0.000000	0.790514	0.000000
1999	0.000000	0.591716	0.000000	0.000000	1.183432	8.875740	4.733728	0.000000	1.183432	2.958580	0.000000	1.183432	0.000000	0.000000	0.591716
2000	0.574713	1.149425	1.724138	0.000000	2.298851	12.643678	14.367816	0.000000	1.149425	2.873563	0.000000	1.149425	0.574713	0.000000	0.574713
2001	0.505051	1.010101	1.515152	0.000000	2.020202	15.151515	4.545455	0.000000	0.505051	2.525253	1.010101	2.525253	0.000000	1.515152	0.000000
2002	0.314465	0.314465	1.886792	0.000000	0.000000	9.748428	9.748428	0.314465	0.628931	0.314465	0.314465	2.830189	0.628931	1.572327	0.628931
2003	0.000000	0.810811	0.810811	0.000000	1.081081	12.162162	9.189189	0.270270	1.081081	1.081081	1.081081	1.621622	0.000000	1.621622	1.351351
2004	0.000000	1.377410	1.928375	0.275482	1.101928	11.294766	4.958678	0.000000	0.550964	0.826446	0.550964	0.550964	0.000000	0.000000	0.275482
2005	0.348432	1.045296	1.742160	0.000000	1.045296	11.498258	6.271777	0.348432	1.045296	2.090592	1.045296	0.348432	0.000000	0.348432	2.090592
2006	0.331126	0.662252	3.311258	0.000000	0.331126	10.927152	9.602649	0.331126	0.000000	1.324503	0.331126	1.655629	0.331126	1.324503	1.655629
2007	0.307692	1.846154	2.461538	0.000000	0.000000	15.076923	8.000000	0.307692	0.615385	1.230769	0.307692	1.538462	0.000000	0.000000	0.923077
2008	0.273973	1.095890	7.397260	0.273973	0.000000	17.534247	13.150685	0.273973	1.095890	1.095890	0.000000	0.821918	0.273973	2.465753	0.821918
2009	0.304878	0.609756	1.524390	0.000000	0.304878	16.768293	13.414634	0.609756	0.609756	1.524390	0.000000	0.914634	0.304878	3.048780	0.000000
2010	0.783290	1.305483	3.655352	0.000000	0.000000	12.793734	9.399478	0.783290	1.044386	2.088773	0.261097	0.000000	0.522193	2.088773	0.783290
2011	0.302115	1.510574	5.135952	0.000000	0.302115	14.199396	10.574018	1.510574	0.604230	2.114804	0.000000	1.208459	0.000000	0.604230	1.208459
2012	0.501253	2.005013	4.761905	0.000000	1.754386	14.035088	14.536341	0.250627	1.754386	3.508772	1.002506	0.250627	0.250627	3.508772	0.000000
2013	0.000000	2.895753	4.054054	0.000000	0.579151	13.513514	8.494208	1.351351	1.158301	1.158301	0.579151	0.772201	0.000000	2.316602	0.386100
2014	0.497512	1.658375	5.140962	0.000000	1.492537	15.422886	15.588723	0.663350	0.995025	2.155887	0.497512	2.155887	0.165837	2.653400	0.995025
2015	0.125945	0.755668	3.904282	0.000000	0.503778	13.602015	8.816121	0.881612	1.007557	1.511335	0.125945	1.133501	0.125945	0.881612	0.503778
2016	0.085034	1.870748	5.187075	0.000000	0.765306	14.200680	11.479592	1.020408	1.275510	2.891156	0.510204	0.680272	0.000000	1.445578	0.425170
2017	0.363071	1.089212	7.002075	0.051867	0.726141	14.885892	11.047718	0.363071	0.518672	1.659751	0.518672	0.570539	0.000000	1.970954	0.363071
2018	0.216450	1.028139	5.790043	0.054113	0.622294	13.474026	9.036797	0.351732	0.514069	1.704545	0.487013	0.730519	0.081169	1.650433	0.351732
2019	0.174473	1.046839	6.187089	0.040263	0.617367	12.897598	9.139713	0.308683	0.697893	2.254731	0.577104	0.577104	0.040263	1.704469	0.228157
2020	0.161031	0.885668	6.280193	0.056361	0.442834	12.012882	8.993559	0.305958	0.748792	2.012882	0.418680	0.724638	0.072464	1.980676	0.402576
2021	0.195312	1.013184	5.572510	0.061035	0.469971	11.749268	8.734131	0.262451	0.897217	2.270508	0.482178	0.860596	0.048828	1.843262	0.421143
2022	0.152424	0.875209	5.620022	0.098338	0.486774	10.782771	8.540663	0.255679	0.830957	2.271610	0.476940	0.722785	0.059003	1.740584	0.467106
2023	0.161013	0.933877	5.914556	0.037570	0.585015	10.852297	8.447832	0.187849	0.810434	2.323959	0.477673	0.611851	0.053671	1.690640	0.493774

