Granular AI Risk Modelling from Tasks to Occupations And Workforce Implications

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy
in
Analytics

by

Dawei Xu

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW, 2007, Australia

September 2025

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Dawei Xu* declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science, Faculty of Engineering and Information Technology* at the University of Technology Sydney.

The thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship doi.org/10.82133/C42F-K220.

Production Note:
Signature removed prior to publication.

SIGNATURE: ________________

Dawei Xu

DATE: 1st September, 2025

PLACE: Sydney, Australia

ABSTRACT

The rise of artificial intelligence (AI) is reshaping the nature of work at an unprecedented pace. Understanding this transformation is critical, yet most existing assessments rely on coarse occupation-level measures or static proxies of technological risk, failing to capture the evolving and heterogeneous relationships between technical progress and human labor. These limitations create blind spots in identifying which tasks and occupations are most exposed to change at a granular level. To address this gap, this thesis develops a dynamic, multi-layered framework for modeling AI risk across tasks, occupations, and career trajectories by integrating fine-grained job task data, longitudinal measures of AI performance, and advanced modeling techniques including graph neural networks (GNNs) and large language models (LLMs).

Specifically, the thesis first constructs a graph-based occupation–skill network, leveraging GNNs to model how machine risk diffuses across interdependent jobs. Second, it applies LLM-driven classification to over 13,000 job tasks, distinguishing substitution, complementarity, and negligible effects, thereby uncovering substantial within-occupation heterogeneity. Third, it introduces an ontology-anchored AI Exposure Index that aligns human tasks with a curated, benchmark-centered knowledge graph of AI capabilities, enhanced with momentum-weighted measures of technical progress and research attention. Finally, the thesis proposes an interpretable career-mobility model as a foundational extension of its overarching objective to examine AI risk at increasing levels of granularity, linking exposure patterns to individual job transitions and longer-term career development.

Through comprehensive experiments on large-scale labor datasets, the thesis demonstrates that AI-related risk is both more pervasive and uneven than previously recognized, with significant implications for individual career trajectories, workforce development, and public policy. Overall, the findings advance measurement of AI's labor-market impact by bridging job content, technological dynamics, and mobility processes, providing a foundation for task-aware and capability-aligned labor analytics in an era of rapid AI innovation.

DEDICATION

To myself for my perseverance, to my family for their supports and to all who truly helped me, for their unwavering encouragement. . .

Acknowledgments

My PhD journey is never what it first appears. I began with naïve confidence, blind to the toughness, the torment, the countless nights of doubt. Yet as I emerged from the storm, I discovered its hidden rewards. It has taught me not only knowledge and skills, but also something deeper: to stay calm when the pressure builds, to keep moving when the weight feels unbearable. It has been a mirror, forcing me to think more about who I am, what I own, what I am good at, and what kind of world I stand in. These lessons, painful and precious, now anchor my future—both in career and in life.

To my parents and my family, I owe the deepest gratitude. They may not understand the meaning of research, nor the strange language of academia, yet their trust and support has never wavered which has been the strongest foundation beneath my feet.

I am grateful to my principal supervisor, Professor Guandong Xu, who opened the door of research to me. He guided me into this field and, more importantly, reminded me of the value and spirit of research. These are not small things: they shaped my direction, corrected me when I went astray, and offered steady support along the way. I also extend my gratitude to my co-supervisor, Professor Marian-Andrei Rizoiu, whose showed me the importance of asking the right research questions—questions that matter, questions that hold value.

My gratitude also extends to external scholars whose wisdom and kindness have shaped this journey. Dr. Jason Chen offered not only sharp academic guidance but also care that reached beyond the boundaries of research. He has been a life friend throughout this journey, someone I could turn to when faced with both academic challenges and personal difficulties. I am equally indebted to Professor Xitong Li, Dr. Jingyuan Yang, and Dr. Hao Zhong, whose support was both generous and steadfast. From the very first draft to the final polished paper, they walked beside me with patience and insight.

This journey would not have been possible without my teammates. Dr. Haoran Yang offered valuable academic advice. Dr. En Yu shared with me the discipline of workouts, a different but vital form of endurance. Above all, I thank Dr. Xiangmeng Wang, whose steady presence and support have carried me through both academic challenges and everyday life during this journey. Beyond academia, I am grateful to my friends: those in Australia who helped me explore Sydney, those in China who encouraged me in the darkest moments, and those in the United States who gave me bright memories—sparks that restored my confidence and urged me onward.

Finally, I wish to thank everyone who has been part of this PhD journey. Each form of support, whether visible or quiet, has helped me reach this moment. This thesis may carry my name, but it carries within it the traces of all of you.

List of Publications

Related to the Thesis

1. D. Xu, H. Yang, M.-A. Rizoiu, and G. Xu, "Being automated or not? risk identification of occupations with graph neural networks," in International Conference on Advanced Data Mining and Applications, Springer, 2022, pp. 520–534.

2. D. Xu, H. Yang, M.-A. Rizoiu, and G. Xu, "From occupations to tasks: A new perspective on automatability prediction using bert," in 2023 10th International Conference on Behavioural and Social Computing (BESC), 2023, pp. 1–7.

3. D. Xu, J. Li, M.-A. Rizoiu, and G. Xu. "An Interpretable Measure of AI Exposure: Linking Workforce Tasks to Dynamic AI Capabilities," 2025. INFORMS Journal on Computing. - Submitted & Under Review.

4. D. Xu, J. Yang, H. Zhong, X. Li, G. Xu. "Who Thrives, Who Stays? Predicting Talent Potential from Level-Aware Career Moves." 2025. - Finished & To be submitted

Table of Contents

List of Figures

Figure **Page**

List of Tables

Chapter 1

Introduction

This chapter begins by examining the impact of AI on the job market, focusing on job displacement and creation, the demand for new skills, and evolving wage dynamics. It then underscores the significance of investigating these effects and reviews the current research in the field. Subsequently, the chapter identifies existing research gaps and outlines the contributions of this thesis. Finally, an overview of the thesis structure is provided.

1.1 Research Background

1.1.1 Review of the AI-Driven Job Market

Artificial intelligence (AI) has moved rapidly from a speculative technological prospect to a tangible force reshaping how work is organised, performed, and valued across the global job market. Its influence is no longer limited to isolated pilot applications or laboratory demonstrations; instead, AI systems are increasingly embedded in everyday production processes, decision-making workflows, and service delivery across a wide range of job tasks and occupations.

A growing body of surveys and academic research has examined these dynamics from multiple perspectives, highlighting AI's heterogeneous effects on employment, productivity, wages, and skills. On the employment side, projections suggest substantial job churn rather than uniform displacement: AI may displace 85 million jobs globally by 2025, it also anticipates the creation of 97 million new roles, yielding a net increase driven by growing demand for AI-, data-, and robotics-related occupations, alongside heightened vulnerability in routine and administrative work (World Economic Forum, 2025). Empirical estimates of automation risk vary widely, with occupation-level studies suggesting that nearly half of U.S. jobs may be exposed to automation (Frey & Osborne, 2013), and comparable risks identified in Finland, Germany, and across Europe (Bowles, 2014; Brzeski & Burk, 2015; Pajarinen et al., 2015), while task-based analyses using job-level data produce substantially lower estimates—around 9%—though potentially understating exposure by obscuring shared task structures across occupations (Arntz et al., 2017). Beyond employment quantities, AI adoption has been consistently linked to productivity gains, with AI-intensive industries exhibiting higher revenue per worker than less exposed sectors (PwC, 2025). However, its implications for wage distribution remain contested. Some studies associate AI with rising wage inequality, disproportionately benefiting highly skilled and strategically complementary roles (International Monetary Fund, 2025; Jaccoud, 2025), whereas others argue that the automation of routine and selected high-skilled tasks may increase the relative value of distinctly human capabilities—such as creativity, critical thinking, and interpersonal skills—thereby narrowing wage disparities (OECD, 2024a; Santarelli et al., 2025). Complementing these structural shifts, a substantial literature documents rapid changes in workforce skill requirements. Skills in AI-exposed occupations are evolving at an accelerated pace, particularly in roles with higher automation potential (PwC, 2025), with job-posting analyses revealing growing demand for AI-related technical competencies alongside adaptable, collaborative, and problem-solving skills (Mäkelä & Stephany, 2024; Stephany & Teutloff, 2022). At the same time, routine task proficiency is declining in value, while creative, cognitive, and interpersonal skills command increasing premiums (Brynjolfsson & Mitchell, 2017; Das et al., 2020; Stephany & Teutloff, 2022). Although AI-specific occupations remain a minority, their share is expanding, and skill requirements within AI-intensive sectors are becoming more diversified (OECD, 2024b). Looking ahead, surveys project that up to 40% of core skills will change by 2030, reinforcing a shift toward skills-based hiring, alternative credentials, and continuous learning as central features of adaptation in the AI-driven labor market (UNESCO-UNEVOC, 2021; World Economic Forum, 2025).

