LUISS

Department of Business Management

Decoding the GenAI Workforce:

An NLP and Machine Learning Analysis of Evolving U.S. Labor Market Demands, Featuring a Healthcare Deep Dive

Simone Filosofi
Student ID: 284531

Supervisor: Prof. Irene Finocchi

Academic Year: 2024/2025

“The first rule of any technology used in a business is that automation applied to an efficient operation will magnify the efficiency. The second is that automation applied to an inefficient operation will magnify the inefficiency.”
— Bill Gates

Abstract

The paradigm shift initiated by Generative Artificial Intelligence (GenAI) is reshaping the global labor market with unprecedented speed and scope. This thesis presents a rigorous, data-driven investigation into the contemporary impact of GenAI, moving beyond speculative discourse to deliver empirical insights into actualized talent demands. Leveraging a computational pipeline, we analyzed 2,726 unique U.S. job advertisements (May 2023 - March 2025) scraped from prominent online platforms. Our methodology incorporated advanced Natural Language Processing techniques, including Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, Cosine Similarity and Levenshtein Distance-based fuzzy string matching (via Rapidfuzz) for robust job title standardization against the O*NET occupational database. Unmatched titles were subsequently classified using a trained Support Vector Machine (SVM) model. Latent Dirichlet Allocation (LDA) was then employed on job descriptions to uncover latent thematic skill clusters and industry-wide adoption patterns. Crucially, a dedicated LDA analysis was subsequently applied to the healthcare sector, unearthing domain-specific trends and skill requirements. Key findings reveal a quantifiable evolution in demanded roles, a significant pivot towards in-house GenAI development, the notable emergence of ’Intern’ positions, and dynamic shifts in compensation structures. Our focused analysis of the healthcare field further illuminated distinct sectoral adoption dynamics and the specialized GenAI-related competencies being prioritized. This research offers a granular, empirically-grounded perspective on the evolving skillsets, occupational structures, and strategic hiring priorities defining the nascent GenAI-driven workforce, with specific insights into its transformative role within the healthcare industry, and compares current results with findings from previous years’ studies to contextualize observed trends.

Contents

1 Introduction

In this initial chapter, we will establish the foundational context for this research by exploring the recent advancements in Artificial Intelligence, particularly Generative AI and their burgeoning impact on the global job market. Subsequently, we will articulate the primary purpose of this thesis: to conduct a data-driven analysis of AI job advertisements to understand current talent demands, role distributions and cross-industry adoption.

1.1 Background

The dawn of the 21st century has been characterized by relentless technological advancement, with Artificial Intelligence (AI) emerging as a dominant force reshaping industries, economies and societies globally. While AI, in its broader sense, has been a subject of research and application for decades, encompassing innovations from machine learning algorithms to robotic process automation, a recent and profound innovation has captured global attention: Generative AI (GenAI). Unlike its predecessors, which primarily focused on analytical, predictive, or task-execution capabilities, GenAI possesses the remarkable ability to create novel content – text, images, audio, code and complex data syntheses – that is often indistinguishable from, or even surpasses, human-generated output (Goodfellow et al., 2014; OpenAI, 2023). This paradigm shift from analytical to creative AI marks a pivotal moment, heralding an era of unprecedented potential and profound uncertainty, particularly concerning its impact on the labor market.

The public's widespread encounter with GenAI arguably began with the launch of models like OpenAI's ChatGPT in late 2022, which rapidly demonstrated the power of large language models (LLMs) to a global audience. This was swiftly followed by a proliferation of sophisticated tools such as DALL-E 2, Midjourney, Stable Diffusion for image generation and GitHub Copilot for code assistance, among others. These developments are not merely incremental improvements: they represent a step-change in AI capabilities, as noted by scholars who see these tools poised to "radically transform economies and societies worldwide", holding "both promise and peril" (George, 2024, p.18). This innovation is distinct because it democratizes content creation and complex problem-solving assistance to an unprecedented degree, moving beyond routine task automation into realms previously considered exclusive to human cognition and creativity.

Historically, technological advancements have always reshaped labor markets, from the mechanization of agriculture to the computerization of office work in the 20th century. Economists have long studied this interplay, often finding that "rather than long-term net job destruction, [...] productivity enhancements open new opportunities" (George, 2024, p.18), a cycle Schumpeter termed 'creative destruction'. However, the prevailing scholarly viewpoint is that while GenAI, like AI more broadly, will revolutionize work, it might not result in long-term net job losses but rather in significant job shifting (George, 2024; Autor, 2015). Tyson and Zysman, 2022, characterize AI as "routine-biased technological change on steroids", suggesting an intensification of automation's adverse effects on labor, including employment polarization and stagnant wage growth for certain skill segments, yet remaining optimistic about interventions. The core argument is that AI is "projected to have a net impact of job shifting rather than job loss by increasing productivity, accelerating economic growth, changing the structure of jobs and allowing sectoral employment transitions" (George, 2024, p.17).

Despite this generally optimistic outlook on net employment, significant concerns persist regarding the nature of this transition and its distributional consequences. The speed, scope and scale of GenAI's encroachment into cognitive tasks are arguably different from previous technological waves (Brynjolfsson and McAfee, 2014; Acemoglu and Restrepo, 2019). Professions involving writing, design, coding, and even scientific research, are now experiencing direct augmentation or potential displacement by GenAI tools. This raises critical questions about:

Skill Demand and Reskilling: As GenAI automates certain cognitive tasks, the demand for complementary human skills – such as critical thinking, complex problem-solving, emotional intelligence and AI management – is expected to rise (Deming, 2017; De Mauro et al., 2018). This necessitates a massive reskilling and upskilling effort, a challenge highlighted by the "imperative for continual skills adaptation" (George, 2024, p.30).

Wage Inequality and Polarization: The impact of GenAI on wages is likely to be uneven. Individuals whose skills are complemented by GenAI may see their productivity and wages increase, while those whose tasks are substitutable could face wage stagnation or decline, potentially exacerbating existing inequalities (Acemoglu and Autor, 2011).

Productivity Paradox: While GenAI promises significant productivity gains, there is an ongoing debate about whether these gains will translate into broad-based economic prosperity or be captured by a select few, a concern echoing the "productivity paradox" observed with earlier IT innovations (Brynjolfsson et al., 2019).

Ethical and Societal Implications: Beyond direct economic impacts, the rise of GenAI brings forth ethical dilemmas concerning bias in AI-generated content, intellectual property rights, the spread of misinformation and the potential for a "black box" problem where AI decision-making processes are opaque (Bender et al., 2021; European Commission, 2021). The perception of AI itself also varies significantly between experts and the public, influencing adoption and governance (Brauner et al., 2024).

Human-AI Collaboration: The future of work is increasingly envisioned as a collaborative ecosystem where humans and AI agents work in tandem. Understanding and designing effective Human-AI Interaction (HAII) becomes paramount to harness the full potential of GenAI while mitigating its risks (Xu and Dainoff, 2021; Shneiderman, 2020).

The emerging landscape requires a nuanced understanding that moves beyond simplistic narratives of utopia or dystopia. As models like DeepSeek demonstrate increasingly sophisticated mathematical and logical reasoning (Jahin et al., 2025) and tools like ChatGPT find applications in specialized domains such as financial forecasting (Bi and Xiao, 2023), it becomes clear that the impact will be multifaceted and sector-specific. The challenge, therefore, is not to predict a singular outcome but to analyze the complex dynamics at play and to identify pathways for an equitable and productive integration of GenAI into the labor market.

1.2 Purpose of the thesis

The primary purpose of this thesis is to conduct a rigorous, data-driven and empirical analysis of the contemporary labor market's dynamic response to the rapid proliferation and increasing sophistication of Generative Artificial Intelligence (GenAI). While the background discussion has highlighted the transformative potential and widespread discourse surrounding GenAI, this research moves beyond general projections to provide a granular examination of actualized talent demands. Leveraging a comprehensive and recent dataset of online job advertisements sourced from prominent platforms, this study specifically aims to identify and meticulously characterize the emerging roles and responsibilities explicitly or implicitly requiring GenAI expertise. Furthermore, it seeks to map the distribution of these AI-related roles across a diverse spectrum of industries and company profiles, thereby illuminating the breadth and depth of GenAI adoption. A crucial objective is also to delineate the key technical and soft skills, alongside the desired experience levels, currently sought by employers, offering a detailed snapshot of the qualifications defining this rapidly evolving technological frontier within the workforce.

2 Literature Review

As written in the introduction, the rapid proliferation and increasing sophistication of Artificial Intelligence (AI), particularly its advanced generative forms (GenAI), have become defining features of the contemporary technological landscape, prompting extensive academic and public discourse on its profound implications for the future of work and the structure of the labor market. This review synthesizes the existing body of research, examining the theoretical frameworks, methodological approaches and empirical findings that illuminate our current understanding of AI's transformative role. It will trace the evolution of scholarly thought from general principles of automation and technological change to the unique challenges and opportunities presented by GenAI. Ultimately, this chapter serves to contextualize the present thesis, which distinguishes itself through a rigorous quantitative investigation leveraging contemporary numerical data, a perspective that aims to provide empirical grounding to a field often characterized by rapid innovation and, consequently, a degree of predictive uncertainty.

2.1 The Emergence of Generative AI: A Paradigm Shift in Artificial Intelligence

Generative AI represents a significant departure from prior AI paradigms. Its defining characteristic lies in the remarkable capability of these systems to not merely analyze or interpret existing data, but to create novel and original content. By learning intricate patterns and underlying structures from vast and diverse datasets - encompassing text, visual imagery, auditory information and even complex code - GenAI models can produce new outputs that often exhibit a striking resemblance to, and sometimes surpass, human-generated work in quality and complexity (Aydın and Karaarslan, 2023; Ellingrud et al., 2023). These systems are underpinned by sophisticated deep learning architectures, most notably neural networks and advanced transformer models, which are meticulously trained on immense corpora of curated information. This extensive training enables them to generate contextually appropriate, coherent and often nuanced responses to a wide array of user inputs and complex prompts (Lim et al., 2023).

The public and professional spheres have witnessed the rapid ascent of influential GenAI tools. Conversational platforms such as OpenAI's ChatGPT and Anthropic's Claude have demonstrated an unprecedented ability to engage in fluid, human-like dialogue and perform a multitude of language-based tasks (Felten et al., 2023; Lim et al., 2023). Concurrently, tools like GitHub Copilot have begun to revolutionize software development by assisting with code generation, while image synthesis models such as DALL-E and Midjourney have unlocked new frontiers in visual content creation. This fundamental capacity for original creation, rather than the predominantly analytical or predictive functions of earlier AI systems, positions GenAI as a distinct and potentially more pervasively disruptive technological force within the labor market and beyond.

2.2 Anticipated Labor Market Dynamics: Projections of Disruption, Transformation and Productivity

A strong and growing consensus within the scholarly community and among economic analysts posits that the labor market is on the cusp of, or already undergoing, substantial disruption attributable to GenAI's advanced capabilities in automating a wide range of human tasks and augmenting human performance in others. This anticipated disruption is largely predicated on the significant productivity enhancements that GenAI is expected to unlock across numerous industries (Brynjolfsson et al., 2023; Brynjolfsson et al., 2018; Choudhury et al., 2020; Noy and Zhang, 2023; Peng et al., 2023). Beyond direct effects on task execution, GenAI is also poised to fundamentally reshape human-machine interaction paradigms, fostering new modes of collaboration and workflow integration (Frank et al., 2019; Sturm et al., 2021). Furthermore, the adoption of GenAI is compelling businesses to re-evaluate and develop novel business models and strategic approaches to leverage its capabilities and maintain competitive advantage (Caner and Bhatti, 2020 Kitsios and Kamariotou, 2021).

Macro-level economic projections from influential institutions attempt to quantify the potential scale of this impending transformation. For instance, an analysis by Goldman Sachs (Hatzius, 2023) made headlines by suggesting that a substantial percentage of existing job roles in developed economies like the U.S. and Europe face some degree of automation potential due to AI. Their report further indicated that a significant portion of current work tasks could be directly handled or significantly altered by GenAI, potentially impacting hundreds of millions of full-time equivalent jobs globally. Similarly, research from the McKinsey Global Institute (Ellingrud et al., 2023) estimates that GenAI-accelerated automation could influence a considerable share of working hours across diverse sectors in the United States by the year 2030. Their analysis points towards significant shifts between occupational categories, with anticipated growth in fields requiring scientific, technological and healthcare expertise, while roles in customer service and administrative support are identified as facing higher risks of displacement or significant task restructuring. Surveys capturing business sentiment, such as the comprehensive report by the World Economic Forum, 2023, further reinforce these expectations, revealing widespread intentions among companies globally to adopt AI technologies, albeit with varied outlooks on net job creation versus potential job losses attributable to these advancements.