Figure 4: High exposure coefficient table over patent publishing years and occupation. Defined as proportion of patents with cosine similarities above 0.76 for a given occupation and expressed as percentage.

Comparison across occupations

Interestingly, the occupation variances in Fig 3 and 4 seem quite unintuitive. Some can be explained away by the flaws of using patents as a proxy for innovation – for example, blockchain engineers, computer network architects, and computer and information systems managers having very high proportions of high and medium exposure indicates that there exist a lot of patents in their field, but not necessarily that they're exposed more to automation.

However, even among certain occupations, many occupation variances are difficult to interpret. Tax preparers don't seem to be indicated to be very automatable, despite software products existing currently to deal with tax preparation, such as TurboTax. This could potentially be because the proportion of patents dealing with tax is small, but those few patents are very effective. Meanwhile, unintuitively, zoologists and ship engineers are given an arbitrarily large percentage of AI automation exposure.

In the Conclusions & Future Work section, we further discuss how this could be likely due to flaws in (a) cosine similarities of embeddings as a potentially unreliable metric and (b) the fallacy of relying on patents to measure innovation, especially in the software domain.

Comparison across time

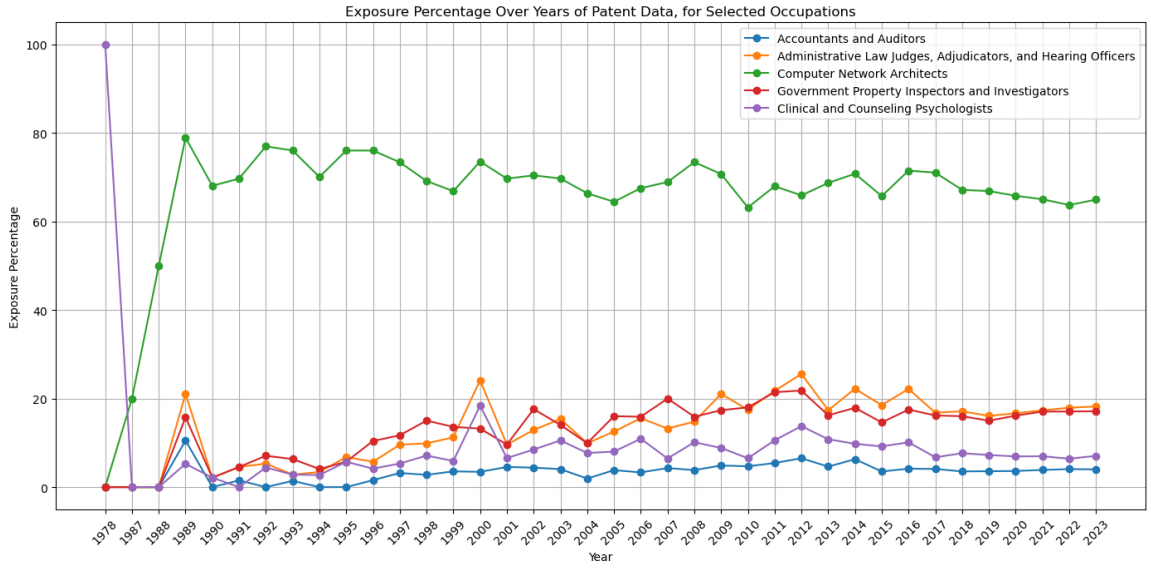

Figure 5: Medium exposure (proportion of patents with cosine similarities above 0.76 for a given occupation), plotted for five occupations over continuous range of patent data.