Beyond aggregate projections and cross-sectional statistics, a growing empirical literature has manifested the concrete illustrations of how AI is transforming the labor market settings across a wide range of sectors. In healthcare, AI systems are increasingly embedded in high-skill clinical workflows, with task-based analyses showing that applications such as medical image interpretation and report drafting could reduce radiologists' workloads by up to one-third without generating near-term job losses, instead reallocating time toward complex cases, patient interaction, and clinical judgment, while raising demand for skills related to AI oversight and interpretation (Langlotz, 2025; Osman et al., 2025). Customer service provides one of the clearest contemporary examples of task-level augmentation, as generative AI assistants substantially increase agent productivity—particularly for less experienced workers—by disseminating best practices in real time, compressing skill differences, and shifting human effort away from routine information retrieval toward exception handling and interpersonal interaction (Brynjolfsson et al., 2025). Similar patterns of augmentation are observed in education, where AI tools increasingly automate lesson preparation, grading, and administrative tasks, allowing teachers to reallocate time toward instructional design and student engagement, while simultaneously shifting skill requirements toward prompt design, evaluation of AI-generated content, and hybrid pedagogical competencies (Milano et al., 2023; Peláez-Sánchez et al., 2024). In manufacturing and logistics, earlier waves of automation demonstrated the displacement potential of robotics for routine manual work, yet more recent AI deployments—such as predictive maintenance, computer vision–based quality control, and warehouse robotics—have primarily driven gradual workforce reorganisation through attrition, retraining, and task redistribution, with productivity gains outpacing direct employment reductions (Brynjolfsson et al., 2025; Y. Li et al., 2024). Finally, in software development, AI coding assistants have dramatically accelerated task completion and reduced barriers for novice programmers, while early labour market evidence suggests a contraction in entry-level hiring alongside growth in senior roles, consistent with the automation of routine coding tasks and the rising importance of system design, AI orchestration, and validation skills (Peng et al., 2023).

Taken together, this body of evidence indicates that AI's impact on the labor market is neither uniform nor adequately captured by simplistic or overly aggregated measurement frameworks. Instead, AI reshapes work through selective automation and augmentation of tasks, unevenly redistributing productivity gains, skill requirements, and employment opportunities both within and across occupations. These patterns underscore the need for analytical frameworks that move beyond aggregate indicators to examine how AI capabilities interact with the specific tasks that constitute jobs, motivating a more granular, task-centered approach to measuring AI exposure and its labor market implications.

1.1.2 Understanding and Measuring AI's Impact on the Job Market

Understanding the impact of artificial intelligence (AI) on the job market is increasingly critical for policymakers, businesses, and workers alike. A robust evidence base is essential for anticipating employment shifts, enabling the design of effective retraining initiatives and social protection policies to mitigate displacement risks and support smooth workforce transitions (Acemoglu & Restrepo, 2019; Arntz et al., 2016; Lane & Saint-Martin, 2021). Without informed intervention, inadequacies in existing social safety nets risk amplifying the negative consequences of AI-driven labor disruptions (Brynjolfsson & McAfee, 2014; Webb, 2019). For workers, understanding AI's influence highlights the urgent need to acquire new, in-demand skills to stay relevant in rapidly evolving job landscapes (Felten et al., 2019; Frey & Osborne, 2013). At the societal level, addressing the ethical and distributive implications of AI is crucial to ensure the broad sharing of productivity gains, preventing exacerbation of existing inequalities and supporting inclusive economic growth (Acemoglu, 2021; Chui et al., 2016; Jarrahi, 2018).

Given the complex and multidimensional nature of AI's impact on employment, researchers have developed a diverse range of methodologies to measure and investigate its effects. These include task-based approaches—such as patent-task linking, suitability for machine learning (SML), and language model-driven assessments—which analyze granular task-level exposure using resources like ONET, USPTO patents, or models such as GPT-4 (Brynjolfsson et al., 2018; Eloundou et al., 2023; Webb, 2019). Ability-based methods, exemplified by AI Occupational Exposure (AIOE), map AI capabilities to occupational skills, leveraging datasets like ONET and PIAAC to provide cross-national insights, though often limited by data granularity (Acemoglu et al., 2022; Felten et al., 2019). Survey-based methods capture worker and employer perspectives on displacement risks and AI adoption, providing subjective yet valuable insight into organizational readiness and workforce sentiment (Bessen et al., 2019; Chui et al., 2016). Econometric models, utilizing firm-level and cross-country datasets such as PATSTAT, ORBIS, and EU-LFS, quantify economic outcomes and structural labor market changes but may overlook occupational heterogeneity (Acemoglu & Restrepo, 2017; Lane & Saint-Martin, 2021). Lastly, scenario-based case studies—produced by entities like the U.S. Bureau of Labor Statistics and the Tony Blair Institute—employ historical analysis and future projections to inform practical policy debates, despite the inherent uncertainty in forecasting (Manyika et al., 2017). This diverse methodological landscape underscores the necessity for a holistic synthesis of evidence to guide stakeholders through the ongoing transformation of work in the AI era.

1.2 Research Scopes

1.2.1 Positioning This Research

A growing body of evidence supports the move from occupation-level assessments to more granular, task-based frameworks. The central idea, first introduced by D. Autor et al. (2003) and further developed by Acemoglu and Autor (2011), is that jobs are best understood as bundles of tasks—specific activities requiring various combinations of skills, knowledge, and abilities. This "task-approach" recognizes that technological change, including AI, rarely automates or augments entire jobs at once; rather, it substitutes or complements discrete tasks within jobs. As a result, job transformation is inherently uneven: within the same occupation, some workers may see their core tasks automated while others shift to new, complementary responsibilities (Acemoglu & Autor, 2011; D. Autor & Handel, 2009; D. Autor et al., 2003). Empirical studies now confirm that significant variation exists in the automability of tasks even within standardized occupations (e.g., managers in different firms or doctors in different settings) (Eloundou et al., 2023; Felten et al., 2019).

Adopting a task-based approach enables a more nuanced analysis of AI's potential risks and opportunities. By decomposing occupations into their constituent tasks, and mapping these to AI capabilities (through benchmarks, patent analysis, or expert rubrics), researchers can more precisely estimate which segments of the workforce are exposed to automation risk or augmentation potential. For example, recent research has leveraged detailed O*NET task statements and AI benchmark progress to construct occupation- and task-level "exposure" scores, highlighting that higher-income and white-collar jobs are increasingly affected by advances in general-purpose technologies like large language models (Eloundou et al., 2023; Felten et al., 2019; Martínez-Plumed et al., 2020; Tolan et al., 2021). Other studies have shown that focusing on task content allows for the identification of abilities (such as reasoning, perception, or manual dexterity) most likely to be influenced by AI, as well as areas where human comparative advantage is likely to persist (Brynjolfsson et al., 2018; Frey & Osborne, 2013).

However, the task-approach is not without limitations. It requires high-quality data on the actual tasks performed across diverse work settings and robust methods for mapping tasks to AI capabilities, which may involve expert judgment or machine learning models whose predictions are sensitive to changing AI benchmarks and real-world deployment. Further, task-based analyses may underestimate the importance of context—such as organizational change, demand elasticity, and institutional factors—that shape the realized impact of AI on employment (Frank et al., 2019; Moro et al., 2021).

In this research, we advance the study of AI's impact on the workforce through a suite of complementary approaches that capture the multi-layered structure of work and career dynamics. First, I model the occupation–skill space as a graph, uncovering how automation risk propagates through shared capabilities and revealing network-level spillovers often missed by occupation-centric measures. Second, I harness advanced language models to parse the semantic content of job tasks, enabling a fine-grained taxonomy that distinguishes between substitution, complementarity, and negligible exposure. Third, I connect the labor market perspective with the technological frontier by systematically aligning job tasks with AI task ontologies and benchmarks, creating a scalable and transparent framework for quantifying exposure to evolving AI capabilities. Finally, we extend the analysis to the individual level by introducing a interpretable career mobility framework that models career trajectories and stability, laying the groundwork for linking AI exposure with individual career outcomes. Together, these contributions move beyond static, occupation-level forecasts and offer a comprehensive, multi-level framework for understanding how AI reshapes tasks, occupations, and individual trajectories, equipping policymakers, educators, and workers with more actionable insights for adaptation in an AI-driven world.

1.2.2 Identified Research Gaps

After an exhaustive review of the literature presented in Chapter 2, I identified four persistent research gaps in how scholars model the labour-market risks posed by AI. Each gap is directly addressed by the following empirical methods that comprise this thesis.

Gap 1: Network-aware spillovers. Most existing indices treat every occupation as an isolated unit, implicitly assuming that risk cannot *spill over* to neighbouring jobs that share the same skills or tasks (Acemoglu & Restrepo, 2017; Frank et al., 2019). This assumption is untenable: once an AI system automates a critical task in one occupation, any other job relying on that task faces an elevated threat. My first paper fill this gap by constructing a bipartite occupation–skill graph from O*NET, where edges represent required skills. A Graph Convolutional Network (GCN) diffuses sparse "automatable / non-automatable" labels through these shared-skill edges, quantifying how vulnerability *cascades* through the network. The resulting risk frontier uncovers clusters of occupations—often invisible in occupation-centric taxonomies—that warrant joint reskilling and policy attention.

Gap 2: Fine-grained semantic understanding of tasks. Prevailing measures map occupations to broad routine–versus–cognitive bins without *reading* the textual content of job tasks (Arntz et al., 2016; Brynjolfsson & Mitchell, 2017; Frey & Osborne, 2013; Tolan et al., 2021). This coarse mapping overlooks crucial distinctions—for example, "*write a creative story*" versus "*write a technical report*." In my second paper, I fine-tune a BERT model on 13 000 O*NET task statements, achieving sentence-level semantic parsing of each task's *action* and *outcome*. This approach exposes within-occupation heterogeneity, reveals complementary versus substitutable subtasks, and supplies a labelled dataset that future researchers can reuse to probe evolving AI capabilities.