This narrative of disruption, however, is not monolithic and is often nuanced by arguments emphasizing transformation and augmentation over outright replacement. Influential work by Frey and Osborne, 2017, while acknowledging that GenAI has extended the reach of automation into tasks demanding creativity and sophisticated social intelligence, posits that its capabilities remain comparatively limited in unstructured physical environments and roles requiring deep, in-person interaction. Consequently, they argue that GenAI is more likely to transform the nature of existing jobs - by augmenting human capabilities and restructuring task compositions - rather than leading to widespread, full-scale replacement of human workers. This perspective suggests a continuing and perhaps even enhanced, role for human skills, particularly those involving complex problem-solving, critical thinking, emotional intelligence and interpersonal engagement. This view aligns with historical interpretations of technological impact on labor, as detailed by Autor, 2015, and finds resonance in AI-specific analyses by George, 2024, both of whom emphasize the dynamic of job shifting, task recomposition and the concomitant emergence of new occupational demands. Tyson and Zysman, 2022, further contribute to this nuanced understanding by framing AI as an accelerator of existing "routine-biased technological change", underscoring that while the effects may be intensified, the ultimate societal and labor market outcomes are not predetermined and will be significantly shaped by "intelligent policies" and proactive adaptation strategies.

2.3 Methodological Frameworks for Assessing AI's Labor Market Impact

The academic pursuit of understanding and quantifying AI's influence on employment and work, has led to the development and refinement of several key research methodologies. These approaches provide structured frameworks for analyzing the complex interplay between technological capabilities and labor market structures.

2.3.1 Task-Based Analysis of AI Exposure and Occupational Vulnerability

A dominant and highly influential methodological stream involves estimating the potential for AI to affect various occupations by meticulously examining the constituent tasks that define those roles. This approach typically begins with a detailed theoretical framework that delineates AI's current and near-future capabilities, which are then systematically mapped to the array of tasks performed by human workers across the economy, often utilizing comprehensive occupational databases like the U.S. Department of Labor's O*NET. Aggregated measures of this AI exposure at the occupational level are subsequently used to evaluate broader labor market implications, such as potential displacement, wage effects and shifts in skill demand. Foundational work by Frey and Osborne, 2017, exemplified this by using expert assessments and statistical modeling based on O*NET data to predict the probability of computerization for a wide range of occupations, identifying tasks reliant on perception, manipulation, creative intelligence and social intelligence as key "bottlenecks." Building on this, Felten et al., 2018, 2021, introduced the AI Occupational Exposure (AIOE) measure, directly linking specific AI capabilities like image recognition and natural language processing to the human abilities required for different jobs and have since updated their analyses to reflect the advancements brought by LLMs. Another significant contribution is Webb's (2020) novel approach, which employed Natural Language Processing to measure the textual overlap between task descriptions within AI patents and those in O*NET, providing a data-driven measure of which occupational tasks were being directly targeted by AI innovation. Similarly, Brynjolfsson et al., 2018, developed a "Suitability for Machine Learning" (SML) index by having experts rate a vast number of detailed work tasks based on their amenability to current machine learning capabilities, subsequently aggregating these to the occupational level. More recently, researchers like Eloundou et al., 2023, have specifically focused on the impact potential of Large Language Models, using both human annotators and GPT-4 itself to evaluate the exposure of O*NET tasks to these advanced AI systems. These task-based methodologies provide crucial, albeit often predictive, insights into which segments of the labor force are most likely to experience AI-driven changes.

2.3.2 Analysis of Job Market Data: Tracking Demand for AI-Related Competencies and Roles

A second major methodological approach focuses on the direct analysis of real-world labor market data, predominantly online job advertisements and vacancy postings, to identify and track the evolving demand for AI-specific skills, emerging AI-related occupations and shifts in the skill composition of existing roles. This method offers a more contemporaneous view of how employers are adapting to technological advancements. For example, Acemoglu et al., 2022, analyzed a massive dataset of online job vacancies in the US from 2010 to 2018, identifying "AI-exposed" establishments and "AI-using" vacancies to examine how AI adoption within firms affected their overall hiring patterns and the specific skill requirements within those job postings. International comparative work by Squicciarini and Nachtigall, 2021, utilized a large dataset of online job vacancies from Burning Glass Technologies across several countries to track the growth in demand for AI skills, noting the types of skills requested and the industries driving this demand. They highlighted not only the rise in technical AI skills but also the increasing importance of complementary non-technical skills such as communication and problem-solving. Other studies, such as those by Kortum et al., 2022, have used job posting data to compare skill demands in different AI subfields, for example, differentiating between requirements for computer vision and NLP roles. These analyses of job market data provide valuable, demand-side evidence of AI's unfolding impact on skill requirements and occupational structures.

2.3.3 Econometric Analyses of Early Generative AI Impacts on Labor Outcomes

The recent and rapid proliferation of powerful, publicly accessible GenAI tools, particularly since late 2022 with the widespread adoption of models like ChatGPT, has spurred a nascent but crucial stream of research employing econometric techniques to assess their immediate, observable impacts on labor market outcomes. These studies often leverage quasi-experimental designs to isolate the effects of GenAI. For instance, Hui et al., 2024 and independently Liu et al., 2023, both utilized a difference-in-differences (DiD) methodological approach on data from large online labor markets (freelancer platforms). They compared changes in job postings, earnings and task complexity for occupations highly exposed to ChatGPT's capabilities (such as writing, translation and certain programming tasks) versus less exposed occupations, examining trends before and after ChatGPT's public release. Their findings generally pointed to significant short-term declines in job availability and earnings for freelancers in the most AI-affected occupations, alongside shifts in the nature of the work demanded. While not an observational econometric study of labor market outcomes, the experimental work by Noy and Zhang, 2023, which conducted a randomized controlled trial (RCT) with professionals performing writing tasks with and without access to ChatGPT, provided crucial micro-evidence on GenAI's substantial productivity effects, a key channel through which it is expected to influence broader labor market dynamics.

These early econometric studies are vital for providing initial, data-driven assessments of GenAI's real-world consequences, moving beyond prediction to direct observation.

2.4 Evolving Understandings of AI's Impact and the Imperative for Data-Driven Quantitative Insights

Beyond direct econometric and task-based analyses of employment, the literature increasingly acknowledges the critical importance of broader contextual factors in shaping AI's ultimate impact on the labor market and society. The field of Human-AI Interaction (HAII), as discussed by Xu and Dainoff, 2021, highlights the necessary evolution from traditional Human-Computer Interaction (HCI) principles. They emphasize that the unique characteristics of AI systems - such as their capacity for autonomous learning, potential for non-deterministic outputs and often opaque decision-making processes - demand new design paradigms that prioritize human control, transparency and ethical considerations. Furthermore, societal perceptions and trust, as explored by Brauner et al., 2024, significantly influence AI's adoption and governance, while the continuously advancing cognitive capabilities of AI models, demonstrated in areas like mathematical reasoning by Jahin et al. (2025), necessitate ongoing re-evaluation of human-AI collaboration. This comprehensive review of such contextual factors and existing impact assessment methodologies reveals a vibrant and rapidly expanding body of research dedicated to understanding the multifaceted interactions between Artificial Intelligence and the labor market. However, while this body of research is extensive and offers crucial foundational knowledge, a careful examination, particularly concerning the very latest wave of sophisticated and widely accessible GenAI tools, indicates that comprehensive, large-scale quantitative analyses leveraging broad, contemporary numerical datasets are still relatively nascent. Many existing studies, by necessity of data availability or research focus, either predate the full societal and economic impact of current GenAI models or are more qualitative or predictive in nature. The present thesis is specifically designed to contribute to this evolving discourse. It does so by integrating key elements from the diverse research methodologies explored in this literature review, thereby addressing the identified need for contemporary, broad-based quantitative evidence and aiming to provide a data-grounded assessment of GenAI's tangible and evolving influence on the modern labor market.

3 Methodology

This section details the systematic approach undertaken to collect, pre-process, and analyze job market data, focusing on roles related to Generative AI. The methodology encompasses data acquisition from online job platforms, a multi-stage job title standardization process and a two-step topic modeling approach to uncover thematic trends and skill requirements.

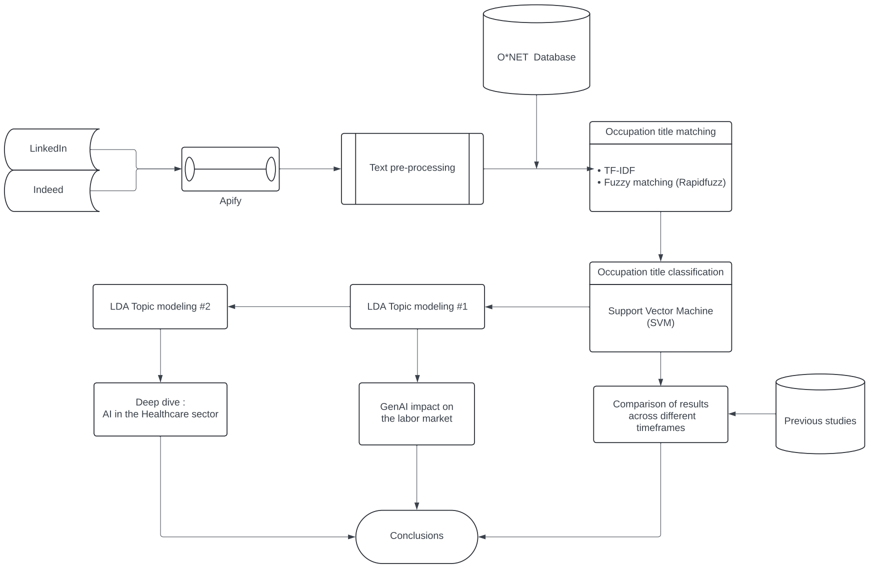

O*NET Database
LinkedIn
Indeed
Apify
Text pre-processing
Occupation title matching
• TF-IDF
• Fuzzy matching (Rapidfuzz)
Occupation title classification
Support Vector Machine (SVM)
LDA Topic modeling #2
LDA Topic modeling #1
Deep dive : AI in the Healthcare sector
GenAI impact on the labor market
Comparison of results across different timeframes
Previous studies
Conclusions

Figure 1: Research methodology steps

3.1 Techniques

To provide readers with a clear understanding of the analytical framework, this section outlines the core computational techniques utilized in this research, as depicted in Figure 1. These methods range from initial data acquisition strategies to advanced textual analysis and classification models, forming the backbone of our investigation.

3.1.1 Apify Actors for Data Scraping

Automated data collection was performed using Apify, a cloud platform designed for web scraping and automation. Apify utilizes Actors, which are serverless microservices packaged as Docker containers, to execute data extraction tasks. Each Actor encapsulates arbitrary code along with defined schemas for its JSON input and output.

An Actor's definition includes a Dockerfile (specifying the code and runtime environment), metadata (such as name and version), documentation (README) and the input/output schemas that detail its operational parameters and expected results.

This architecture simplifies execution and integration, as the Apify platform automatically generates a user interface (UI) and API for each Actor based on its definition. The platform also manages the underlying infrastructure, including execution, scaling and storage, abstracting these complexities from the developer. Actors operate in isolated Docker containers with access to built-in storage systems (e.g., key-value stores, datasets). They can be scheduled, triggered by webhooks, call other Actors or integrate with external services, enabling the construction of complex scraping workflows from modular components. For this research, the specific Actor "LinkedIn Jobs Scraper" (https://apify.com/bebity/linkedin-jobs-scraper) was utilized to systematically gather job advertisement data from LinkedIn, while for Indeed "Indeed Scraper" was used (https://apify.com/misceres/indeed-scraper).

3.1.2 TF-IDF and Cosine Similarity for Textual Analysis

Term Frequency–Inverse Document Frequency (TF-IDF) is a foundational statistical technique in information retrieval and text mining used to quantify the importance of a word within a document relative to a collection of documents (corpus). The TF-IDF weight, for a term t in a document d, is the product of two metrics:

- **Term Frequency (TF)**: Typically the raw count or normalized frequency of term t in document d: it measures how often a term appears in a specific document.

- **Inverse Document Frequency (IDF)**: Calculated as log(N/df_t), where N is the total number of documents in the corpus and df_t is the number of documents containing term t (nlp.stanford.edu). IDF down-weights common terms (appearing in many documents) and up-weights rarer, more specific terms.