We graph the medium exposure percentage of five randomly selected occupations over time in Fig 5, and find that the temporal differences are generally interpretable. For example, it seems to be that generally, exposure percentage of selected occupations increases over time. There are some outliers (e.g. for 1978, clearly, the sole AI-related patent had to do with clinical practice), but outside of that, the trend is clear: the proportion of AI-filed patents registered seems to be increasing in the occupations selected. However, the computer network architecture occupation is seeing a gradually lower proportion of AI-related patent exposure, even if the absolute number of patents filed in that area is going up. This demonstrates a flaw with the metric; namely, that it is relative to patents filed each year, and does not take into account the broader trend of AI patent-filing increasing generally.

4 Examples of AI Exposure Measure in Econometric Analyses

In this section, we discuss various use cases of this measure in econometric analyses, including different settings where AI exposure could be useful – drawing upon principles learned in class.

Example of use directly in OLS

When using OLS to predict the economic impact of artificial intelligence (AI), we might, for example, look at the relationship between AI exposure and average wages in the tech industry. Here, the AI exposure measure is constructed based on the relevance of recent patents to the tech sector. By running an OLS regression with wage as the dependent variable and AI exposure as one of the independent variables (alongside others like education level, experience, and region), we aim to estimate how changes in AI exposure might predict changes in wages. This method assumes a direct, observable relationship between the variables and relies on historical data to simulate future wage distributions under different AI exposure scenarios.

One way to test this would be to develop a measure of AI exposure for various occupations based on the frequency and relevance of recent patents (from 2021 to 2023) to these occupations. Then, apply the estimated coefficients from historical regressions to simulate how the wage distribution might shift under different scenarios of AI exposure. This simulation assumes that the relationship between exposure to past technologies and labor market outcomes can be extrapolated to AI.

To do so, we'd construct a model like:

$$\text{Wage}_i = \beta_0 + \beta_1 \cdot \text{AIExposure}_i + \beta_2 \cdot \text{EducationLevel}_i + \beta_3 \cdot \text{Experience}_i + \beta_4 \cdot \text{Region}_i + \epsilon_i$$

Here, Wage_i is the dependent variable representing the wage of individual i, AIExposure_i is the independent variable of interest measuring the extent of AI exposure, and β_1 is the coefficient estimating the impact of AI exposure on wages. The model also includes control variables like EducationLevel, Experience, and Region with respective coefficients $\beta_2, \beta_3, \text{and} \beta_4$, and ϵ_i is the error term. This model aims to isolate the effect of AI exposure on wages while controlling for other factors.

Example of endogenity concerns from direct OLS

This may occur if there are unobserved factors influencing both AI exposure and labor market outcomes that are not included in the model.

For example, let's say we're examining the impact of AI exposure on productivity in manufacturing. Suppose there is that firms with higher productivity are more likely to invest in AI, creating a simultaneity issue where productivity influences AI exposure just as much as AI exposure influences productivity. Additionally, there might be unobserved factors like managerial skill or company culture influencing both productivity and the likelihood of adopting AI, leading to omitted variable bias. Measurement error could also be a concern if our measure of AI exposure doesn't accurately capture the true extent of AI integration in the firm's processes. These endogeneity issues could lead to biased and inconsistent estimates of the relationship between AI exposure and productivity.

Suppose the true model is:

$$\text{Productivity}_i = \alpha_0 + \alpha_1 \cdot \text{AIExposure}_i + \alpha_2 \cdot \text{ManagerialSkill}_i + \alpha_3 \cdot \text{CompanyCulture}_i + u_i$$

Here, Productivity_i is the dependent variable, and AIExposure_i is the potentially endogenous independent variable. α_1 measures the effect of AI exposure on productivity. However, simultaneity, omitted variables (ManagerialSkill and CompanyCulture), and measurement error in AIExposure_i might cause the error term u_i to be correlated with AIExposure_i, introducing bias in the estimation of α_1.