Gap 3: Task-based alignment between AI capabilities and job tasks. Much of the existing literature relies on static, proxy-based measures—such as patent overlap, expert assessments, abstract abilities, or one-off benchmark scores—to approximate AI exposure (Felten et al., 2019; Frey & Osborne, 2013; Tolan et al., 2021; Webb, 2019). These approaches treat technological change as a uniform or time-invariant shock and rarely establish a direct, transparent link between concrete AI capabilities and the specific job tasks they may affect. As a result, they obscure both the heterogeneous structure of work and the rapidly evolving trajectory of AI performance, limiting interpretability and policy relevance. To address this gap, my third paper introduces a task-based model that directly aligns O*NET job tasks with concrete AI tasks in the ITO knowledge graph (Blagec et al., 2022), producing a forward-looking, interpretable measure of task-level AI exposure that captures the evolving trajectory of AI capabilities beyond static proxies.

Gap 4: Individual-level AI exposure assessment. Despite growing interest in task- and occupation-level exposure, no existing framework systematically links AI risk to the *individual* level. Current indices implicitly assume that all workers within an occupation face identical exposure, overlooking heterogeneity in career histories, mobility opportunities, and talent potential. Yet in reality, individuals differ markedly in how they traverse organizations, accumulate skills, and adapt to shifting technological frontiers. My fourth paper begins to address this omission by introducing LADDER-GNN, a level-aware career trajectory model that predicts *Ability to Thrive* and *Willingness to Stay*. This provides the foundation for integrating individual-level career dynamics with task- and occupation-level AI exposure scores.

1.3 Research Questions

Building on the gaps identified in Section 1.2.2, the thesis is guided by the following research questions:

RQ1 How can the rich *contextual* and *semantic* information harnessed to forecast occupational risk to AI with greater accuracy than traditional baselines?

RQ2 How can the rapidly evolving spectrum of AI capabilities be mapped onto thousands of human job tasks in a way that is both *scalable* (minimal manual labelling) and *transparent* (benchmark-level traceability)?

RQ3 How to build a structured framework that quantifies AI exposure from individual tasks to entire occupations while incorporating the evolving dynamics of AI capabilities?

RQ4 What novel insights into task-, occupation- and sector-level vulnerability emerge when these network-aware, capability-weighted indices are introduced with established exposure measures?

RQ5 How can task and occupation level AI exposure be connected to individual career trajectories to assess workforce resilience at the micro level?

1.4 Contributions

This thesis advances the study of artificial-intelligence (AI) risk in the labour market along conceptual, methodological, and empirical, dimensions. Its core contributions are summarised below.

Conceptual contribution. This thesis introduces a *multi-resolution, dynamic framework* that links machine and AI risk debate across three nested levels of analysis: whole *occupations* (Chapter 3), their constituent *skills and tasks* (Chapter 3 and Chapter 4), and the *aligned AI capabilities* that can now perform those tasks (Chapter 5). By fusing these layers, it unifies strands of the literature that have traditionally been studied in isolation, while the incorporation of year-on-year benchmark momentum reframes AI risk as a *moving frontier* rather than a static snapshot, thus aligning risk measurement with the accelerating pace of technological progress. In addition, the thesis extends the classic *task-based approach* by embedding each task in a rich semantic representation, weighting it by real-world salience, and explicitly linking it to concrete AI benchmarks; this enriches the micro-level lens through which labour economists and policymakers assess how technological change translates into occupational vulnerability. Finally, an interpretable career mobility model LADDER-GNN is proposed to connect AI's impact assessment with individual-level career trajectories.

Methodological contribution. This thesis advances automation-risk modelling through a sequence of complementary methods. First, it develops a graph-based approach that captures how automation risk diffuses through shared skills, uncovering spillover pathways invisible to occupation-centric indices (Chapter 3). Second, it introduces natural-language understanding into the analysis by fine-tuning transformer models on job task descriptions, producing a task-level taxonomy that distinguishes between substitution, complementarity, and negligible automation risk (Chapter 4). Third, it establishes a scalable and transparent job task-AI alignment framework that leverages large language models to connect human tasks with AI benchmarks from the Intelligence Task Ontology, enabling traceable mappings between workforce activities and technical capabilities (Chapter 5). Forth, it proposed a graph-to-sequence interpretable career mobility model to bridge the AI impact with individual career trajectories. Finally, these elements are integrated into a dynamic exposure index that combines measures of technological momentum, task salience, and cross-domain coverage, offering a forward-looking and interpretable assessment of AI's evolving impact on work.

Empirical contributions. By applying the proposed models to large-scale, real-world data, this thesis generates several novel empirical insights. First, the graph-based framework in Chapter 3 uncovers a high-risk frontier of occupations that share vulnerable skills, revealing clusters where automation risk cascades across interconnected roles. Second, Chapter 4 highlights the heterogeneity of risk within occupations, showing that even jobs traditionally viewed as safe contain mixtures of tasks that may be substituted, complemented, or remain unaffected by AI. Third, the capability-weighted index in Chapter 5 reveals exposure patterns that had been overlooked, including unexpected vulnerabilities in education-related roles, while also confirming that blue-collar and middle-income work remain at the technological frontier. Finally, Chapter 6 extends the analysis to the individual level, introducing a framework for predicting talent potential and demonstrating how measures of career mobility and stability can be linked to AI exposure. This integration raises new questions about whether high-potential individuals are systematically shielded from automation risks, or whether they cluster in highly exposed roles where adaptability becomes critical. Taken together, these findings show that the thesis not only reconciles prior inconsistencies in exposure measurement but also advances a richer, multi-level understanding of AI's impact—spanning tasks, occupations, and individual career trajectories.

1.5 Thesis Structure

This dissertation is submitted as a conventional thesis under a common research theme (University of Technology Sydney, 2024). During my doctoral candidature I completed four full papers, publishing two in influential conferences while the third was submitted to a high-impact journal; a fourth paper is currently finished and ready to be submitted at another top-tier journal. Reflecting this publication-focused trajectory, my research papers are presented as independent chapters, framed by an overarching introduction, a comprehensive literature review, methodologies and a concluding chapter. The thesis therefore comprises seven chapters, structured as follows:

- Chapter 1 outlines the context, motivation, central research questions, and overall contributions related to understanding and investigating AI's impact on job market.

- Chapter 2 synthesises occupation-, skill-, and task-level studies of ai risk; reviews graph, NLP, and benchmark-driven methodologies; and identifies the research gaps addressed in subsequent chapters.

- Chapter 3 introduces a GCN Network over the occupation–skill bipartite graph, revealing the high-risk occupational frontier and validating results against BLS employment-decline projections. (RQ1 & RQ4)

- Chapter 4 presents a BERT-powered taxonomy of O*NET tasks, quantifies substitution, complementarity, and negligibility, and uncovers within-occupation heterogeneity across Australian industries. (RQ1 & RQ4)

- Chapter 5 aligns human tasks with ITO AI tasks via LLMs few-shot reasoning, introduces gain-popularity weighting, and demonstrates the index's explanatory power—highlighting the unexpected vulnerability of education occupations. (RQ2, RQ3 & RQ4)

- Chapter 6 extends the analysis to the individual level, introducing LADDER-GNN -a framework for predicting talent potential and demonstrating how measures of career mobility and stability can be linked to AI exposure. (RQ5)