The concept of IDF was pioneered by Jones, 1972 as "term specificity", highlighting that less frequent terms often carry more informational content for distinguishing between documents. For instance, a term present in all documents has an IDF of 0, rendering its TF-IDF score 0, indicating no discriminative power. Conversely, a term frequent in a particular document but rare across the corpus receives a high TF-IDF score, marking it as a key descriptor for that document.

TF-IDF forms the basis of the classic vector space model, where documents are represented as vectors of term weights. To compare these document vectors, **Cosine Similarity** is employed. It measures the cosine of the angle between two TF-IDF vectors (v_1 and v_2), calculated as:

$$\cos(\theta) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \, \|\mathbf{v}_2\|}$$

This is the dot product of the vectors divided by the product of their magnitudes (nlp.stanford.edu). The result is a normalized similarity score, typically between 0 and 1 for non-negative TF-IDF vectors, indicating the degree of content overlap.

The combination of TF-IDF weighting and cosine similarity has been a cornerstone of information retrieval for decades, used in early search engines and document retrieval systems, to rank documents by relevance. Its enduring utility stems from its ability to balance topic-specific terms against ubiquitous ones, often serving as a baseline or component in modern NLP pipelines for tasks like text classification, clustering, keyword extraction and recommender systems.

3.1.3 Fuzzy String Matching and Rapidfuzz (Levenshtein Distance)

To identify and group similar but not identical strings (e.g., job titles with minor variations), fuzzy string matching techniques were employed. These methods find strings that are approximately equal, accommodating typos or slight differences. The most common metric for this is the Levenshtein distance (or edit distance), defined as the minimum number of single-character edits (insertions, deletions or substitutions) required to change one string into another. Each edit typically has a cost of 1. A distance of 0 signifies identical strings, while larger distances indicate greater dissimilarity. The Levenshtein distance is commonly computed using a dynamic programming approach.

For this research, the Rapidfuzz library was used. Rapidfuzz is a high-performance Python library, inspired by FuzzyWuzzy, that implements Levenshtein distance and other string similarity metrics in optimized C++ with Python bindings, offering significant speed advantages over pure-Python implementations (https://github.com/rapidfuz z/RapidFuzz).

$$d(i,j) = \begin{cases} 0, & i=0,\ j=0, \\ i, & j=0,\ 1 \le i \le m, \\ j, & i=0,\ 1 \le j \le n, \\ \min\left\{ \begin{array}{l} d(i-1,j)+1, \\ d(i,j-1)+1, \\ d(i-1,j-1)+\delta(a_i,b_j) \end{array} \right\}, & \text{otherwise} \end{cases}$$

$$\delta(a_i,b_j) = \begin{cases} 0, & a_i = b_j, \\ 1, & a_i \neq b_j. \end{cases}$$

In essence, a table d(i,j) is constructed where each entry represents the minimum edits to transform the first i characters of string a into the first j characters of string b. The value d(m,n) provides the Levenshtein distance between the full strings. RapidFuzz also provides normalized similarity scores (e.g., a ratio score), often scaled from 0 to 100, for a more intuitive percentage-based measure of string similarity.

3.1.4 Latent Dirichlet Allocation (LDA) for Topic Modeling

To uncover latent thematic structures within the corpus of job descriptions, Latent Dirichlet Allocation (LDA) was utilized. Introduced by Blei et al., 2003, LDA is a generative probabilistic model for collections of discrete data, such as text corpora. The fundamental assumption of LDA is that each document is a mixture of a finite set of latent "topics" and each topic is characterized by a probability distribution over words.

The generative process imagined by LDA is as follows: for each document, a distribution over topics (θ) is drawn from a Dirichlet prior, then, for each word in that document, a topic is selected according to θ and the word is generated from that chosen topic's word distribution. A key output of LDA is a set of interpretable topics, where each topic is represented as a list of terms with associated probabilities. These topics often correspond to coherent semantic themes (e.g., a topic might heavily feature "data", "model", "algorithm" indicating a theme around data science).

Due to its ability to produce such interpretable thematic structures, LDA has become a cornerstone of topic modeling in machine learning. It has found applications across diverse fields, including natural language processing, digital humanities, bioinformatics and social media analysis. Theoretically, LDA is a three-level hierarchical Bayesian model (corpus $\rightarrow$ documents $\rightarrow$ words) that uses Dirichlet priors, which provide mathematical convenience for inference. The core of LDA is its joint probability over all hidden and observed variables in a corpus of D documents with K topics:

$$p(\mathbf{w},\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi} \mid \alpha,\beta) \;=\; \underbrace{\prod_{k=1}^{K} p(\phi_k \,|\, \beta)}_{\substack{\text{pick a word distribution} \\ \text{for each topic}}} \quad\times\quad \underbrace{\prod_{d=1}^{D} p(\theta_d \,|\, \alpha)}_{\substack{\text{pick topic proportions} \\ \text{for each document}}} \quad\times\; \underbrace{\prod_{d=1}^{D}\prod_{n=1}^{N_d} p(z_{d,n} \,|\, \theta_d)\, p(w_{d,n} \,|\, \phi_{z_{d,n}})}_{\substack{\text{for each word: choose} \\ \text{a topic, then a word}}}$$

3.1.5 Support Vector Machine (SVM) Classification Model

Support Vector Machines (SVMs) are a class of supervised learning models, introduced in the 1990s, known for their robust theoretical foundations and strong performance in classification and regression tasks. At its core, an SVM operates as a max-margin classifier. Given labeled training data from two classes, the SVM algorithm identifies a decision boundary (a hyperplane in the feature space) that maximizes the margin - the distance between the hyperplane and the nearest data points (support vectors) of each class. This maximized margin is believed to enhance generalization to unseen data. The training instances that lie on the margin boundaries are termed support vectors, as they are critical in defining the optimal hyperplane's position.

The decision function of an SVM for a new data point x can be expressed as:

$$f(x) \;=\; \operatorname{sign}\Bigl(\sum_{i=1}^{n} \alpha_i\, y_i\, K(x_i,x) \;+\; b\Bigr)$$

Where:

- x_i **(Support Vectors):** the training examples closest to the decision boundary. Only these vectors receive nonzero coefficients α_i and thus define the boundary.

- α_i **(Weights):** learned coefficients obtained by solving the SVM's quadratic optimization problem. A larger α_i means the corresponding support vector x_i has more influence on the boundary.

- y_i **(Labels):** the true class label ($+1$ or -1) of each support vector x_i. Multiplying by y_i ensures points from opposite classes push the boundary in opposite directions.

- $K(x_i, x)$ **(Kernel Function):** a similarity measure between a support vector x_i and a new point x. For a linear SVM, $K(x_i, x) = x_i^\top x$; for nonlinear separation, one common choice is the RBF kernel:

$$K(x_i, x) = \exp\bigl(-\gamma \|x_i - x\|^2\bigr),$$

which implicitly projects data into a higher-dimensional space (the "kernel trick").

- b **(Bias):** a constant offset that shifts the decision boundary away from the origin, allowing it to separate datasets not centered at zero optimally.

When classifying a new point x, the SVM computes the weighted sum of similarities to each support vector (adjusted by their labels), adds the bias and takes the sign. If $f(x) = +1$, the point is assigned to the positive class, if -1, to the negative class. In essence, SVMs select a handful of pivotal examples and use them, via the chosen similarity measure, to draw the clearest possible boundary between classes.

3.2 Data collection and preprocessing

Following this concise overview of the methodologies employed in this research, we will now proceed to the practical application and analysis.

The initial phase involved the acquisition of job advertisements from LinkedIn and Indeed, two prominent online job portals. This was achieved using Apify, a cloud-based platform designed for large-scale web scraping, data extraction, and browser automation. To ensure the relevance of the collected data to the research focus on Generative AI, the scraping process was filtered using a curated list of keywords, that were required to be present within the job description and included terms such as 'GenerativeAI', 'GenAI', 'ChatGPT', 'Claude', 'LLM', 'Gemini', and related variations.

Due to resource constraints for broader geographical scraping, the geographical scope of the data collection was confined to the United States and, consequently, the subsequent analysis and findings are specific to the U.S. job market. The scraping process yielded job advertisements posted between May 2023 and March 2025.

Upon completion of the scraping tasks, the raw data, obtained as multiple CSV files, was consolidated to form a comprehensive dataset. Exploratory Data Analysis (EDA) and preprocessing steps were then undertaken. A critical preprocessing task was the removal of duplicate entries, that could arise from identical job postings scraped multiple times or from the same firm advertising a single position across both LinkedIn and Indeed. After deduplication and initial cleaning, the final dataset comprised 2,726 unique job postings.

While Named Entity Recognition (NER) is a technique widely employed in job market data analysis for identifying and classifying named entities such as skills, platforms, organizations and locations (Nadeau and Sekine, 2007; Upadhyay et al., 2021; Zhao et al., 2015), the specific API I used for scraping provided a structured output with numerous columns. This detailed output, which included specific information about URLs, company identifiers, application links and other metadata, obviated the need for a NER pipeline on all fields. From the extensive set of available columns, the following were selected as most pertinent for this analysis:

- `'title'`: The job title as advertised;
- `'location'`: The geographical location of the job;
- `'postedTime'`: The timestamp or date when the job was posted;
- `'publishedAt'`: The date the job advertisement was published/made live;
- `'description'`: The full text of the job advertisement;
- `'experienceLevel'`: The level of experience required (e.g., entry, senior);
- `'contractType'`: The type of employment (e.g., full-time, contract);
- `'sector'`: The industry or sector of the hiring company;
- `'salary'`: The advertised salary or salary range, if available;

The 'description' column, containing the main textual content of the job ads, underwent further preprocessing: all text was converted to lowercase to ensure uniformity and punctuation was removed using Python's str.translate method in conjunction with string.punctuation. This step is common in Natural Language Processing (NLP) as it helps to reduce the dimensionality of the data and standardize words, preventing variations (e.g., "ai." vs "AI") from being treated as distinct tokens.

3.3 Approach to Job Title identification

A crucial task in the systematic study of job posting data is the identification and classification of these postings based on their job titles, which are obviously often expressed in natural language. The inherent variability and complexity of natural language necessitate the application of NLP and Machine Learning (ML) techniques for robust standardization (Nasser and Alzaanin, 2020; Rahhal et al., 2023). This standardization is particularly important as many roles that are functionally similar may be advertised under a variety of different titles. For instance, within this dataset, job titles such as 'AI Engineer', 'Machine Learning Specialist' and 'Deep Learning Researcher' might refer to very similar underlying roles.

To address this challenge, a multi-step approach was implemented, as illustrated in Figure 1. The foundation of this approach was the Occupational Information Network (O*NET) database, which served as the reference occupational structure. O*NET, developed in the mid-1990s under the sponsorship of the U.S. Department of Labor, provides comprehensive and standardized descriptors for approximately 1,000 occupations across the U.S. economy. For this research, the 2019 version of the O*NET database, being the latest version available at the time of analysis, was utilized. The database comprises 36 distinct CSV files, offering standardized occupation titles and detailed data on job requirements and worker attributes, organized within a three-level hierarchical structure for categorizing occupations. Its utility in labor market data analysis is well-documented (Burrus et al., 2013; Peterson et al., 2001; Tippins and Hilton, 2010), and it has been cross-referenced with international occupation databases such as the International Standard Classification of Occupations (ISCO) (Hardy et al., 2018) and European Skills, Competences, Qualifications and Occupations (ESCO) (European Commission and U.S. Department of Labor, 2022). Prior to its use but also after the analysis, a manual review of the O*NET database was conducted, and certain occupations that were deemed irrelevant to the context of AI-related jobs or that for some reasons resulted highly likely to cause false positives in the matching process (e.g., 'Priest'), were removed to enhance matching accuracy.

The process of mapping job vacancy titles from the dataset to the closest O*NET occupation titles involved several stages:

1. **TF-IDF Cosine Similarity** Initially, a cosine similarity search was performed on the TF-IDF (Term Frequency-Inverse Document Frequency) matrices derived from the job titles in our dataset and the O*NET occupation titles. A high similarity threshold of 0.95 was set to ensure that only strings with a very high degree of lexical similarity were matched. This stringent threshold was chosen to minimize false positives in this initial pass, prioritizing precision. As expected, due to the nature of TF-IDF which relies on exact word matches and their frequencies, this method performed sub-optimally for a diverse set of natural language job titles, leaving 2263 out of 2726 unique job titles (83.50%) unmatched.