Example of use as a proxy variable

This AI automation measure may be used to address measurement error or omitted variable bias by finding a variable that is correlated with the unobservable but not with the error term. If some direct measurement of automation is flawed, this metric may serve as a proxy variable closely related to the true exposure, used in its place to capture its effect on the labor market.

For example, if one wants to understand how AI integration affects job satisfaction, it may be difficult to measure AI integration like the following:

$$\text{JobSatisfaction}_i = \beta_0 + \beta_1 \cdot \text{AIIntegration}_i + \beta_2 \cdot \text{ControlVariables}_i + \epsilon_i$$

Instead, we may introduce a proxy, in the form of our AI exposure method:

$$\text{JobSatisfaction}_i = \alpha_0 + \alpha_1 \cdot \text{AIExposureMetric}_i + \alpha_2 \cdot \text{ControlVariables}_i + u_i$$

In this model, $\text{AIExposureMetric}_i$ serves as a proxy for the actual, complex, and difficult-to-measure concept of AI integration in the workplace. This metric, derived from patent data and occupational descriptions, reflects the degree to which an individual's occupation is likely to be affected by AI.

Example of use as an instrumental variable (IV) estimator

The AI exposure measure can be used as an instrument for the potentially endogenous explanatory variable in the model, where this instrument is not correlated with the error term in the regression.Then, one would regress the endogenous variable (technology exposure) on the instrument (the constructed AI exposure measure) and other exogenous variables – and then use the predicted values from this regression as a proxy for the actual, potentially endogenous values in the main OLS regression.

In an IV approach, suppose we're looking at how AI exposure affects employment levels in the automotive industry. We might suspect that firms experiencing higher growth are both more likely to adopt AI and increase employment, which introduces endogeneity. As an instrument, we could use AI research patents related to the firm's specialty, which are presumably related to the firm's AI exposure but not directly to their employment decisions (except through AI). The IV estimator then helps us to isolate the variation in AI exposure that is independent of the firm's employment levels, providing a more accurate estimate of the causal effect of AI on employment.

5 Conclusions & Future Work

Limitations

For the following reasons, these methodologies and results should be interpreted with caution:

1. **Cosine Similarities of Embeddings as a Reliable Metric:** Notably, the cosine similarities seem to be quite close and within a limited range – indicating that all patent and occupational data tends to be quite similarly related. However, given such constant cosine similiarities between texts, it could be possible that the differences may be due to other features of the text other than relevant content. Further investigation is needed to ensure that (a) the embeddings can unlock discern meaningful differences and (b) the model is not locking onto spurious cues, like phrasing or sentiment, to classify differences.

2. **Patents as a Reliable Metric of Innovation:** The assumption that patents can be a reliable metric of innovation is deeply flawed.

(a) Most software solutions developed are protected under copyright law, not patent law (cop, 2023).

(b) Furthermore, most modern AI research is published publically on arXiv or Github, where it is open, free, and not patentable. Furthermore, many open-source models developed by major companies and countries (e.g. Meta, UAE, EleutherAI, StabilityAI, MosaicML) are publically available on HuggingFace.

(c) Generally, AI research tends to be far more driven by disruption rather than intellectual property discussion (a prominent researcher at NeurIPS 2023 in December, Luke Sernau, commented in the Scaling, Alignment, & Open-Source AI workshop that "the rate at which a research institution leaks info is proportional to as hard as they try to keep thm"). The moats developed by big AI companies center not around architectures and training techniques, but rather scale in compute and data used for models as well as market access.

(d) Because of all of the above, the patents published may be dealing with the few edge cases where AI research is patentable, and overrepresenting those. For example, hardware tends to be patentable – therefore, research relevant to computer networks are more likely to be patented than language APIs, thus explaining the previously-observed (unreasonably) high exposure of computer network engineers among AI patent data.

To fix this, I could also include other types of language data in my analysis, such as web-scraping open-source software and activity on Github or HuggingFace, tracking usage of specific key software(s) for various sectors (e.g. Microsoft Word, ChatGPT, Google, Github Copilot), or analyzing top computer science arXiv and conferences (which is where many significant advancements are made and published).