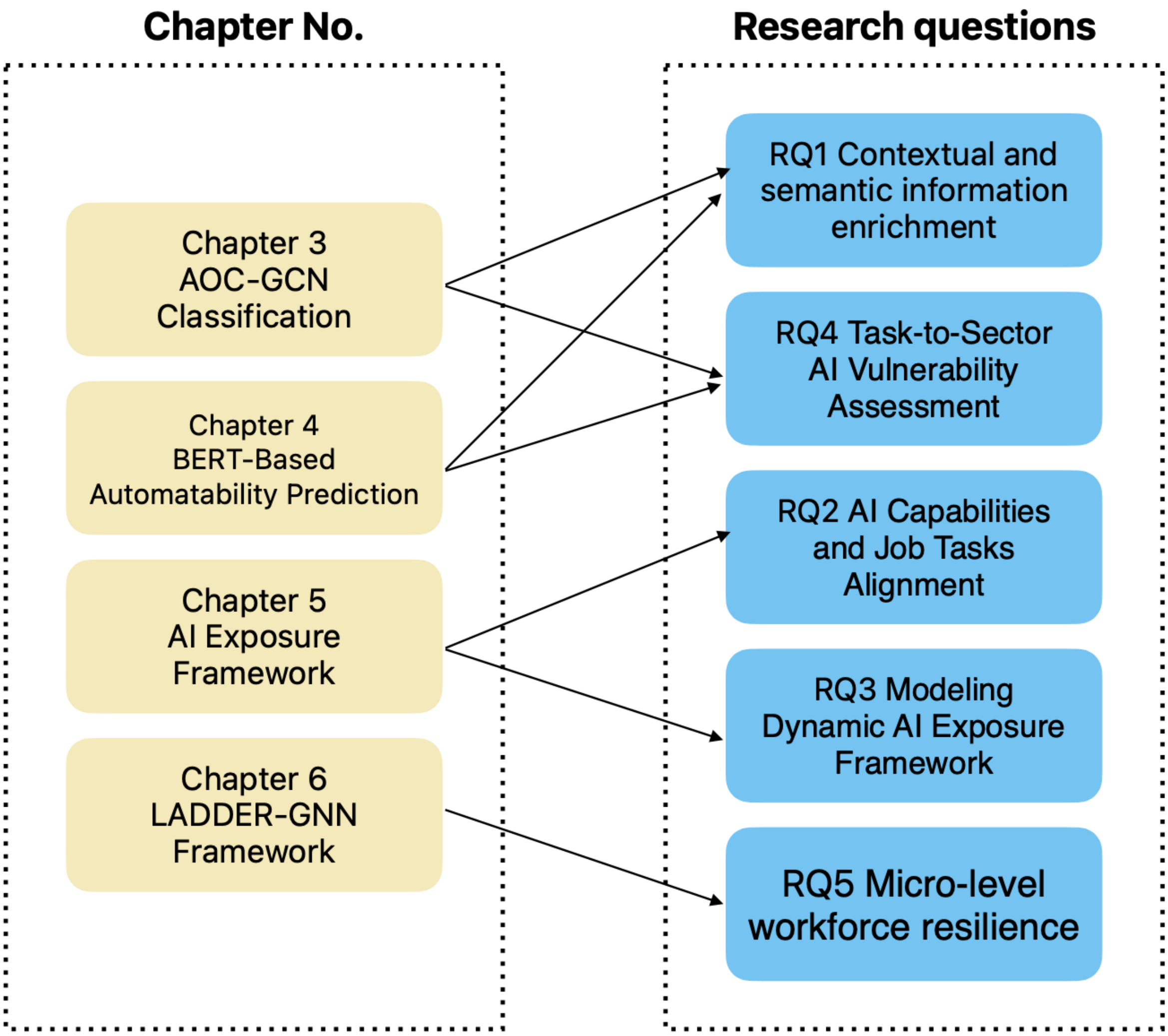


Figure 1.1: Thesis organization

- Chapter 7 maps a detailed research agenda that includes longitudinal labour-market forecasting, real-time AI-capability tracking, cross-country validation; and outlines concrete policy initiatives for government, industry, and higher education—particularly in the areas of targeted reskilling programmes and curriculum redesign.

Chapter 2

Literature Review

This chapter introduces the literature review underpinning this thesis, covering the historical perspective on technological advances and their labor market impacts, including significant periods such as the Industrial Revolution, automation, and the IT revolution. It further examines current trends in artificial intelligence (AI) and robotics, along with their increasingly pervasive impacts on the labor market. To effectively assess these impacts, the chapter reviews various methods developed to measure AI's risks to employment, such as task-based approaches, ability-based analyses, econometric modeling, expert forecasting, and ontology-based frameworks. Additionally, it discusses advanced AI techniques like deep learning models, Graph Neural Networks (GNNs), and large language models (LLMs, e.g., GPT), which significantly enhance our ability to predict and understand AI-driven job transformations.

2.1 Technologies and Job Market

Technological change has long been a central force shaping the structure and dynamics of labour markets. Across successive waves of innovation—from mechanisation and electrification to computerisation and, more recently, artificial intelligence—new technologies have systematically altered how work is organised, what tasks are performed, and which skills are valued. These transformations occur not simply through the replacement of labour, but through more nuanced processes of task reconfiguration, productivity augmentation, and the reallocation of human effort across occupations and sectors. Importantly, technological advances affect labour markets at multiple levels: by modifying task content within jobs, reshaping occupational boundaries, and influencing sectoral composition over time. Contemporary AI-driven technologies thus represent a continuation of this broader historical trajectory, albeit with distinctive characteristics arising from their ability to replicate or augment cognitive and decision-making tasks. Understanding the labour-market implications of AI therefore requires situating it within the longer tradition of technological change, drawing conceptual and empirical insights from earlier technological revolutions while recognising the novel mechanisms through which modern digital and AI-based systems interact with work.

2.1.1 Historical Technological Change

Technological change has shaped labour markets since the onset of industrialisation, with concerns about machines displacing human labour dating back to the Industrial Revolution. Early resistance, most notably the Luddite protests against mechanised looms in nineteenth-century England, reflected anxieties that new machinery would render workers redundant (Brynjolfsson & McAfee, 2015). The British Industrial Revolution provides a foundational illustration of these dynamics: steam power, mechanised spinning and weaving, and the factory system dramatically increased output per worker—by around 46% between 1780 and 1840—yet real wages stagnated for decades in what is known as "Engels' pause" (Allen, 2009). Skilled artisans experienced displacement, fuelling social unrest and lending support to early warnings of technological unemployment, including Ricardo's concerns and Marx's notion of an industrial reserve army (Mokyr et al., 2015; Ricardo, 2015). Early industrial labour was further characterised by long hours, hazardous conditions, and extensive child and female employment, prompting later regulatory intervention and union mobilisation (Allen, 2009; National Research Council, 1983). Over time, however, the labour-market effects of industrial technologies evolved: from the mid-nineteenth century onward, real wages rose rapidly and eventually outpaced productivity growth as capital deepening, mass education, new complementary occupations, and collective bargaining enabled workers to capture a greater share of technological gains (Allen, 2009). These historical patterns gave rise to two recurring narratives—one emphasising job loss and social disruption, and the other highlighting productivity growth, new industries, and rising living standards (Brynjolfsson & McAfee, 2015; Mokyr et al., 2015). Task-based economic models reconcile these views by distinguishing between automation, which replaces labour in specific tasks, and the creation of new tasks in which labour retains a comparative advantage, showing that employment, wages, and labour's income share depend on the balance between task destruction and task creation (Acemoglu & Restrepo, 2018; D. H. Autor, 2015). In this interpretation, early mechanisation was initially labour-saving but ultimately employment-expanding through demand growth and new labour-intensive activities (Acemoglu & Restrepo, 2019; Brynjolfsson & Mitchell, 2017). Historical and comparative evidence nonetheless underscores that these effects were uneven and contingent, shaped by institutions, skills, and development timing, with late-industrialising economies often experiencing "premature deindustrialisation" when imported technologies failed to absorb surplus labour (Anelli et al., 2021; Braverman, 1998; Goldin & Katz, 2008; Rodrik, 2015).

Technological change intensified in the early to mid-twentieth century with electrification, the internal combustion engine, chemical engineering, and early computing. Assembly-line production raised productivity but fragmented work into narrowly defined, repetitive tasks, reinforcing concerns about deskilling (Braverman, 1998). Mechanisation in agriculture produced even larger employment shifts, reducing farm labour from roughly 40% of the U.S. workforce in 1900 to under 2% by 2000, while displaced workers were largely absorbed into manufacturing and services (D. H. Autor, 2015). Post-war advances in computing revived fears of a "laborless" economy, reflected in prominent policy reports during the 1960s (Brynjolfsson & Mitchell, 2017). In practice, long-run unemployment did not rise persistently; instead, automation reconfigured work by shifting human effort toward customer-facing, supervisory, and judgment-intensive tasks (D. H. Autor, 2015). Tinbergen described this as a "race between education and technology," with expanding secondary and tertiary education supplying the skills demanded by new technologies and supporting broad-based wage growth and declining inequality in the mid-century period (Goldin & Katz, 2008). Although critics such as Leontief warned of eventual human obsolescence once machines mastered both physical and cognitive tasks (National Research Council, 1983), most evidence pointed toward complementarity rather than wholesale substitution (D. H. Autor, 2015; Brynjolfsson & Mitchell, 2017).

The late twentieth-century information technology (IT) revolution introduced a distinct pattern of labour-market adjustment. The diffusion of computers, software, and digital networks proved strongly skill-biased, raising demand and wages for highly educated workers while eroding middle-skill employment (D. H. Autor et al., 1998; Katz & Murphy, 1992). Routine-biased technological change theory explains this shift by noting that computers automate codifiable, repetitive tasks while complementing non-routine cognitive and interpersonal activities (D. Autor et al., 2003). Empirical studies document declining clerical and production employment alongside growth at the top and bottom of the skill distribution, producing job polarization in the United States and Europe (D. H. Autor & Dorn, 2013; Goos & Manning, 2003). Computerisation has also been linked to rising wage inequality, declining labour shares, and increased capital substitution in routine-intensive tasks (Acemoglu & Restrepo, 2018; Karabarbounis & Neiman, 2013). These trends were amplified by globalization and institutional changes such as declining unionisation (Acemoglu et al., 2022; Krugman, 1994). Beyond wages and employment, IT altered job quality through digital monitoring, work intensification, and the expansion of gig-based labour, reviving debates over precarity and control (Acemoglu & Restrepo, 2017; Mokyr et al., 2015). In developing economies, digital technologies facilitated rapid service-sector growth while accelerating automation in manufacturing, reinforcing concerns about premature deindustrialisation (Rodrik, 2015). While technology alone cannot explain all labour-market trends (D. H. Autor, 2015; Card & Dinardo, 2002), the IT era firmly established task-based frameworks as essential for analysing how technological progress reshapes work and distributional outcomes, setting the stage for contemporary debates on artificial intelligence and its potential to extend automation into new domains (Frey & Osborne, 2013).