2. **Fuzzy String Matching** As a subsequent step, a partial ratio fuzzy string search was applied using the Rapidfuzz library in Python, which is based on the Levenshtein distance (Navarro, 2001). This technique calculates the similarity between the job titles in the advertisements and the O*NET occupation titles based on edit distance. A threshold of 0.95 for the normalized Levenshtein distance (or partial ratio score) was maintained to ensure highly similar string matches.

Combining these two methods, O*NET occupation matches were found for nearly 90% of the job postings, resulting in a significant improvement. Only 317 job titles out of the total 2,726 job postings (11.63%) remained unmatched. All matches generated by these methods were manually reviewed to ensure their quality and accuracy.

To account for the remaining 11.63% of unmatched job titles and to ensure comprehensive classification, a supervised machine learning approach was adopted. The successfully matched job titles (88.37%) and their corresponding O*NET categories were used as training data for a classification model. To prepare the data for this model, the job *descriptions* associated with these titles underwent further cleaning, including the removal of common stop words and lemmatization to reduce words to their base or dictionary form. The cleaned descriptions were then vectorized using TF-IDF to convert the textual data into numerical features suitable for machine learning.

A linear kernel **Support Vector Machine (SVM)** classification model was trained. SVMs are supervised learning models that are effective for classification tasks by finding an optimal hyperplane that best separates data points belonging to different classes in a high-dimensional space (Cortes and Vapnik, 1995) - they are widely utilized in labor market data classification (Goindani et al., 2017; Javed et al., 2015; Rahhal et al., 2023). The O*NET occupation title served as the target variable and a one-versus-rest (OvR) strategy was employed to handle the multi-class nature of the problem, where a separate binary classifier is trained for each class against all other classes. The tuned model achieved an accuracy of 0.97 and a false positive rate of 0.00028 on a held-out test set.

Model robustness was further assessed using 5-fold cross-validation during the training phase. The cross-validation accuracy scores obtained were [0.95969423, 0.97289785, 0.95691452, 0.95899931, 0.9659249]: these consistent scores indicate good generalization performance.

The trained SVM model was then applied to the remaining 317 unmatched job postings to predict and assign appropriate O*NET occupation titles. The predictions made by the model were also manually cross-checked for quality and accuracy. Once all job postings in the dataset were assigned a standardized O*NET occupation title, the O*NET hierarchical structure was used to classify each job posting into Occupation Family Levels 1, 2 and 3, progressing from the most specific (Level 3) to the broadest (Level 1) occupational categories. To give you an example of such hierarchy: a "Machine Learning Engineer" (Level 3) is part of the "Data scientists" (Level 2), that falls under "Computer and Mathematical" occupations (Level 1) .

3.4 Topic modeling

Following the job title standardization, a two-step Latent Dirichlet Allocation (LDA) topic modeling approach was performed to uncover underlying thematic structures within the job descriptions. Latent Dirichlet Allocation (LDA) is a generative probabilistic model used for discovering abstract "topics" that occur in a collection of documents (Blei et al., 2003). It assumes that each document is a mixture of a small number of topics, and that each word's presence is attributable to one of the document's topics. LDA works by attempting to learn the topic representation of each document and the word distributions for each topic from an unlabeled corpus.

The analysis up to this point, based on job titles, provided valuable insights into the types of roles and positions most in demand, however, to link these roles to the specific sectors or application domains where Generative AI skills are required, further analysis of the job descriptions was necessary. The aim was not only to identify, for example, that there is a demand for 300 data scientists, but also to understand the contexts and fields in which these data scientists are sought.

The first LDA model was applied to the entire dataset. The keywords used to guide this initial topic extraction, focusing on the context of Generative AI, were: ['chatgpt', 'chat gpt', 'llm', 'ai', 'generativeai', 'genai', 'prompt engineering']. This selection of keywords allowed the model to identify words frequently co-occurring with these terms within the job descriptions, thereby revealing potential sectors or purposes for which Generative AI expertise is required in each specific job.

In LDA, the number of topics (clusters) to be discovered can be specified in advance, along with the number of representative words for each topic. For this analysis, each topic was characterized by a sample of 15 words, while to determine the optimal number of topics, an iterative process was employed, evaluating models with a varying number of topics (from 4 to 10) and, for each iteration, the model quality was assessed using coherence and perplexity metrics. Coherence measures the semantic similarity between high-scoring words within a topic, with higher scores indicating more interpretable topics (Mimno et al., 2011), perplexity instead, assesses the model's generalization ability on unseen data, with lower values typically indicating better performance (Blei et al., 2003). In addition to these quantitative metrics, the output word clusters were manually examined to ensure that the derived topics were both statistically sound and semantically meaningful.

The coherence and perplexity results for the different numbers of topics are presented in Table 1.

Table 1: Summary of model evaluation

num_topics	**coherence**	**perplexity**
3	0.383669	-7.516706
4	0.420116	-7.421278
5	0.405068	-7.387199
6	0.376369	-7.363272
7	0.403437	-7.339949
8	0.408025	-7.322713
9	0.453346	-7.315741
10	0.439433	-7.328115

The values in Table 1 are consistent with those expected when fitting LDA to a moderately sized, real-world text corpus. Coherence scores in the range of 0.35–0.45 suggest that the top words within each topic co-occur with sufficient consistency to form semantically meaningful clusters, without overfitting to overly narrow themes. A negative log perplexity around –7 reflects the model's reasonable uncertainty in predicting held-out documents; lower (more negative) values would typically indicate overconfident and less generalizable topics. The coherence curve exhibited a discernible peak at nine topics, suggesting that this number offered a good balance between topic interpretability and granularity. These diagnostics collectively validated the stability of the nine-topic solution and the overall soundness of the LDA implementation.

The nine-topic solution yielded a list of the 15 most frequent words for each cluster, along with their respective probabilities of appearance.

Based on a careful examination of the vocabulary characterizing each topic, and with the assistance of generative AI tools like ChatGPT to help discern logical connections between the constituent words, a descriptive name was assigned to each of the nine identified topics:

- 0: Healthcare & Medical Communications
- 1: AI Business Strategy & Leadership
- 2: Core Skills & Job Qualifications
- 3: AI & Machine Learning Engineering
- 4: Digital Marketing & Sales Strategy
- 5: AI Project Strategy & Management
- 6: Diversity, Equity & Inclusion in Employment

- 7: Sales & Recruitment Operations
- 8: Generative AI Research & Innovation

While these results provided sufficient information to begin interpreting data patterns for the thesis, certain identified clusters, particularly “Healthcare & Medical Communications” and "Diversity, Equity & Inclusion in Employment" prompted further investigation. Consequently, a second stage of LDA was performed on selected topics from this initial clustering to delve deeper into specific Generative AI-related skills and applications within these domains.

Upon completion of the analysis of all nine derived topics, just one specific cluster, "Healthcare & Medical Communications," was in the end chosen to be kept in Chapter 4 ("Results and discussion"), for subsequent in-depth, granular examination.

The methodology for this second LDA phase mirrored the process detailed for the first LDA, with one key difference: the set of keywords used to guide the topic extraction was changed. The new keywords were: ‘competencies’, ’ability’, ’technology’, ’technologies’, ’experience’. This change was implemented because the objective of this second LDA was to uncover specific insights into the actual application of AI-related competencies and technologies within the previously identified high-level thematic clusters. The process of hyperparameter tuning (number of topics, words per topic) and evaluation using coherence, perplexity, and manual inspection was repeated for each of the selected sub-corpora.

4 Results and discussion

This chapter presents and discusses the key findings derived from the analysis of job advertisement data. It begins with an examination of job title clustering to reveal broad trends in AI talent demand, followed by an exploration of the companies driving this demand and the experience levels they seek. Subsequently, a Latent Dirichlet Allocation (LDA) topic modeling approach is employed to uncover deeper thematic insights from job descriptions, culminating in a focused analysis of the healthcare sector to illustrate AI's specific industry applications and talent acquisition challenges.

4.1 Beyond Pure Tech: Family-Level Clusters Reveal AI's Cross-Industry Reach

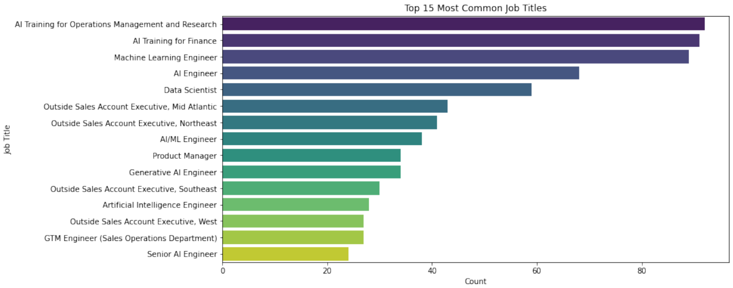


Figure 2: Top original occupation titles

Initially, we will examine the results of our job title matching. As illustrated in Figure 2, which plots the top 15 original job titles, a significant number of titles referred to essentially the same occupation. For instance, 'Machine Learning Engineer', 'AI Engineer', 'AI/ML Engineer' and 'Artificial Intelligence Engineer' all denote the same job role but were advertised with slight variations in nomenclature. Similarly, titles like 'Outside Sales Account Executive, Mid Atlantic', 'Outside Sales Account Executive, Northeast' and 'Outside Sales Account Executive, Southeast' represent the exact same job distinguished only by the geographical area. This observation underscores the necessity of applying the techniques mentioned in the previous chapter to group similar or identical jobs under a standardized title, referencing the O*NET database. The results of this consolidation are visible in Figure 3, where repetitions are eliminated.

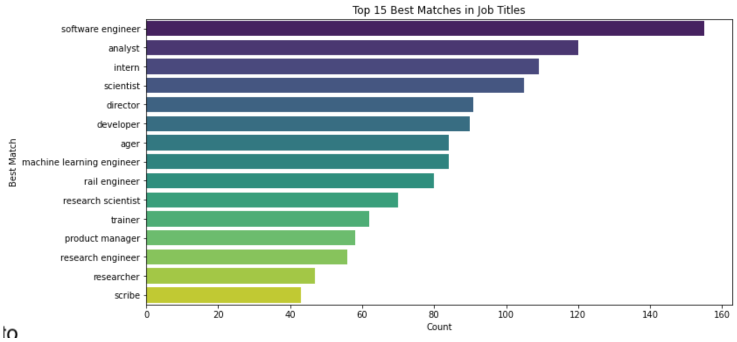


Figure 3: Top O*NET occupation titles

A striking observation from Figure 3 is the surprisingly high count for "Intern", which ranks as the third most sought-after position. This indicates a large volume of entry-level postings, highlighting that many companies in this field are actively seeking young, talented students, thereby demonstrating a belief in the capabilities of the new generation in Generative AI (GenAI) and Artificial Intelligence. Overall, engineering and technical roles dominate the top 15; positions such as 'Developer', 'Machine Learning Engineer' and 'Research Scientist' all appear with counts in the 70–90 range. However, we also find some less technical roles, such as 'Manager', 'Product Manager', and 'Scribe'. This latter term might raise questions, but upon checking the database and the classification performed during the matching process, the original roles identified as Scribes include 'Content Scribes' (working on prompt-engineering LLMs to produce first drafts of articles, then polishing them), 'Legal Scribes' (leveraging generative models to draft contracts or case summaries from recorded audio) and 'Medical Scribes' (using AI to auto-transcribe doctor–patient conversations, leaving the human "scribe" to correct errors and structure notes).

4.1.1 Employer Landscape & Experience Levels: AI Across Industries

Next, an investigation into the main companies seeking these roles was conducted, and the results were not particularly surprising. Figure 4 provides evidence that what once seemed like a race among a handful of big tech players has now become an all-out sprint involving startups, platforms, and legacy corporations alike - an unmistakable sign that the demand for AI talent currently outstrips supply across the board. Indeed, the list includes not only giants such as Meta, OpenAI and TikTok, but also some relatively small or emerging firms, not necessarily strictly AI-oriented, that are exploring the application of these instruments in other sectors.

Even Athelas in health-tech and the financial powerhouse JPMorgan Chase feature in the top 15, underlining that every industry, from medicine to banking, is scrambling to build generative AI capability.