3. **Relative, Rather Than Absolute, Patent Metric Calculation:** There is an assumption made with the metric used that the proportion of AI-related patents determine its exposure; however, when comparing across different years, the absolute number of AI-related patents filed can increase while its relative exposure can decrease due to the broader trend of AI patent-filing increasing trend. Measuring absolute exposure or normalizing regularly can be used to solve this issue.

Conclusion

Outside of this, future work could generally include explore different embedding models (or training a custom occupation-focused embedding model with a specific loss function), incorporate more granular occupational data, and experiment with alternative statistical similarity metrics to refine the exposure metric further.

In conclusion, our research offers a novel approach to quantifying the exposure of various occupations to AI automation. By analyzing the semantic similarity between patent data and occupational information, I have explored a potential technique that can help policymakers, researchers, and industry professionals forecast and prepare for the impacts of AI on the labor market.

Appendix

SQL Query Used to Filter Patent Data in Google Patents Public Data

```
SELECT
  title.text AS title,
  abstract.text AS abstract,
```

```
  publication_number as patent_number,
  publication_date
FROM
  `patents-public-data.patents.publications`,
  UNNEST(title_localized) AS title,
  UNNEST(abstract_localized) AS abstract
WHERE
  (title.text LIKE '%neural network%'
  OR abstract.text LIKE '%neural network%'
  OR title.text LIKE '%machine learning%'
  OR abstract.text LIKE '%machine learning%'
  OR title.text LIKE '%deep learning%'
  OR abstract.text LIKE '%deep learning%'
  OR title.text LIKE '%artificial intelligence%'
  OR abstract.text LIKE '%artificial intelligence%')
  AND title.language = 'en'
  AND abstract.language = 'en'
  AND country_code = 'US'
```

Commentary on Libraries Used and Engineering Details

In the provided Jupyter notebook in my Github repository, a comprehensive data processing and analysis pipeline is established, primarily focusing on patent data and its correlation with various occupations over time.

Data Cleaning and Initial Processing: The process begins by importing the `pandas` library and defining the `process_raw_data` function. This function reads raw patent data from a CSV file, eliminates any duplicate entries, and concatenates the title and abstract of each patent into a new 'all_text' column. The publication dates are converted into datetime objects to extract the year, and the data is then split and saved into separate CSV files for each year. This step ensures that the dataset is clean, organized, and prepared for more detailed analysis.

Embedding and Similarity Calculation: To explore the relationship between patents and occupations, the notebook employs OpenAI's GPT embedding API. It uses a combination of `pandas`, `openai`, and `concurrent.futures` for parallel processing to generate embeddings for each patent's text. With these embeddings, it calculates the cosine similarity between patents and various occupations. These similarities aim to quantify the relevance or association between the content of patents and the skills or knowledge required in different occupations.

Statistical Analysis and Further Visualization: After calculating cosine similarities, the notebook performs statistical analysis to understand the distribution of similarities for each occupation across different years. Using `seaborn` and `matplotlib`, it creates distribution plots for cosine similarities, marking key statistics like mean and median. In a more detailed analysis, I calculate mean cosine similarity for each year and occupation, providing a granular view of the evolving landscape. It also determines the percentage of patents that fall under "exposure" and "high exposure" categories based on predefined thresholds of cosine similarity. Then, I use `matplotlib` and `pandas` to create line plots, showing the trends of medium exposure percentage over the years for selected occupations such as Accountants & Auditors, Administrative Law Judges, and Computer Network Architects.

References

(2023). 17 U.S.C. § 101. United States Code. Copyright Act, Section 101.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive NLP tasks.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Thongtan, T. and Phienthrakul, T. (2019). Sentiment classification using document embeddings trained with cosine similarity. In Alva-Manchego, F., Choi, E., and Khashabi, D., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 407–414, Florence, Italy. Association for Computational Linguistics.

Webb, M. (2019). The impact of artificial intelligence on the labor market. Available at SSRN: https://ssrn.com/abstract=3482150 or http://dx.doi.org/10.2139/ssrn.3482150.