2.1.2 The Age of Automation and Artificial Intelligence

Artificial Intelligence (AI) is broadly defined as the field of computer science dedicated to creating systems capable of performing tasks that typically require human intelligence, such as reasoning, learning, perception, and decision-making (McCarthy et al., 2006). Modern AI encompasses a range of technologies, including machine learning (ML), deep learning, natural language processing, and robotics. These systems are increasingly able to analyze data, recognize patterns, understand language, and interact with the physical world in sophisticated ways. The concept of automation, meanwhile, refers to the use of technology to perform tasks without human intervention. While automation has a long history—stretching from mechanical looms to assembly lines and industrial robots—recent advances in AI have greatly expanded the range of tasks that can be automated. Traditionally, automation focused on routine, repetitive, and rule-based activities, particularly in manufacturing and clerical work. AI, by contrast, enables the automation of more complex, non-routine, and even cognitive tasks that were once considered the exclusive domain of humans (D. H. Autor, 2015; Brynjolfsson & Mitchell, 2017). It is important to recognize that while all AI-driven systems that perform work can be considered automation, not all automation involves AI. Early forms of automation relied on fixed rules or simple mechanical actions, whereas AI-based automation leverages algorithms that can learn from data and adapt to new situations. Thus, AI can be understood as a subset and an enabler of the broader process of automation, driving what some scholars describe as a new wave of "intelligent automation" that blurs the boundaries between physical, digital, and human systems. Accordingly, in Chapters 3 and 4 of this thesis, the term automation is used in a general sense to denote the full spectrum of task-performing technologies, encompassing both traditional non-AI systems and AI-driven methods. Within this framework, AI is treated as a distinct but nested subset of automation, representing a newer class of learning-based and adaptive technologies whose labour-market effects are analysed separately when appropriate. This terminological distinction ensures conceptual consistency across chapters and clarifies that empirical measures of automation exposure capture a broader technological phenomenon, while AI-specific analyses isolate the unique mechanisms introduced by modern intelligent systems.

2.1.2.1 Automation (early 2000s–2010s)

At the turn of the millennium, automation accelerated sharply across advanced and emerging economies, driven by falling costs of industrial robots, sensors, and digital control systems. Industrial robots became increasingly prevalent in manufacturing, particularly in automotive, electronics, and metal industries; by 2020, more than 2.7 million robots were operating in factories worldwide (International Federation of Robotics, 2020). Automation also expanded beyond factory floors into logistics, warehousing, and service operations through technologies such as automated storage systems, barcode and RFID tracking, and algorithmic inventory management, reflecting the broader transition toward "Industry 4.0" and cyber-physical production systems (Khan et al., 2023). In parallel, software-based automation advanced rapidly. From the early 2010s, robotic process automation (RPA) tools were adopted to execute routine, rule-based digital tasks—such as data entry, payroll processing, and customer service workflows—enabling firms to streamline back-office operations without physical capital investment (Lacity & Willcocks, 2017). Empirical evidence indicates that these technologies had substantial but uneven labour-market effects. Exposure to industrial robots was associated with declines in manufacturing employment and wages in routine-intensive regions, particularly in the United States and parts of Europe (Acemoglu & Restrepo, 2017; Graetz & Michaels, 2015). However, automation did not generate aggregate mass unemployment: displaced workers partially reallocated into construction, logistics, and service sectors, while new occupations emerged in robot maintenance, systems integration, and process supervision (D. H. Autor, 2015; Bessen et al., 2019). As in earlier technological transitions, automation primarily reshaped the task composition of jobs rather than eliminating work altogether, reinforcing long-standing patterns of adjustment and skill adaptation.

2.1.2.2 Machine learning and Artificial Intelligence (2010s onward)

After decades of incremental progress, artificial intelligence experienced a marked resurgence in the 2010s, driven by advances in machine learning—particularly deep neural networks—alongside rapid growth in computing power, cloud infrastructure, and large-scale digital data (LeCun et al., 2015). Breakthroughs such as DeepMind's AlphaGo defeating the world Go champion in 2016 demonstrated that AI systems could outperform humans in complex, non-routine domains previously considered resistant to automation (Silver et al., 2016). By the late 2010s, AI achieved substantial gains in image recognition, speech processing, natural language understanding, and prediction tasks, enabling widespread deployment across both consumer and enterprise applications (Jordan & Mitchell, 2015). Crucially, these technologies began to diffuse into core business processes, augmenting or partially automating knowledge work rather than primarily replacing manual labour. Machine learning systems were increasingly used in medical imaging diagnostics, fraud detection, recommendation systems, and customer service chatbots, altering task requirements within professional occupations (Brynjolfsson et al., 2018; Topol, 2019). This diffusion reshaped labour demand: empirical evidence from job postings shows that demand for AI-related skills grew rapidly during the 2010s, accompanied by wage premia and strong complementarities with advanced education (Acemoglu et al., 2022; Deming & Kahn, 2018). Rather than simply displacing workers, AI shifted the skill structure of employment and generated new occupational categories—such as data scientists, machine-learning engineers, and AI product managers—highlighting a transition from task automation toward task augmentation and recombination.

2.1.2.3 The Era of Generative AI

In the early 2020s, artificial intelligence entered a qualitatively new phase with the rise of generative AI—models capable of producing novel text, images, code, audio, and video at a level approaching or, in narrow domains, rivaling human performance. This shift was propelled by large language models (LLMs) trained on massive datasets using transformer architectures, culminating in the public release of systems such as ChatGPT in late 2022 (Achiam et al., 2023; Brown et al., 2020). In parallel, image-generation models such as DALL·E 2, Midjourney, and Stable Diffusion demonstrated that generative systems could perform creative tasks once assumed to be uniquely human, including illustration, design, and visual ideation (Ramesh et al., 2022). Similar advances extended to software development, where tools such as GitHub Copilot enabled automated code generation and completion, significantly reducing time spent on routine programming tasks (M. Chen, 2021). Importantly, most generative AI systems are deployed not as autonomous agents but as copilots, designed to augment human decision-making and creativity rather than replace it outright. This paradigm of human–AI collaboration—already visible in domains such as legal research, marketing content creation, and software engineering—signals a shift toward workflow-level augmentation, where AI handles coordination, drafting, and synthesis while humans retain strategic judgment and accountability (Brynjolfsson et al., 2025). Ongoing developments in multimodal and integrated systems, such as Google's Gemini, further suggest that generative AI will become embedded across productivity ecosystems, enabling real-time assistance across documents, communication, and planning tasks.

The diffusion of generative AI has profound implications for labour markets, particularly for knowledge-intensive and creative occupations that were previously less exposed to automation. Unlike earlier waves of technology that primarily displaced routine manual or clerical tasks, generative AI directly affects content creation, analysis, and problem-solving activities central to professional work. Task-level exposure estimates indicate that a substantial share of the workforce is affected: OpenAI researchers estimate that around 80% of U.S. workers have at least 10% of their tasks exposed to GPT-type models, while nearly one-fifth may see half or more of their tasks impacted (Eloundou et al., 2023). Early evidence suggests that generative AI can raise productivity in affected occupations—for example, improving output quality and speed in writing, programming, and customer support—while simultaneously compressing skill gradients by enabling less-experienced workers to perform more complex tasks (Brynjolfsson et al., 2025; Noy & Zhang, 2023). These dynamics raise concerns about task displacement and entry-level job erosion in fields such as software development, design, journalism, and legal research, even as they create demand for new complementary roles related to AI supervision, system integration, and human–AI interaction. Aggregate employment effects remain uncertain, but prevailing projections emphasize job reallocation rather than net job destruction. Consistent with historical patterns of technological change, generative AI appears less likely to eliminate work wholesale than to reconfigure task structures, amplify individual productivity, and accelerate skill-biased demand, reinforcing the need for continuous reskilling and institutional adaptation.

2.2 AI's Risk Assessment: A Methodological Review

A wide array of methodological approaches has emerged to measure or predict how AI-driven automation threaten jobs. These methods range from quantitative econometric analyses to qualitative case studies, from forward-looking forecasts to theoretical frameworks. By surveying empirical (data-driven) methods alongside theoretical and qualitative approaches, we can gain a comprehensive view of what each reveals about AI's role in job displacement, creation, and transformation.

2.2.1 Theoretical Foundations

Before delving into empirical measurement, we briefly review theoretical frameworks that guide the measurement of AI's impact on jobs. The modern economics of automation builds on the *task-based* view of production, which decomposes jobs into discrete activities assigned either to labour or to machines. Within this framework, AI influences labour demand through two forces: the *automation of existing tasks*, which substitutes capital for workers, and the *creation of new human-intensive tasks*, which raises demand where humans retain comparative advantage (Acemoglu & Autor, 2011; Acemoglu & Restrepo, 2018). In the canonical "race between man and machine" model, firms repeatedly weigh the cost of automating a task against the payoff from inventing a new one for labour (Acemoglu & Autor, 2011; Card & Dinardo, 2002). Heavy automation initially reduces wages and labour's share of income; yet falling wages eventually make further automation less profitable, encouraging innovation that restores labour demand. Hence, equilibrium forces can limit long-run displacement, though short-run outcomes may feature sizable job losses and rising inequality, especially for low-skill groups whose tasks are easiest to automate (Acemoglu & Restrepo, 2018). The model also highlights a potential efficiency gap: because each firm internalises only its private cost savings, the economy can end up with *excess* automation relative to the social optimum, implying a role for policy intervention. A complementary strand treats machine learning as a *general-purpose technology* whose ultimate impact depends on complementarities with other innovations rather than one-for-one job substitution (Brynjolfsson & Mitchell, 2017; Goldfarb et al., 2023). Using their "suitability-for-machine-learning" (SML) rubric, Brynjolfsson and Mitchell (2017) demonstrate that most occupations are bundles of heterogeneous tasks: some (e.g. structured perception or routine prediction) are highly automatable, while others (e.g. creative problem-solving or social interaction) remain distinctly human. Partial task automation can therefore (i) augment productivity by freeing workers to concentrate on residual, higher-value tasks or (ii) depress employment if the automated share dominates job content. Theoretical work thus converges on three insights: (1) AI's employment effects are inherently *task-level* and heterogeneous; (2) displacement and augmentation can coexist within the same occupation; and (3) aggregate welfare hinges on the pace and direction of innovation, the distribution of skills, and the policies that mediate the private–social gap in automation incentives. These principles motivate the empirical strategies surveyed in subsequent sections.