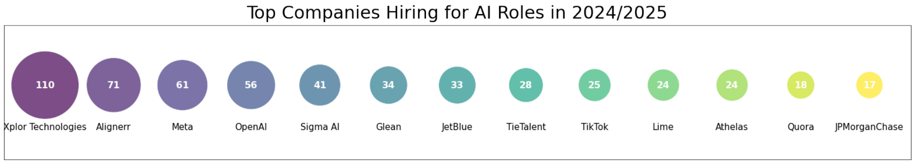


Figure 4: Companies with the most ads

As shown in Figure 5, this widespread adoption is also reflected in the experience levels companies are seeking. While internship-level positions are present, the vast majority of requests are for Mid-Senior positions (>40%), indicating a necessity to integrate AI and GenAI tools into mid-level decision-making processes. However, 'Entry Level' positions rank second, showing that companies also want to integrate these tools from the ground up. This strategy aims to build a workforce with a foundational understanding that, in the coming years, will allow employees who have scaled up to managerial positions to possess intrinsic knowledge of how to leverage these technologies - a significant competitive advantage over companies that did not act as promptly in this direction.

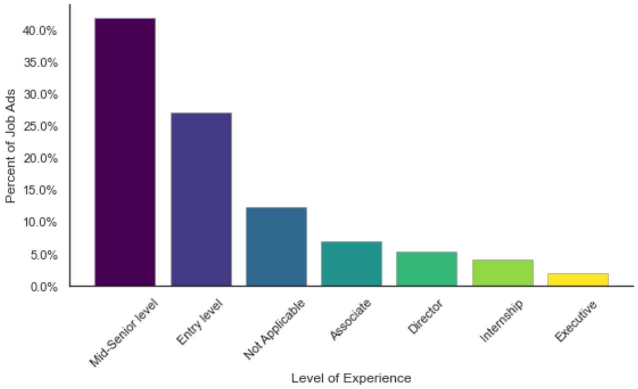


Figure 5: Degree of expertise for jobs

Thus far, we have examined the top positions among job advertisements and the companies seeking them. It has been found that many of these technical occupations are not solely for development aimed at improving technology for its own sake, as companies like OpenAI or Meta primarily do. Instead, there is a plethora of smaller companies, specialized in their respective sectors, that aim to improve their performance by integrating AI and Generative AI.

Grouping job ads one level further, up to Family 2 level (Figure 6), further supports our thesis. Roles such as developers, software engineers, analysts, scientists and researchers are indeed hired to contribute in their specific technical fields, as seen from the top two groups. However, all the other roles listed below represent different and more specific sectors or areas of interest. Among the top 15 Family Level 2 groups, we find AI applied for geography analysis, advertising and promotions, financial risk assessment, photographic processes and healthcare diagnosis. AI's reach even extends to fields that, at first glance, seem disconnected, such as history, which appears in 15th place with almost 100 ads.

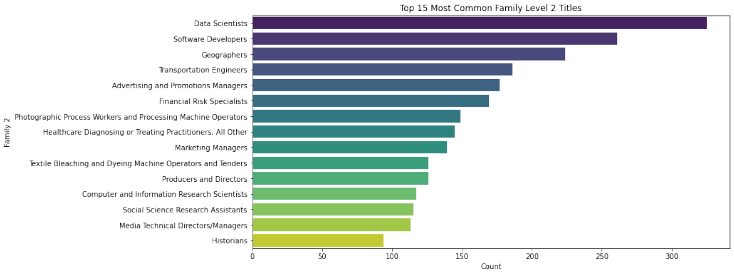


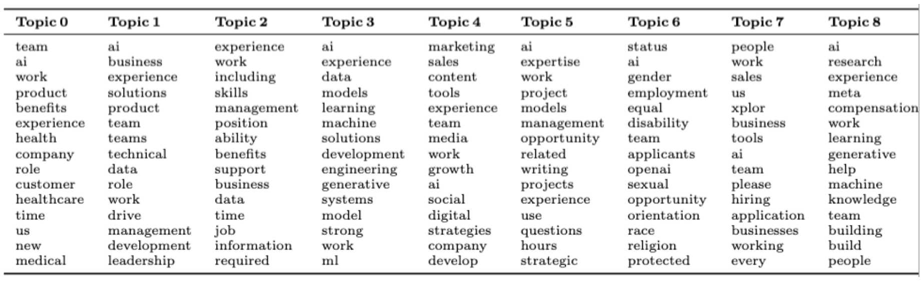
Figure 6: Family Level 2 occupation titles

This is clear evidence of how AI and GenAI are becoming integral to all aspects of society, spanning the most disparate sectors. Entities that do not wish to be forgotten or left behind by progress are actively trying to adapt.

4.1.2 From Job Titles to Job Texts: LDA-Driven Topic Discovery

The preceding section utilized O*NET database classifications for a hierarchical interpretation of job advertisements. However, to achieve a more nuanced understanding beyond job titles and company names, a deeper analysis of the actual job descriptions was undertaken.

Indeed, as detailed in the Methodology (Chapter 3), Latent Dirichlet Allocation (LDA) was performed on the descriptions of each ad to derive more in-depth insights that could help identify interesting patterns. This initial LDA resulted in the nine topics listed in Figure 7. The table displays the nine distinct clusters obtained, with the 15 most frequently found terms from the descriptions of the ads belonging to each cluster listed below it.



Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
team	ai	experience	ai	marketing	ai	status	people	ai
ai	business	work	experience	sales	expertise	ai	work	research
work	experience	including	data	content	work	gender	sales	experience
product	solutions	skills	models	tools	project	employment	us	meta
benefits	product	management	learning	experience	models	equal	xplor	compensation
experience	team	position	machine	team	management	disability	business	work
health	teams	ability	solutions	media	opportunity	team	tools	learning
company	technical	benefits	development	work	related	applicants	ai	generative
role	data	support	engineering	growth	writing	openai	team	help
customer	role	business	generative	ai	projects	sexual	please	machine
healthcare	work	data	systems	social	experience	opportunity	hiring	knowledge
time	drive	time	model	digital	use	orientation	application	team
us	management	job	strong	strategies	questions	race	businesses	building
new	development	information	work	company	hours	religion	working	build
medical	leadership	required	ml	develop	strategic	protected	every	people

Figure 7: LDA clusters with most frequent terms

Leveraging Generative AI tools, all scraped job descriptions were reviewed. By cross-referencing these descriptions with the terms highlighted in Figure 7, nine specific titles were generated, one for each cluster:

- **"Healthcare & Medical Communications"**: Characterized by keywords like "team", "ai", "health" and "medical", this clearly points to roles integrating AI within healthcare contexts. The interpretation suggests these positions focus on enhancing communication, customer service and product offerings specifically within medical or health-related industries.

- **"AI Business Strategy & Leadership"**: Features terms such as "ai", "business", "experience" and "leadership". This highlights strategic and leadership roles where AI is leveraged to drive business solutions, requiring a blend of technical expertise and managerial skills.

- **"Core Skills & Job Qualifications"**: Groups keywords like "experience", "skills", "management" and "required". This forms a general cluster centered on essential job qualifications, skills and experience requirements that are broadly applicable across various roles.

- **"AI & Machine Learning Engineering"**: Is defined by "ai", "data", "models", "machine learning" and "engineering". This signifies highly technical roles centered on AI, machine learning, and data-driven engineering, predominantly for development and research positions.

- **"Digital Marketing & Sales Strategy"**: Includes "marketing", "sales", "content", "ai" and "digital". The interpretation indicates roles in marketing, sales and digital strategy where modern AI tools and content creation are pivotal for driving business growth.

- **"AI Project Strategy & Management"**: Brings together "ai", "project", "models", "management" and "strategic". This points to positions that combine technical AI expertise with project management and strategic planning, often involving hands-on leadership in AI projects.

- **"Diversity, Equity & Inclusion in Employment"**: Stands out with keywords like "gender", "employment", "equal", "disability" and "race". This cluster is focused on issues of diversity, equity and inclusion, addressing policies and practices designed to ensure fair employment across various demographics.

- **"Sales & Recruitment Operations"**: Emphasizes "people", "sales", "hiring" and "ai". This suggests roles centered on customer interaction, sales, and recruitment, likely in environments where AI tools support both hiring processes and broader business operations.

- **"Generative AI Research & Innovation"**: Is distinguished by "ai", "research", "generative", "learning" and "build." This indicates highly research-oriented positions focused on generative AI, innovation and learning, probably involving the development of next-generation AI solutions.

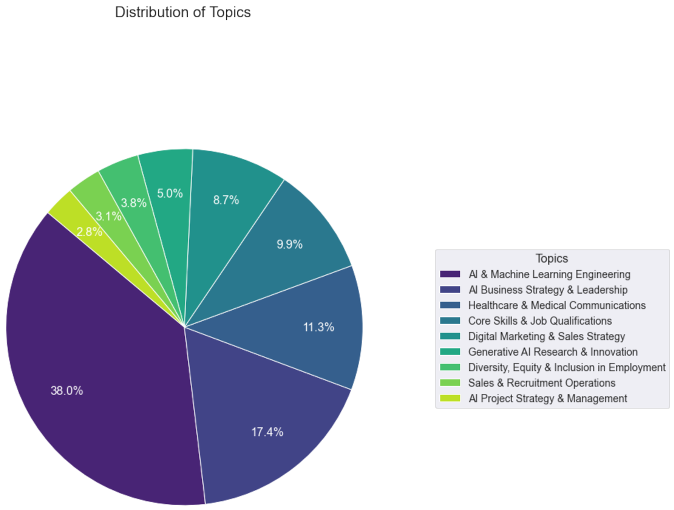


Figure 8: Distribution of LDA clusters

Looking at the statistics (Figure 8), the topic “AI & Machine Learning Engineering” significantly prevails over the others with 38%, followed by “AI Business Strategy & Leadership” with 17.4% and “Healthcare & Medical Communications” with 11.3%.

There is actually no surprise in the results shown in Figure 8. In our current period, where Generative AI is experiencing a moment of great and somewhat uncontrolled expansion, it is expected that engineering roles would dominate. This rapid growth has been judged by many scholars and critics to be a bubble, potentially speculative in nature. For instance, some analysts suggest that the rapid valuation increases in AI companies, driven by hype rather than immediate profitability, mirror past speculative bubbles.

Nevertheless, even if the LDA results are not exactly the same as those obtained during the previous job title analysis, our basic thesis is respected and therefore validated: AI is not used only in companies focused on pure technology innovation and by purely technical figures. It is nowadays required as a skill and competence necessary for different roles, as shown in the pie chart: Digital Marketing, Sales Strategy, Recruitment Operations (which, according to some studies cited in Chapter 2, has been one of the sectors most affected by AI even in past years), Management and Business Strategy.

Examining the growth over time of job ad postings, as shown below in Figure 9, reveals that the more time passes, the more new positions are opened. This demonstrates that although the AI sector has experienced strong growth in the last two and a half years - particularly driven by the spread of GenAI models released by OpenAI, Anthropic, Meta and others - the sector does not currently seem to be experiencing slowdowns or declines. This, therefore, allows us to raise several questions: How much longer can the sector grow at this pace? How sustainable will this growth be? And finally, will it indeed prove to be a bubble ready to burst in the near or distant future? However, addressing these questions is beyond the scope of this study, as it would require different data and methodologies. These remain pertinent avenues for future research.

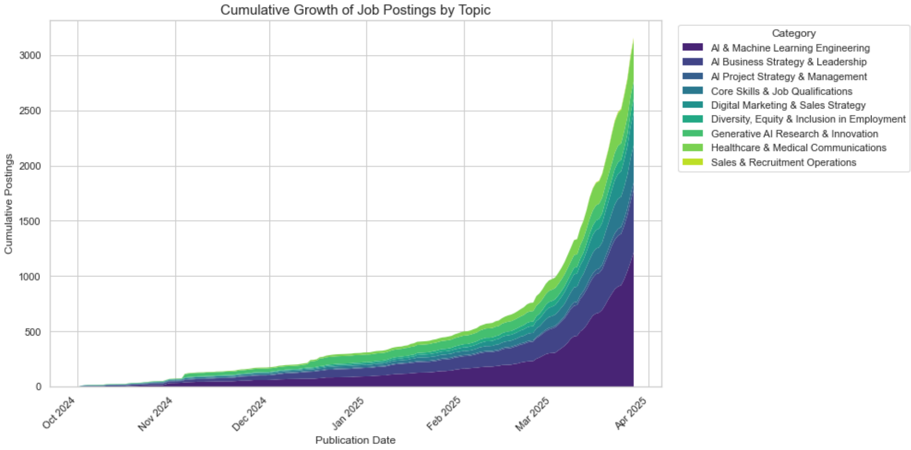


Figure 9: Job posting growth by LDA topic over time

An important consideration when looking at Figure 9 is that, having collected the data over a relatively short period of two weeks and not continuously across the entire timeframe considered, the growth that appears exponential here may, in reality, have been more linear, as previous studies suggest (Ahmadi, 2024). The reason behind this potential difference is that at the moment of data collection, a fraction of the total ads posted, especially the older ones, were likely already closed. This is why, in this specific graph, only posts starting from October 2024 are plotted; although older ads exist (starting from March 2023), the amount of that data is not as relevant with respect to the total, as it comprises few data points spread over a long period.

4.2 Healthcare sector

Following this initial broad thematic analysis of the entire dataset, a decision was made to conduct a more focused investigation into a specific sector. While the LDA revealed several potentially interesting topics, such as Topic 6: "Diversity, Equity & Inclusion in Employment", the relatively small sample size associated with such niche clusters did not permit the extraction of substantial or sufficiently robust insights for a deep-dive analysis. Conversely, other prominent topics, such as "AI & Machine Learning Engineering", despite offering the largest sample among all clusters (38%), presented a different challenge: an in-depth analysis risked reiterating well-established, common-domain knowledge about core engineering roles, which would be redundant and less aligned with the research's aim of uncovering novel, sector-specific applications and nuanced talent demands.

For these reasons and more, among the nine topics identified in the initial LDA (as shown in Section 4.1.2 / Figure 8), the "Healthcare & Medical Communications" topic was selected for a deeper analysis. The primary motivation for this choice lies in the critical, human-centric nature of the healthcare sector: understanding how Generative AI is being adopted within this domain can provide invaluable insights into broader societal impacts, ethical considerations and the specific, often unique, talent requirements emerging from this technological shift. Moreover, its significant representation, accounting for 11.3% of the job advertisements in the initial LDA topic modeling, clearly demonstrates substantial AI integration activity and provides a sufficiently large sub-corpus for more granular investigation, promising richer and more specific findings.

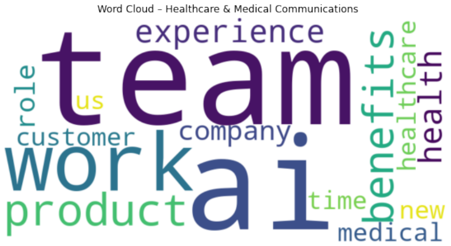


Figure 10: Healthcare & Medical Communications' Wordcloud

4.2.1 Mapping the True Ecosystem of AI-Healthcare Employers

To delve deeper into this interesting sector, the first step was to check the companies that published such job advertisements. The results were largely in line with expectations, while a few others raised suspicions that were subsequently resolved.

As noticeable from Table 2, all the companies align with the healthcare sector, except for two: Netflix and TikTok. Why are these companies listed here? Upon searching the descriptions of these clearly misclassified announcements, the reason for this error was found: in these ads, the companies listed health insurance as a benefit for the candidate, specifying the conditions and advantages, which likely triggered healthcare-related keywords.

Regarding the other companies, to ensure they were genuinely related to this sector and not misclassified like Netflix and TikTok, and also to satisfy the interest of the most curious readers, a brief investigation found the following:

- **Athelas**: A health-tech startup developing portable devices and software for rapid blood diagnostics and chronic disease management.

- **LILT AI**: An enterprise AI company specializing in computer-assisted translation and localization solutions.

- **Commure**: A healthcare interoperability platform provider that helps hospitals and clinics integrate disparate applications into a unified system.

Table 2: Company Counts for AI-Related Job Postings

Company Name	**Count**
Athelas	24
LILT AI	16
Commure	14
Seen Health	10
JPMorganChase	10
Oracle	9
Klarity	8
Abridge	6
Chef Robotics	6
Thomson Reuters	6
LTIMindtree	5

- **Seen Health**: Uses machine learning and NLP to automate clinical documentation and streamline physicians' workflows.

- **JPMorgan Chase**: One of the world's largest banking and financial services firms, active in investment banking, asset management and consumer finance (potentially involved via healthcare financing or investment).

- **Oracle**: An enterprise software giant known for its relational database, cloud infrastructure and business application suites, offering services also to large hospitals and firms involved in healthcare.

- **Klarity**: A startup offering AI-driven contract analytics tools to automate legal document review and clause extraction (relevant for healthcare contracts).

- **Abridge**: A digital assistant for doctors that automatically transcribes and summarizes patient-doctor conversations.

- **Chef Robotics**: Designs and builds robotic arms for commercial kitchens but has expanded into the medical sector in recent years (e.g., for lab automation or assistive robotics).

- **Thomson Reuters**: An international news and information services company providing data, analysis and software for finance, legal and media professionals (with applications in healthcare informatics and regulatory information).

- **LTIMindtree**: An IT consulting and services firm (part of Larsen & Toubro) specializing in digital transformation, software development and system integration for various sectors, including healthcare.

Essentially, having excluded Netflix and TikTok, each of these companies is relevant in the AI-healthcare sector because they represent a diverse ecosystem pushing innovation. Some directly develop AI-powered solutions for clinical diagnostics (Athelas), workflow automation (Seen Health, Abridge) and interoperability (Commure). Others provide crucial enabling technologies like AI translation (LILT AI), data infrastructure (Oracle), contract analysis (Klarity), robotics (Chef Robotics) and IT consulting (LTIMindtree), or support the sector through finance (JPMorgan Chase) and information services (Thomson Reuters). Collectively, they drive efficiency, improve care and expand the technological capabilities within healthcare.

4.2.2 Subcluster Analysis of Generative AI Skill Combinations in Healthcare Roles

To understand which skills related to the Generative AI world are sought in the medical sector, a second LDA was performed at this point. This involved changing the keywords (as specified in Chapter 3) to group occupations into subclusters based on specific skills and abilities related to AI, or required in conjunction with AI, to understand the most requested combinations for the 'roles of the future'. Four different subclusters were obtained, to each of which, specific, related names have been assigned, as shown in Table 3. This table allows us to examine these emerging skill profiles.

Table 3: Main terms associated to topics

End-to-End Healthcare	HealthTech, Marketing & Finance	Compliance & Training	Localization & Sales
patient	client	learning	lilt
healthcare	financial	eligible	sale
commure	care	user	city
care	strategy	gender	hour
end	manager	week	month
provider	creative	legal	account
organization	marketing	—	translation

- **"End-to-end Healthcare Solutions"**: Clearly highlights roles at the core of healthcare delivery. Keywords like "patient", "healthcare", "care" and "provider", along with "end" (suggesting "end-to-end solutions"), point to positions focused on developing or managing comprehensive digital platforms. These roles are crucial for improving the user experience and streamlining workflows for both patients and healthcare organizations like hospitals and clinics.

- **"HealthTech Business & Strategy"**: Shifts focus to the business and strategic side. The combination of "clien", "financial", "strategy" and "marketing" suggests roles in client management, account management or strategic marketing, particularly within HealthTech companies that may have divisions offering financial services (e.g., health insurance, billing) or that require sophisticated marketing for their medical products and services.

- **"Compliance, Training & User Experience"**: Underscores essential support functions within the AI-healthcare ecosystem. Terms such as "learning", "eligible", "user" and "legal" indicate positions related to internal training, e-learning development, ensuring compliance with complex healthcare regulations (like patient eligibility for trials or data privacy laws) and conducting user research to enhance the usability and effectiveness of healthcare solutions.

- **"Specialized Localization & Sales"**: Reveals a very specific and growing niche. The presence of "Lilt" (the company mentioned from previous results), "sale", "account" and "translation" strongly indicates roles in sales and account management for companies providing AI-assisted translation services tailored for the medical sector. This points to the increasing need for localizing medical content, clinical software, and patient communication for global audiences.

Collectively, these subclusters demonstrate that the AI-healthcare job market is dynamic and diverse. As also discovered in previous sections of this chapter, it extends beyond pure AI development to encompass roles crucial for direct patient-provider platform engagement, strategic business growth, regulatory adherence, user-centric design and specialized global communication. This indicates a sector that is not only innovating technologically but also actively building the commercial, operational, and support structures necessary to deploy AI effectively and responsibly in healthcare settings. Moreover, this further investigation was useful to prove that the previously formulated theses are correct, by verifying them directly from the text of the ads descriptions, as well as basing the analysis on predictions made previously, based on the companies that had published the ads. In this way, the correct classifications from our first LDA were confirmed and found to be in line with all the general analysis made in the first part of the chapter.

4.2.3 Investigating the Persistent Demand for Healthcare AI Roles

Now, referring back to Figure 9 in the previous subchapter 4.1.2, an examination of the statistics for the "Healthcare & Medical Communications" topic, reveals a particular trend that caught attention: the number of ads from December 2024 to April 2025 assumed values ranging between 298 and 405, presenting the lowest spread across these four months (excluding topics counting for less than or exactly 5% of the sample).

The interpretations for such behavior - that the demand for positions in the medical sector remains basically constant for a longer period compared to other sectors - can be more than one. Firstly, (1) the demand for AI skills in this sector might be so high that as soon as a position is occupied, a new ad is published, indicating a continuous need for new figures with GenAI competencies. Secondly, (2) for some reasons, the competencies required for certain specific roles in the healthcare sector are not being met, and the positions are difficult to cover and occupy. Since the aim here is not to offer opinions, but rather data-driven insights, the analysis in this area was taken a bit further to provide more reliable answers.

The answer appears to be a mixture of these two hypotheses. Applying different filters based on industry benchmarks, it was found that some positions, such as "Product Manager" (which deals more with management decisions) or "Generative AI Engineer" (for which only computer science competencies are requested), are covered relatively quickly. In contrast, others - in almost all cases, fairly high-level positions such as "Director of Central Operations" (published 2024-03-17), "Senior Data Scientist" and "Engineering Manager, Deep Learning" - have been open for many months, with no candidate yet considered suitable to fill them.

Delving more into the descriptions of such roles also revealed an explanation for this difficulty in finding the right candidate: companies are looking for candidates with a very strong medical profile who, at the same time, possess solid abilities in the tech sector, including skills in Deep Learning, Machine Learning and knowledge of LLMs. Such combinations are commonly difficult to find in people who have spent most of their lives specializing in the healthcare sector. Moreover, to fill high-prestige positions, companies usually prefer someone with many years of experience in the sector. Currently, there are very few experienced doctors with significant ML abilities or computer scientists with extensive experience in the healthcare industry.

4.3 Comparison with previous studies

While the chapter up to this point has detailed the interpretation of this study's primary findings in relation to the initial research questions, it is now pertinent to contextualize our work against existing literature, in order to draw interesting comparisons. A particularly relevant study by Professor Ahmadi, 2024, shared similar initial objectives and foundational analytical steps with the present research. Both studies aimed to understand aspects of the Generative AI job market; however, our analytical pathways diverged after the initial data processing. Professor Ahmadi's work focused on quantifying the required experience levels for GenAI skills, categorizing job advertisements by the "degree of centrality of ChatGPT skill sets", ranging from 'general familiarity' to 'advanced functionalities.'

Although our secondary analysis differs, focusing on some specific work fields, Professor Ahmadi's findings, based on data preceding ours, provide a valuable temporal benchmark. This allows for an examination of how demand for specific GenAI-related job roles may have evolved. The validity of this comparison is supported by key methodological alignments:

- **Data Retrieval Keywords**: Despite different scraping tools, both studies employed a common set of keywords (e.g., 'chatgpt', 'llm', 'genai') for job ad filtering;

- **Data Sources**: Both analyses drew data from the same prominent job platforms, LinkedIn and Indeed;

- **Occupational Classification**: Both studies utilized the O*NET database as the reference for categorizing jobs into occupational families.

These shared foundational elements enable a meaningful comparison of early-stage findings concerning the types of roles in demand, even as our deeper analyses explore different facets of the GenAI labor landscape.

4.3.1 Family Level 1 Change in Structure

Examining the cumulative distribution of job advertisements by O*NET Family Level 1 categories (Figures 11 and 12), a clear evolution in employer priorities emerges.

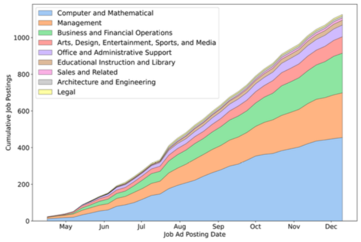


Figure 11: Cumulative Ads growth (2023)

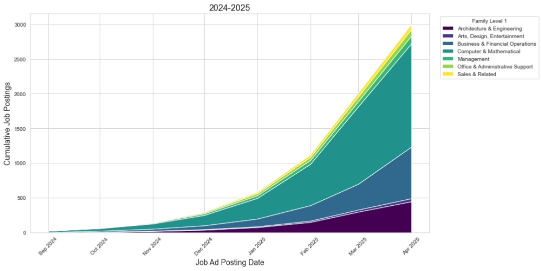


Figure 12: Cumulative Ads growth (2024-2025)

In both 2023 and 2025, two of the top three most requested occupational categories remain consistent: “Computer and Mathematical” roles overwhelmingly dominate, followed by “Business and Financial Operations”. This persistent high demand underscores the foundational importance of technical expertise and business acumen in the burgeoning GenAI field. However, a notable shift occurs in the third top category: between 2023 and 2025, we observe a significant decline in the prominence of "Management" roles, which are largely supplanted by the aforementioned "Business and Financial Operations" category.