2.2.2 Occupational AI Risk Assessment

One of the earliest empirical approaches to measuring AI's impact on jobs was to classify entire occupations by their risk of automation. Occupation-based approaches assign a probability or score of automation risk to each occupation as a whole, typically under the assumption that if a sufficient share of tasks in that job can be automated, the occupation itself is at risk.

A seminal study in this category is Frey and Osborne (2013), who examined the computerization potential of 702 detailed occupations. They developed a novel methodology using a Gaussian process classifier, trained on expert assessments of whether certain example occupations were automatable, to estimate the probability of automation for each occupation. Strikingly, their analysis concluded that about 47% of total US employment was at high risk of automation (over a decade or two) given foreseeable AI and robotic technologies. They also found that automation risk was negatively correlated with wages and education levels, meaning lower-paid, lower-education jobs were more likely to be automated.

However, these occupation-based extrapolations have been criticized for assuming a uniform impact on all workers within an occupation and for ignoring adaptive responses (Arntz et al., 2016; Frank et al., 2019; Moro et al., 2021). The occupation-level approach implicitly treats each occupation as a fixed bundle of tasks that will be entirely automated once technology is capable of performing those tasks. In reality, there is significant heterogeneity in tasks performed by different workers under the same occupation title, and firms may reorganize job roles rather than simply eliminate them (Arntz et al., 2016). As a result, occupation-based risk estimates tend to represent an upper bound on job automation potential – a theoretical maximum if technology were adopted wholesale and no new tasks emerged. Indeed, subsequent research has found substantially lower proportions of jobs at high risk when accounting for task variation within occupations (Arntz et al., 2016). We turn to those task-based approaches next. But the occupation-based studies remain useful for highlighting which *occupations* are most susceptible. For example, Frey and Osborne

(2013) identified jobs such as telemarketers, cashiers, and routine production and clerical roles as having very high automation probabilities (above 0.9), whereas occupations requiring creative intelligence, complex perception, or social intelligence (e.g. teachers, healthcare workers, arts-related occupations) were assessed as relatively safe. These broad patterns have been consistent with expert intuition and have helped focus attention on the types of work AI and robotics are most likely to transform. The strength of occupation-based measures lies in their simplicity and broad coverage of the workforce. Policy makers found such estimates easy to communicate ("X% of jobs are at high risk"). However, the limitation is their coarse granularity and deterministic view of occupations – they do not capture partial automation of tasks or variation in how jobs can change.

In summary, occupation-level risk assessments provide a macro-scale picture of automation potential and have been applied globally, but they likely overestimate near-term job losses from AI. In the past decade, no country has experienced workforce contraction on the scale of 47% despite rapid technological progress. In fact, empirical employment data show that total employment has continued to grow in most countries, even as AI-related technologies diffuse. For example, an OECD study by Arntz et al. (2016) examined what happened to employment in occupations that were rated as "high risk" by such models over the 2010s. They found no net decline in jobs in those high-risk categories at the national level – all 21 OECD countries in their sample saw overall employment growth. However, they did observe that employment grew more slowly in the high-risk occupations (around 6% growth over the decade) compared to low-risk occupations (18% growth). This suggests that while the dire predictions of mass job obliteration have not materialized, occupations flagged as automation-prone have indeed lagged behind in job creation. It underscores that occupation-based risk measures capture real vulnerabilities, but actual outcomes depend on many moderating factors (technology adoption rates, labor supply changes, etc.). Next, we look at more fine-grained task-based approaches that refine these estimates.

2.2.3 Task-Based Methods AI Risk Estimation

Task-based approaches aim to measure AI's impact by analyzing the specific tasks or duties that make up jobs, rather than treating entire occupations as monolithic. This methodology recognizes that within any given occupation, some tasks may be easily automated by AI, while others are difficult to automate (often those requiring human judgment, creativity, or interpersonal skills). By evaluating tasks at a granular level, researchers can estimate what fraction of a job's activities are technically automatable and thereby infer the job's overall exposure to AI.

A pioneering task-based study is the work by Arntz et al. (2016). In contrast to Frey and Osborne's occupation-level assumption, Arntz et al. (2016) argued that "we are likely overestimating automation risk when neglecting the substantial heterogeneity of tasks within occupations". They used micro-data from the OECD's Survey of Adult Skills (PIAAC), which contains detailed information on the tasks that individual workers perform in their jobs. By mapping Frey and Osborne (2013) list of automatable tasks to the PIAAC data, Arntz et al. (2016) estimated the proportion of tasks (for each person) that could be automated. They then calculated the share of jobs that have at least e.g. 70% of their tasks automatable. Using this individual-level task variation, they found that on average across 21 OECD countries, only about 9% of jobs are at high risk of automation, far lower than the 47% figure for the US found by Frey and Osborne. For the United States specifically, their estimate was similarly around 10% of jobs highly susceptible. Moreover, they reported considerable cross-country differences: for example, the share of jobs at high risk was around 6% in Korea and 12% in Austria, reflecting differences in job task composition and possibly technology adoption lags.

In a follow-up study, Nedelkoska and Quintini (2018) (OECD) refined these numbers and suggested roughly 14% of jobs on average are highly automatable (with an additional 32% of jobs facing significant change in task composition). These OECD studies confirm that taking into account the diversity of tasks leads to substantially lower automation risk estimates than occupation-based approaches. Task-based assessments have also highlighted that the risk is concentrated in certain categories of tasks. Routine, codifiable tasks – whether manual or cognitive – are the most susceptible. For instance, repetitive assembly or basic data processing tasks can often be learned by AI or performed by robots. In contrast, tasks involving problem-solving in unstructured environments, creative thinking, or complex communication are currently far less automatable. This aligns with the earlier literature on routine-biased technical change (D. Autor et al., 2003), but updated for the AI era. Importantly, task-level analysis supports the idea that most occupations will experience partial automation rather than complete replacement. Brynjolfsson and Mitchell (2017) emphasize this point: many jobs consist of a mix of SML (Suitable for Machine Learning) tasks and non-SML tasks. As AI automates the SML subset, workers may shift focus to the remaining tasks, potentially enhancing those aspects of the job that require human skills. Task-based measures can thus inform us not only about potential job losses, but also about how job roles might evolve. A strength of the task-based approach is its micro-level realism and flexibility. It acknowledges heterogeneity among workers and can be updated as new

AI capabilities emerge. It is also well-suited for cross-country analysis because tasks data (like PIAAC or job surveys) exist for many countries, enabling comparison of automation exposure across different labor markets.

However, one limitation is that task-based estimates are still largely about *technical potential* – they assume if a task *can* be automated, it eventually will be. In practice, economic, regulatory, or ethical considerations may prevent or delay automation of certain tasks even if AI is capable. Furthermore, task-level analysis at one point in time does not automatically account for the dynamic adaptation of jobs; as tasks get automated, new tasks or responsibilities might be added to a role, which the initial static analysis wouldn't capture. Nevertheless, task-based approaches provide a more nuanced picture than occupation-level studies and have become a standard in measuring AI's job impact. To illustrate, consider a clerical job that involves data entry, customer emails, and creative report writing. An occupation-based approach might label the entire job as automatable if data entry is automatable. A task-based approach would note that perhaps 40% of the job's tasks (data entry) can be done by AI, but the other 60% (customer service and report writing) are not yet automated, meaning the job is only partially exposed. This opens possibilities: the worker might spend less time on data entry and more on customer interaction once AI tools are introduced, which could even improve productivity and job satisfaction rather than result in unemployment.

In summary, task-based methods (using detailed surveys or databases of occupational tasks) have substantially revised downward the estimates of jobs "at risk" and provided insight into what portions of jobs are most affected by AI. They underscore that AI's impact will be uneven across tasks, even within the same job, and hence that the future of work will likely be characterized by humans working alongside AI, focusing on complementary tasks. These analyses form the basis for many policy discussions on reskilling: if we know which tasks are likely to be automated, we can identify the skills workers will need to cultivate to remain employable in the evolving job landscape.

2.2.4 AI Risk Assessment Using Hybrid Methods

The most recent wave of research has utilized machine learning algorithms and novel data sources to measure AI's impact on jobs, effectively creating hybrid methods that combine human analysis with algorithmic assistance. These approaches often involve mapping AI capabilities to job requirements using data-driven techniques such as natural language processing (NLP) on text corpora (like patents, job descriptions, or AI research papers).

The goal is to develop indices of "AI exposure" or "automation risk" that are grounded in actual technological progress signals, rather than solely relying on human judgment.

A notable example is the methodology of Felten et al. (2019). They proposed a new way to link advances in AI to occupational abilities. Specifically, they leveraged performance data from various AI benchmarks and competitions (what they call "third-party measures of past advances in AI") and mapped these advances to the abilities required in different occupations. In practical terms, they identified major application areas of AI (e.g. image recognition, language understanding, strategic game play) and tracked how rapidly AI performance was improving in each area. Then, using the O*NET database of occupational descriptors, they linked these AI application areas to human abilities and skills relevant to occupations. By aggregating these effects, they constructed an index of AI impact for each occupation and industry. An attractive feature of this approach is that it is based on actual progress in AI (e.g., AI vision accuracy improving over time) rather than experts' future guesses. Felten et al. (2019) demonstrated this method and suggested it could help model how advances in AI affect occupations and industries over time. One finding from their initial exercise was that their AI impact index was moderately correlated with Frey and Osborne (2013) risk probabilities but not identical, and it evolved year by year as AI got better in certain domains.