This transition likely reflects a maturation in how companies are approaching GenAI integration. In the earlier stages (circa 2023), as GenAI was gaining initial traction, there might have been a greater emphasis on high-level "Management" to oversee the exploration and initial strategic positioning of these new technologies within organizations. Companies were likely focused on understanding GenAI’s potential and setting a general direction. By 2025, as understanding has deepened and practical applications have become clearer, the demand appears to have shifted towards roles within ’Business and Financial Operations’. This suggests a move from strategic oversight to more operational and financial integration, focusing on how GenAI can drive specific business outcomes, optimize processes and create tangible financial value. Companies may now be looking for professionals who can not only manage AI projects but also analyze their ROI, integrate them into existing financial workflows and manage the business risks and opportunities associated with AI adoption. Menaing that, while GenAI’s influence may be broad, the direct hiring focus is becoming more specialized and targeted towards these high-impact domains.

The other four O*NET categories, which were already in the minority in 2023, appear to have become proportionally even less requested by 2025 when compared to the top three. The disparity in demand is quite evident even at a cursory glance, indicating a heightened concentration of demand within the core technical and business-oriented fields.

4.3.2 Evolution of Specific Top Occupations: From Utilization to In-House Development

Delving deeper into specific occupations (O*NET Family Level 3), the comparison between the top 15 roles in 2023 and 2025 (Figures 13 and 14) provides further confirmation and refinement of the trends observed at the category level and in the earlier sections of this chapter.

In 2023, the landscape of in-demand roles was considerably more varied. While technical roles like Data Scientists and Software Developers were present, there was also a significant representation of managerial positions, marketing professionals, writers and even business teachers. This suggests that in the initial wave of GenAI adoption, companies were seeking a broader range of talent capable of understanding, applying and communicating the value of existing GenAI tools (like OpenAI’s ChatGPT) across various business functions.

By 2025, however, the profile of top occupations has become markedly more technical and research-oriented. The list is dominated by various types of Engineers (Machine Learning, software, research, rail), Developers, Researchers, Scientists and Analysts. Managerial roles, while still present (e.g., Director, Manager), are less numerous proportionally.

This pronounced shift from a diverse set of roles in 2023 to a more technically focused array in 2025 signals a critical evolution in corporate GenAI strategy. As highlighted in Professor Ahmadi's 2023 analysis, two years ago, Generative AI was already making a significant impact, but I would add that the primary focus was on leveraging off-the-shelf products developed by a few pioneering companies: businesses sought individuals who could effectively use these external tools. The 2025 data, however, points to a more advanced stage: companies are increasingly aiming to develop their own proprietary GenAI solutions or heavily customized internal tools. This transition from consumption to creation is a pivotal development. It indicates a desire for greater control, customization, data security and the development of unique competitive advantages that cannot be achieved solely by using third-party applications. The demand for specialized engineers, scientists and researchers to build these in-house systems explains the increased technical specialization observed. Companies are no longer just exploring GenAI, they are investing in building foundational capabilities.

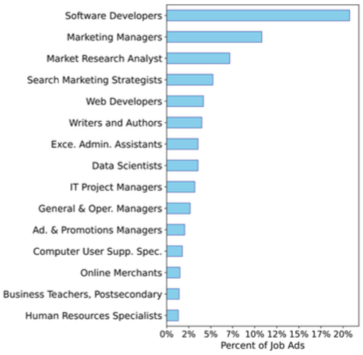


Figure 13: Top O*NET occupation titles (2023)

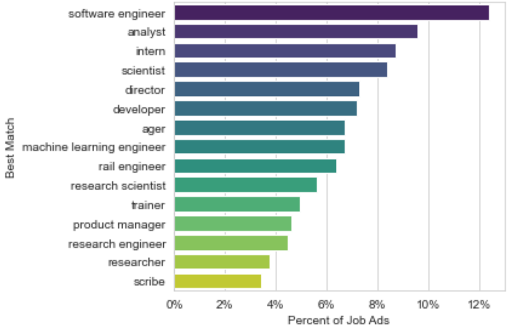


Figure 14: Top O*NET occupation titles (2024-2025)

A particularly noteworthy observation is the emergence and significant ranking of the 'Intern' position, which holds the third spot in 2025 but was completely absent from the top roles in 2023.

This prominent appearance, contrasted with their absence in 2023's top rankings, is a strong indicator of several evolving market dynamics. Firstly, it signifies that the GenAI field is maturing to a point where companies are establishing pipelines for new talent. While in 2023, the focus might have been on acquiring experienced professionals who could immediately contribute to understanding and implementing nascent GenAI technologies, by 2025, with a clearer understanding of foundational GenAI skills and a growing need for a larger talent pool, companies are investing in cultivating talent from the ground up. Secondly, this surge in internships reflects a growing confidence in the long-term viability and importance of GenAI skills: companies are willing to invest in training the next generation. Thirdly, it may also point to a strategy to access fresh perspectives and cutting-edge academic knowledge from students, particularly as the GenAI field continues to evolve rapidly. This proactive approach to talent development was less evident in the earlier, more exploratory phase of 2023.

4.3.3 Contract Type Preferences: Solidifying Full-Time Engagement

A comparison of contract types (Figures 15 and 16) reveals a clear trend towards more stable employment structures.

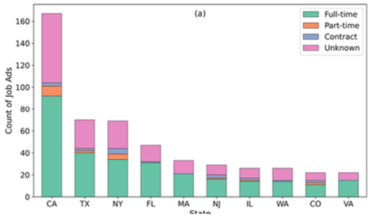


Figure 15: Contract types per Location (2023)

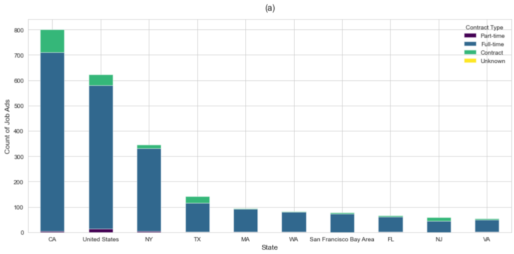


Figure 16: Contract types per Location (2024-2025)

A primary difference is the prevalence of the ’Unknown’ contract type in the 2023 data, suggesting that many job advertisements at that time did not explicitly specify the nature of the contract. However, even when accounting for this, a comparison between full-time and part-time roles across both years confirms that full-time employment remains the strongly preferred and most requested engagement model by companies in the GenAI space. This preference is even more pronounced in 2025, where part-time roles are almost entirely absent from the significant contract types.

The dominance of full-time contracts, especially their increased clarity and prevalence in 2025, suggests that companies view GenAI roles as integral and long-term strategic positions rather than temporary or project-based needs. In 2023, the higher incidence of ’Unknown’ contract types might have reflected some initial uncertainty or flexibility as companies were still defining the scope and permanence of these new roles. By 2025, the near absence of part-time positions and the clear preference for full-time indicate a commitment to building dedicated internal teams with deep expertise in GenAI. This aligns with the earlier observation of a shift towards in-house development, which typically requires sustained, full-time effort from dedicated professionals. Companies are investing in employees they can retain, train and integrate deeply into their operations for sustained GenAI development and deployment.

4.3.4 Salary Dynamics: Escalation and Increased Volatility

An analysis of average annual salaries by O*NET Family Level 1 category reveals significant changes in both compensation levels and distribution between 2023 and 2025.

In 2023, average annual salaries ranged from approximately $50,000 for "Office & Administrative Support" to around $150,000 for "Sales & Related", with core technical roles in "Computer & Mathematical" averaging near $130,000. A key characteristic of the 2023 salary landscape was relatively tight spreads within each occupational family: “Office Support” varied by only ±$7,000, “Business & Financial Operations” by ±$10,000 and “Computer & Mathematical” by about ±$20,000. This uniformity suggests that, at that time, roles and experience levels within each category were fairly homogenous.

By 2025, the salary landscape had transformed dramatically. Almost every category’s mean salary climbed substantially: “Architecture & Engineering” roles now average near $175,000, “Management” around $200,000 and even “Arts, Design, Entertainment, Sports & Media” and “Office & Administrative Support” moved into the mid-$60,000 range. Yet the most striking shift is not merely higher averages but a dramatic expansion in dispersion. “Computer & Mathematical” roles, while averaging close to $225,000, now exhibit an exceptionally wide span—from nearly $0 (driven by unpaid or stipended internships) to over $900,000 for highly specialized senior positions. Likewise, “Business & Financial Operations”, still centered near $55,000 on average, now displays error bars exceeding ±$200,000. In contrast, “Sales & Related” remains around $150,000 but with much tighter variability and both “Management” and “Architecture & Engineering” show moderately widened ranges, reflecting a surge in senior-level and specialist hires.

This explosion in salary volatility - especially within technical and finance families - points to a rapidly evolving market characterized by fierce competition for elite AI talent at one end and robust entry-level hiring at the other. Top-tier AI researchers and architects command premium compensation, sometimes nearing nine-figure salaries, while companies continue to recruit interns and junior staff at comparatively modest rates. Niche specializations, the ongoing maturation of GenAI capabilities and divergent role definitions have all contributed to a fragmented pay structure, where the highest and lowest ends of the spectrum drift ever further apart.

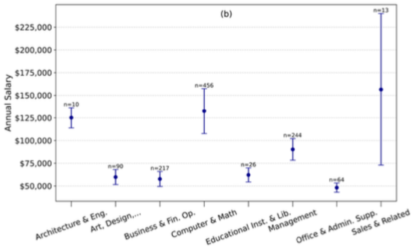


Figure 17: Salary ranges per Family Level 1 (2023)

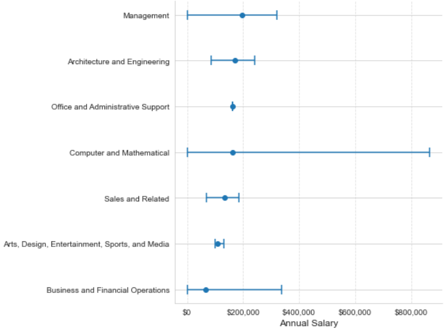


Figure 18: Salary ranges per Family Level 1 (2024-2025)

5 Conclusions

This concluding chapter synthesizes the principal outcomes of this thesis, which aimed to provide a data-driven, empirical analysis of the impact of Generative Artificial Intelligence (GenAI) on the contemporary U.S. labor market. It will begin by summarizing the key findings derived from the analysis of job advertisement data, followed by an acknowledgment of the inherent limitations of this study. Finally, the chapter will explore the broader research implications of the findings and propose potential directions for future investigation in this rapidly evolving field.

5.1 Summary of outcomes

This research systematically investigated the GenAI job market landscape by analyzing 2,726 unique job postings collected between May 2023 and March 2025. The study employed a multi-faceted methodological approach, including job title standardization against the O*NET database, Latent Dirichlet Allocation (LDA) for thematic topic modeling of job descriptions, and comparative analysis with earlier (2023) market data. Key findings from the job title standardization and O*NET family-level analysis revealed:

- A **significant demand for technical roles**, with "Computer and Mathematical" occupations dominating, followed by "Business and Financial Operations".

- A **notable shift from 2023**, with a decline in the relative demand for general 'Management' roles and an increased focus on operational and financial integration of GenAI.

- The **emergence and high ranking of 'Intern' positions** by 2025, indicating a maturing field establishing talent pipelines and investing in early-career professionals.

- A strong **preference for full-time employment** contracts, suggesting a commitment by companies to build stable, long-term GenAI capabilities.

The LDA topic modeling of job descriptions identified nine distinct thematic clusters, with "AI & Machine Learning Engineering" being the most prevalent, followed by "AI Business Strategy & Leadership" and "Healthcare & Medical Communications". This highlighted that while core technical development is paramount, GenAI's application is also significantly influencing business strategy and specific sectors like healthcare.