Another influential approach is by Webb (2019), who developed a method to predict the impact of AI (and other technologies) on occupations by textually analyzing patents and job descriptions. As described in a Brookings Institution summary, Webb (2019) approach was to find the overlap between the verbs and objects used in AI patents and those in job task descriptions. For example, if many AI patents discuss methods to "recognize faces" or "diagnose diseases", then occupations for which "recognize faces" or "diagnose patient's condition" are important tasks (such as security guards for the former, doctors for the latter) would be considered exposed to AI. Webb compiled thousands of such verb-object pairs from a corpus of AI patents. By quantitatively measuring this overlap, he created an "AI exposure" score for each occupation. One of Webb's key findings is that AI exposure is highest for high-skill, high-wage occupations – in contrast to previous waves of automation which mainly threatened lower-skill, routine jobs. Specifically, occupations like chemical engineers, physicists, and market research analysts ranked high on AI exposure, whereas many low-wage service jobs (e.g. restaurant servers, janitors) had very low AI exposure (Webb, 2019). This does not mean that AI will outright eliminate high-skill jobs, but it means AI is capable of taking over certain tasks within those jobs (for example, data analysis tasks done by scientists or drafting basic reports done by lawyers). Webb's work is notable for its objectivity – using patent text as an indicator of where innovators are directing AI capabilities – and it aligns with the notion that AI's comparative advantage is currently in pattern recognition, data processing, and prediction tasks, which are prevalent even in many professional occupations. The Brookings report highlighting Webb's study noted that patent-based measures likely have predictive power because inventors investing in patents expect these innovations to be commercially relevant. Thus, they might be a better gauge of future job impact than asking experts, since the act of patenting involves substantial costs and reflects real technological intent.

In a similar vein, Kogan et al. (2021) link patents to occupations but take it further by examining actual worker outcomes. They created a measure of workers' "technology exposure" by seeing how the tasks in a worker's occupation align with the text of new patents (not limited to AI, but any technology). Using U.S. administrative earnings data, they then studied how workers with different exposure scores fared over time in terms of earnings and job displacement. They found an intriguing polarization effect: workers at the bottom of the wage distribution (low-skill jobs) were more likely to be displaced (lose employment or face wage cuts) if their tasks were exposed to new technologies (consistent with automation of routine manual tasks), and at the same time, workers at the very top of the wage distribution also experienced slower earnings growth if their tasks were exposed. They interpret the latter as high-paid workers' specialized skills becoming obsolete in the face of new tech (for instance, a senior radiologist might see diminished demand for their skill in reading images if AI systems handle more of that work). This finding is important because it suggests AI and related innovations might constrain wage growth even for some highly educated workers, potentially contributing to inequality in a new way – by hollowing out not just middle-skill jobs (as past automation did) but even some high-skill advantages. Kogan et al. (2021)'s model and empirical evidence emphasize that technology-driven labor displacement is not solely a low-skill phenomenon.

Also, Tolan et al. (2021) introduce a cognitively grounded, update-friendly framework for quantifying how advances in AI map onto workplace tasks and, by extension, occupations. They construct a three-layer linkage: 59 generic tasks (from O*NET, PIAAC, EWCS) are connected to 14 underlying cognitive abilities via expert-elicited binary matrices, and those abilities are in turn linked to 328 AI benchmarks through systematic literature review. By weighting task shares in each occupation with a research-intensity proxy for every benchmark (e.g., publication and GitHub activity), they derive a continuous AI-exposure score for 119 ISCO-3 occupations while also allowing the reverse query—identifying which AI subfields most threaten or complement any given task. This bidirectional design, open data release, and separation of relatively stable abilities from rapidly evolving benchmarks make the method more granular and future-proof than earlier task-or benchmark-based measures, though it still relies on expert correspondence matrices and research-activity metrics that may lag real-world deployment.

Finally, the year 2023 saw a new kind of AI-hybrid measurement with the advent of powerful large language models (LLMs) like GPT-4. Eloundou et al. (2023) conducted a study using GPT-4 itself, combined with human expertise, to evaluate how LLMs might affect occupations. They developed a rubric to assess each occupation's exposure to LLM capabilities, essentially asking: if someone had access to GPT-4, what percentage of the tasks in this occupation could be done significantly faster (at equal quality) by the AI? They utilized both human subject matter experts and GPT-4 acting as a reasoning engine to classify tasks. The results made headlines: roughly 80% of the U.S. workforce has at least 10% of their tasks that could be affected by LLMs, and about 19% of workers have at least 50% of their tasks exposed. They also found that exposure was spread across all wage levels, with a tendency for higher-wage occupations to have more tasks that LLMs could help with or take over. This again reinforces the theme that modern AI's reach extends into higher-skilled cognitive work. For example, occupations like accountants, programmers, and writers showed high LLM exposure, whereas physically intensive or outdoor jobs (mechanics, firefighters) showed very low exposure. An interesting nuance from the GPT-based study is the distinction between using an LLM alone versus with complementary software tools: the authors estimate that current LLMs alone could speed up about 15% of all tasks in the economy, but if you integrate LLMs into software (with code execution, tool use, etc.), up to 47–56% of tasks could be done faster by machine. This implies that the impact of AI on productivity (and possibly labor demand) could scale dramatically when these models are embedded in workflows. While this study does not directly measure job loss or gain, it quantifies the technical exposure of tasks in a very granular, contemporary manner using AI itself as an evaluator.

In conclusion, machine learning and AI-assisted classification methods represent the cutting edge of measuring AI's impact on jobs. These methodologies innovatively combine insights from computer science—particularly what current AI capabilities enable—with a nuanced understanding of labor economics, especially regarding the detailed task content of occupations. Nonetheless, existing methods predominantly rely on classical machine learning techniques, textual analysis, and human-in-the-loop expert evaluations, which, while effective, might overlook certain complexities in occupational data. To advance further, future methodological development should integrate state-of-the-art deep learning techniques, including transformer-based language models and graph neural networks (GNNs) to enable a more precise and dynamic assessment of AI's impact across interconnected labor markets. Incorporating these advanced computational methods would enhance the robustness, predictive accuracy, and granularity of assessments, significantly improving our understanding of how AI technologies reshape employment landscapes over time.

2.3 Advanced Data-Driven Modelling Frameworks

In this section we systematically review cutting-edge computational approaches—including convolutional and recurrent deep-learning architectures, transformer-based language models, graph neural networks (GNNs), and the latest generation of large language models (LLMs)—and discuss how each class of methods contributes to a richer, more granular, and temporally responsive measurement of AI's labour-market effects.

2.3.1 Interrelations and Topologies of Entities

Modeling job market problems involve data with rich relational structure, where entities (such as tasks and skills) are interconnected. Modeling these interrelations and topologies explicitly can greatly enhance learning, as it injects relational inductive bias into AI systems. Early approaches to learning on structured data include recursive or graph-based neural networks in the 1990s. For example, Sperduti (1993), Sperduti and Starita (1997), and Sperduti et al. (1995) pioneered neural networks that operate on arbitrary structures, applying supervised learning to classifications of graphs and trees. The formal Graph Neural Network (GNN) model was introduced by Scarselli et al. (2009) in 2009, which generalized neural networks to graph-structured inputs by propagating information along edges until reaching a stable state. Around the same time, Micheli (2009) proposed a constructive neural network for graphs that incrementally aggregates node context, reflecting growing interest in learning from graph topology. These early GNNs established that iterative message passing over graph neighbors could compute meaningful representations for nodes and graphs, laying the groundwork for modern advancements.

In recent years, graph neural networks have surged in popularity and capability. Modern GNN variants can be categorized by how they propagate and transform information on graphs. Convolutional GNNs adapt the notion of convolution to graph neighborhoods: the seminal example is the Graph Convolutional Network (GCN) by Kipf and Welling (2017), which employs a spectral filtering approach to aggregate a node's neighbors' features, enabling semi-supervised node classification with impressive results. Spatial GNNs perform neighborhood aggregation directly in the graph domain; for instance, GraphSAGE introduced by Hamilton et al. (2017) uses sampling and pooling to inductively learn node embeddings for previously unseen nodes or graphs. Attention-based GNNs leverage attention mechanisms to weight the influence of neighboring nodes: Veličković et al. (2018) Graph Attention Network (GAT) assigns learned attention coefficients to neighbors, improving performance on node classification by focusing on the most relevant connections. Other notable developments include message passing frameworks unifying these approaches, and powerful architectures like the Graph Isomorphism Network (GIN) by K. Xu et al. (2019), which demonstrated that a sufficiently expressive GNN can distinguish complex graph structures up to isomorphism. These methods, and many others, are surveyed comprehensively by Wu et al. (2021), who categorize GNNs into recurrent, convolutional, and spatial–temporal models and document their applications across domains.