The **deep-dive analysis into the Healthcare sector** further illustrated this cross-industry reach, identifying sub-clusters such as "End-to-end Healthcare Solutions", "HealthTech Business & Strategy", "Compliance, Training & User Experience" and "Specialized Localization & Sales". This showcased the diverse applications of GenAI within healthcare, extending beyond pure technical development to include patient engagement, strategic growth, regulatory adherence, and global communication. This sector also highlighted challenges in finding candidates with a dual proficiency in deep medical knowledge and advanced AI skills, particularly for senior roles.

The **comparative analysis with 2023 data** (based on Ahmadi, 2024) within Chapter 4.3, underscored a critical evolution:

- A **shift in** top occupations from a broader utilization of existing GenAI tools (e.g., ChatGPT) in 2023 towards a more technically specialized **demand for in-house**

development of proprietary GenAI solutions by 2025.

- A **significant increase in average salaries** across almost all AI-related job categories by 2025, coupled with a dramatic expansion in salary dispersion, especially within "Computer & Mathematical" and "Business & Financial Operations" roles. This reflects intense competition for elite talent alongside growth in entry-level opportunities.

Overall, the findings paint a picture of a GenAI labor market that is rapidly evolving, characterized by increasing technical specialization, a drive towards in-house capability building, significant cross-industry adoption (exemplified by healthcare) and dynamic shifts in talent demand, experience requirements and compensation structures.

5.2 Limitations of the study

While this thesis provides valuable insights into the GenAI labor market, several limitations must be acknowledged:

- Data source specificity: The primary data was sourced from LinkedIn and Indeed job postings. While these are major platforms, they may not capture the entirety of the job market, potentially missing jobs advertised through other channels (e.g., company-specific career pages, niche job boards, recruitment agencies).

- Geographical focus: The data collection was confined to the United States. Findings may not be directly generalizable to other countries with different economic structures, technological adoption rates, or labor market regulations.

- Snapshot in time: Although the data spans nearly two years, the GenAI field is exceptionally dynamic. The findings represent a snapshot during this period and may evolve rapidly. The data collection period of two weeks for the cumulative growth analysis (as mentioned in Chapter 4) also introduces a limitation in perfectly capturing linear vs. exponential growth over longer, continuous periods.

- Reliance on advertised information: The analysis is based on the information provided in job advertisements. This may not always perfectly reflect the actual day-to-day responsibilities, true skill requirements or final compensation packages. There can be a discrepancy between advertised roles and filled positions.

- Interpretive nature of LDA: While coherence and perplexity scores, along with manual validation, were used to select the optimal number of topics in LDA, the interpretation and naming of these topics retain an element of subjectivity.

- Sample size for niche topics: As noted, some identified LDA topics (e.g., "Diversity, Equity & Inclusion in Employment") had smaller sample sizes, limiting the depth of analysis possible for those specific areas.

5.3 Research implication and future directions

This thesis reveals important insights into the evolving Generative AI (GenAI) labor market, with direct consequences for education, policy and future scholarship.

The findings first underscore that **education and training must adapt**. We observed a growing demand for specialized technical AI roles, such as engineers and scientists, alongside strong business skills, often cultivated in-house. This necessitates that educational institutions and training programs develop curricula blending deep AI expertise with business acumen and industry-specific knowledge. Secondly, this study **informs our understanding of labor market transformation** by providing concrete data on how GenAI is actively reshaping jobs, skill demands, and potentially wages, offering empirical evidence beyond speculation. Furthermore, these insights can **guide policy decisions**; our findings on in-demand jobs, critical skills, and talent acquisition challenges, like sourcing hybrid medical-AI experts, can help policymakers design effective workforce development strategies, targeted reskilling programs, and responsible AI integration policies. Finally, the analysis of the GenAI labor market from 2023-2025 **establishes a valuable benchmark**, enabling future research to track trends and measure the ongoing evolution of this dynamic field.

To build on these findings, several promising avenues for future research emerge. **Long-term monitoring** of the GenAI job market is crucial to observe the sustainability of current trends, the maturation of new roles, and potential market shifts. Additionally, **qualitative deep dives**, through interviews with hiring managers and AI professionals, would offer richer context on specific skill needs and hiring challenges. Future work should also investigate the **impact on specific job roles and tasks**, precisely how GenAI is transforming individual occupations, distinguishing between task augmentation and potential displacement. A focused **skill gap analysis** is also needed to quantify the disparity between demand for specific GenAI skills, especially in areas like healthcare-AI, and the available workforce supply. The **ethical and societal dimensions** warrant further examination, including the growth of roles in AI ethics and governance, with a dedicated focus on Diversity, Equity, and Inclusion. Lastly, **global comparative analysis**, expanding research to other major economies, would provide a more comprehensive understanding of regional differences in GenAI's labor market impact.

References

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of labor economics* (pp. 1043–1171, Vol. 4B). Elsevier.

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, *40*(S1), S293–S340. https://doi.org/10.1086/720104

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, *33*(2), 3–30. https://doi.org/10.1257/jep.33.2.3

Ahmadi, M. (2024). Generative ai impact on labor market: Analyzing chatgpt's demand in job advertisements [Available at arXiv:2412.07042].

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. https://doi.org/10.1257/jep.29.3.3

Aydın, Ö., & Karaarslan, E. (2023). Is chatgpt leading generative ai? what is beyond expectations? *Academic Platform Journal of Engineering and Smart Systems*, *11*(3), 118–134. https://doi.org/10.21541/apjess.1293702

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623.

Bi, D., & Xiao. (2023). *Journal*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brauner, J. M., et al. (2024). Managing extreme ai risks amid rapid progress. *Science*, *384*(6698), eadn0117. https://doi.org/10.1126/science.adn0117

Brynjolfsson, E., Li, D., & Raymond, L. (2023, April). *Generative ai at work* (tech. rep. No. Working Paper 31161). National Bureau of Economic Research.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W.W. Norton & Company.

Brynjolfsson, E., Mitchell, T., & Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, *108*, 43–47. https://doi.org/10.1257/pandp.20181000

Brynjolfsson, E., Rock, D., & Syverson, C. (2019). *The productivity j-curve: How intangibles complement general purpose technologies* (tech. rep. No. Working Paper 25148) (Revised Jan 2020). National Bureau of Economic Research.

Burrus, J., Jackson, T., Xi, N., & Steinberg, J. (2013). *Identifying the most important 21st century workforce competencies: An analysis of the occupational information network (o*net)* (Research Report Series No. 2013(2)). ETS.

Caner, S., & Bhatti, F. (2020). A conceptual framework on defining business strategy for artificial intelligence. *Contemporary Management Research*, *16*(3), 175–206. https://doi.org/10.7903/cmr.19970

Choudhury, P., Foroughi, C., & Larson, B. (2020). *Work-from-anywhere: The productivity effects of geographic flexibility* (Working Paper). Harvard Business School.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

De Mauro, A., Greco, M., & Grimaldi, M. (2018). A formal definition of big data based on its essential features. *Library Review*, *65*(3), 122–135. https://doi.org/10.1108/LR-06-2017-0090

Deming, D. J. (2017). The growing importance of social skills in the labor market. *Quarterly Journal of Economics*, *132*(4), 1593–1640. https://doi.org/10.1093/qje/qjx022

Ellingrud, K., Manyika, J., Sneader, K., et al. (2023, July). *Generative ai and the future of work in america* (tech. rep.). McKinsey Global Institute.

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* (tech. rep.). OpenAI.

European Commission. (2021, March). *2030 digital compass: The european way for the digital decade* (Communication COM(2021) 118 final). European Commission.

European Commission & U.S. Department of Labor. (2022). *Esco–o*net crosswalk: A joint mapping of european skills, competences, and occupations to u.s. soc* (Report). European Commission Joint Research Centre.

Felten, E., Raj, M., & Seamans, R. (2018). A method to link advances in artificial intelligence to occupational abilities. *AEA Papers and Proceedings*, *108*, 54–57. https://doi.org/10.1257/pandp.20181033

Felten, E., Raj, M., & Seamans, R. (2023). Occupational heterogeneity in exposure to ai and implications for labor market outcomes.

Frank, M. R., Autor, D., Bessen, J., et al. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, *116*(14), 6531–6539. https://doi.org/10.1073/pnas.1900949116

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerization? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

George, A. S. (2024). Artificial intelligence and the future of work: Job shifting not job loss. *Partners Universal Innovative Research Publication*, *2*(2), 17–37. https://doi.org/10.5281/zenodo.10936490

Goindani, M., Liu, Q., Chao, J., & Jijkoun, V. (2017). Employer industry classification using job postings. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 183–188. https://doi.org/10.1109/ICDMW.2017.30

Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*, 2672–2680.

Hardy, W., Keister, R., & Lewandowski, P. (2018). *Computerization and the labor market in developing economies* (tech. rep. No. Research Paper No. 17). International Labour Organization.

Hatzius, J. (2023, March). *The potentially large effects of artificial intelligence on economic growth* (tech. rep.). Goldman Sachs.

Hui, X., Reshef, O., & Zhou, L. (2024). The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. *Organization Science*. https://doi.org/10.1287/orsc.2023.18441

Javed, F., Luo, Q., McNair, M., Jacob, F., Zhao, M., & Kang, T. S. (2015). Carotene: A job title classification system for the online recruitment domain. *2015 IEEE First International Conference on Big Data Computing Service and Applications (BigDataService)*, 286–293. https://doi.org/10.1109/BigDataService.2015.61

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Kitsios, F., & Kamariotou, M. (2021). Artificial intelligence and business strategy towards digital transformation: A research agenda. *Sustainability*, *13*(16), 8485. https://doi.org/10.3390/su13168485

Kortum, H., Rebstadt, J., & Thomas, O. (2022). Dissection of ai job advertisements: A text mining-based analysis of employee skills in the disciplines computer vision and natural language processing. *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS)*, 5211–5220. https://doi.org/10.24251/HICSS.2022.635

Liu, J., Xu, X. (, Li, Y., & Tan, Y. (2023). "Generate" the future of work through ai: Empirical evidence from online labor markets. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4529739

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 262–272.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

Nasser, I. M., & Alzaanin, A. H. (2020). Machine learning and job posting classification: A comparative study. *International Journal of Engineering and Information Systems*, *4*(9), 6–14.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, *33*(1), 31–88. https://doi.org/10.1145/375360.375365

Noy, S., & Zhang, W. (2023, March). *Experimental evidence on the productivity effects of generative artificial intelligence* (tech. rep.) (Working Paper). MIT Department of Economics.

OpenAI. (2023). *Gpt-4 technical report* (tech. rep.). OpenAI.

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of ai on developer productivity: Evidence from github copilot.

Peterson, N. G., Mumford, M. D., Borman, W. C., et al. (2001). Understanding work using the occupational information network (o*net): Implications for practice and research. *Personnel Psychology*, *54*(2), 451–492. https://doi.org/10.1111/j.1744-6570.2001.tb00062.x

Rahhal, I., Carley, K. M., Kassou, I., & Ghogho, M. (2023). Two stage job title identification system for online job advertisements. *IEEE Access*, *11*, 19073–19092. https://doi.org/10.1109/ACCESS.2023.3247866

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Squicciarini, M., & Nachtigall, D. (2021). *Occupations at risk: Automatability and its earnings and employment effects* (Science, Technology and Industry Working Paper No. 2021/01). OECD.

Sturm, B., et al. (2021). Artificial intelligence and music: Open questions of copyright law and policy [Please verify title/pages]. *Creative Industries Journal*, *14*(2), 130–146.

Tippins, N., & Hilton, M. (Eds.). (2010). *A database for a changing economy: Review of the occupational information network (o*net)*. National Academies Press.

Tyson, L., & Zysman, J. (2022, January). *Beyond automation anxiety: Building a future of shared prosperity* (tech. rep.). Berkeley Roundtable on the International Economy.

Upadhyay, A., et al. (2021). A review of ai adoption in industry: Critical challenges and future research directions. *Journal of Manufacturing Systems*, *60*, 828–841. https://doi.org/10.1016/j.jmsy.2021.02.006

World Economic Forum. (2023). *The future of jobs report 2023* (tech. rep.). World Economic Forum. https://www.weforum.org/publications/the-future-of-jobs-report-2023/

Xu, W., & Dainoff, M. J. (2021). Leveraging ai for ergonomic risk assessment: Opportunities and challenges. *IISE Transactions on Occupational Ergonomics and Human Factors*, *9*(3-4), 213–223.

Zhao, M., Javed, F., Jacob, F., & McNair, M. (2015). SKILL: A system for skill identification and normalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, *29*(2), 4012–4017. https://doi.org/10.1609/aaai.v29i2.19064