Beyond homogeneous graphs, many problems involve multiple types of entities and relations – requiring heterogeneous and relational graph models. For example, knowledge graphs represent different entity types (e.g. tasks, skills, occupations) connected by various relation types. Traditional knowledge representation learning introduced knowledge graph embedding methods: translational approaches like TransE embed entities (Bordes et al., 2013) and relations into a vector space such that relation-specific translations on entity vectors approximate observed triples. Bordes et al. (2013) was a landmark model of this kind, enabling link prediction and completion on multi-relational data by simple vector arithmetic. A plethora of embedding models followed (TransH (Z. Wang et al., 2014), TransR (Lin et al., 2015), DistMult (B. Yang et al., 2015), ComplEx (Trouillon et al., 2016), etc.), and surveys by Nickel et al. (2016) provide an overview of how relational machine learning has addressed multi-relational knowledge in vector space. However, these embedding approaches often treat the graph in an implicit way. Recent GNN-based models instead perform relational message passing, explicitly accounting for different edge types. An example is the Relational GCN (R-GCN) proposed by Schlichtkrull et al. (2018), which extends GCNs by having separate weight matrices for each relation type, enabling propagation along labeled multigraphs (such as knowledge graphs) while controlling over-parameterization. Such approaches have proved effective for tasks like knowledge base completion and entity classification, as they combine the advantages of deep learning with structured knowledge.

Graph-based methods are particularly relevant for matching task descriptions to AI capabilities because this matching often depends on a network of relationships. Rather than treating a task description and a capability description as isolated text blobs, one can model the contextual relationships: for example, tasks might require certain skills or tools, which an AI system may or may not have. Representing this as a graph – e.g. a bipartite graph linking tasks to required capabilities, or a knowledge graph of tasks, skills, and methods – allows algorithms to reason over connectivity and compatibility. Early work in domains like recruitment has used ontologies or static graphs to encode relations: Phan et al. (2021) built a skill ontology to improve resume-job matching, modeling resumes, job requirements, and skills as interconnected nodes and using ontological relations as edges. X. Wang et al. (2021) took a related approach using subject-term graphs for person-job matching, where they represented resumes and jobs in a graph connected through shared key terms or concepts. These knowledge-driven graphs added semantic relationships beyond simple text similarity, but earlier systems often relied on manually crafted or static structures and could struggle with noisy, real-world data.

Contemporary research has demonstrated that learned graph representations can outperform such static approaches in matching tasks. Graph neural networks can dynamically learn which relations and connectivity patterns are important from data, rather than relying only on predefined ontology links. A recent example by Frazzetto et al. (2025) constructed instance-specific graphs for each candidate-job pair to be matched. In their framework, each graph contains nodes representing the candidate, the job posting, and various attributes extracted (e.g. skills, past experience, education), with edges linking these entities based on semantic similarity or membership (for instance, an edge might connect a candidate to a skill node if the skill appears in their resume). By applying different GNN architectures (GCN, GraphSAGE, GAT, GIN, etc.) on these graphs, they learned a compatibility score for each candidate-job pair. The graph-based models substantially outperformed traditional text-based matching (which often used simple feature similarity or keyword matching) as well as plain neural networks that ignore the relational structure. In particular, the GCN achieved a balanced accuracy of 65.4% in predicting successful matches, versus 55.0% for a non-graph multilayer perceptron baseline. Moreover, the GNN was better at identifying rare positive matches (qualified candidates among many), illustrating that modeling the topology of connections (e.g. how a candidate's skills jointly relate to a job's requirements) captures essential signals that flat models miss. Another benefit is interpretability: representing the task-capability matching problem as a graph can enable explanation (e.g. highlighting which skill nodes formed the bridge between a task and a capable agent) and traceability of the decision, an increasingly important aspect in AI selection systems

In summary, methods that account for interrelations and graph topologies – including GNNs and related relational learning techniques – provide a powerful toolkit for matching tasks to AI capabilities. They excel at modeling structural data where meaning emerges from connections between entities, going beyond independent feature descriptions. By leveraging graph representations, these methods can naturally perform reasoning tasks like link prediction (e.g. will a given capability fulfill a given task?) and node classification (e.g. categorizing a task based on its relational context), which align closely with matching problems. As a result, incorporating GNNs or other graph-based methods helps ensure that an AI system's capability matching is sensitive to the rich topology of skills, prerequisites, and relationships that truly determine compatibility.

2.3.2 Text Semantic Understanding

Understanding the semantics of text – that is, capturing the meaning of words, sentences, or documents – is a longstanding goal in AI, crucial for interpreting task descriptions or capability statements. Over the decades, approaches to semantic understanding have evolved from count-based statistics to advanced neural models. Early statistical methods represented text in high-dimensional spaces without deep context. A classic example is Latent Semantic Analysis (LSA), introduced by Deerwester et al. (1990). LSA applies singular value decomposition to word-document co-occurrence matrices to uncover latent concepts, mapping words and documents into a lower-dimensional "semantic space" where those with similar meaning lie near each other. This was a breakthrough in mitigating vocabulary mismatch (synonyms could align in the latent space), but LSA's linear, global nature limited its ability to represent polysemy or context-specific meaning. In the 2000s, probabilistic generative models like Latent Dirichlet Allocation (LDA) further advanced text understanding by modeling documents as mixtures of topics. Blei et al. (2003) assumes each document's words are generated from a set of underlying topics, where each topic is a distribution over words. LDA effectively captures themes and can disambiguate word senses by assigning them to different topics, providing a richer semantic representation than simple counts. However, these "bag-of-words" approaches still treat word order and local context as secondary – they model what words occur together, but not the precise meaning of a sentence as determined by word sequence and syntax.

The rise of neural network models for language brought a paradigm shift. One milestone was the introduction of dense word embeddings in the early 2010s, which encode each word as a low-dimensional vector capturing its linguistic context (Mikolov et al., 2013; Pennington et al., 2014). The Word2Vec models by Mikolov et al. (2013) popularized this idea. Word2Vec's skip-gram model learns word vectors by predicting surrounding words in a sentence, such that words appearing in similar contexts end up with similar vectors. This yielded famous results like the ability to solve analogies using vector arithmetic (e.g. $\mathbf{v}(\text{king}) - \mathbf{v}(\text{man}) + \mathbf{v}(\text{woman}) \approx \mathbf{v}(\text{queen})$), indicating the embeddings captured abstract semantic relations. Similarly, GloVe (Pennington et al., 2014) combined global co-occurrence statistics with local context windows to produce robust word embeddings that excelled on similarity benchmarks. These representations greatly improved downstream tasks compared to one-hot or frequency-based encodings, as semantic similarity and relatedness were implicitly encoded in the geometry of the embedding space. Nonetheless, word embeddings have an inherent limitation: they are context-independent. The word "bank" will have the same vector whether used in river bank or bank account, which means important contextual nuances are lost.

To address context, the next generation of models produced contextual word embeddings. Instead of a fixed vector per word type, these models generate embeddings on the fly for each word instance in its sentence, using deep neural architectures to incorporate neighboring words. Recurrent Neural Networks (RNNs) (Elman, 1990), especially Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) networks, were early choices for this. For example, the ELMo model (Peters et al., 2018) uses a bidirectional LSTM to output context-dependent representations for each word in a sentence. ELMo demonstrated huge gains on tasks like question answering and coreference resolution by providing word representations that vary with the sentence context (the vector for "bank" becomes different in a financial vs. geographical context). This was a significant leap: contextual models effectively perform a simple form of on-the-fly disambiguation and encode some sense of the sentence structure in their output representations. Around the same time, the method of pre-training and fine-tuning began to dominate. ELMo's vectors were learned via unsupervised "language model" objectives (predicting the next word or reconstructing the input) on large corpora, then used in downstream models – an early example of the pre-train/finetune paradigm that soon became ubiquitous.

The true revolution in text understanding came with the introduction of the Transformer architecture by (Vaswani et al., 2017). Transformers dispense with recurrence and instead rely entirely on self-attention mechanisms to process sequences in parallel, attending to relationships between all pairs of words in a sentence. This proved highly effective at capturing long-range dependencies and nuanced interactions in text. With self-attention, a word can directly peek at any other word's representation, enabling, for instance, a noun to readily attend to an adjective that modifies it even if several words apart. The Transformer's capacity to capture complex patterns in language, combined with its efficiency on modern hardware, made it the architecture of choice for large-scale pre-training. The seminal model BERT (Bidirectional Encoder Representations from Transformers) by (Devlin et al., 2018) exemplified this new wave. Following BERT, there has been an explosion of Transformer-based language models, each pushing some boundary. Some focused on training methodology: RoBERTa (Y. Liu et al., 2019) showed that removing the next-sentence objective, training longer with more data and bigger batches, led to even better performance, suggesting BERT was undertrained. Others introduced new pre-training objectives: XLNet (Z. Yang et al., 2019) proposed a permutation-based language modeling objective, effectively capturing bidirectional context like BERT but without masking (it learns by predicting each word in some random permutation of the sequence). XLNet outperformed BERT on several benchmarks by leveraging the benefits of autoregressive modeling while integrating context from both directions. Another notable model, ALBERT (Lan et al., 2020), addressed the size of these models: it introduced factorized embedding parameterization and cross-layer parameter sharing to drastically reduce the number of parameters without sacrificing performance, making deployment more feasible. Simultaneously, models grew in size and scope.

OpenAI's GPT series took transformers in a different direction: instead of an encoder like BERT, GPT models use a unidirectional (autoregressive) Transformer decoder and are trained to predict the next token in huge amounts of text (the language modeling task). The first GPT (Radford et al., 2018) already showed strong performance on many language understanding tasks via fine-tuning, and GPT-2 (Radford et al., 2019) demonstrated uncanny text generation ability along with some zero-shot task performance (despite not being explicitly trained for those tasks). This culminated in GPT-3 (Brown et al., 2020), a 175-billion-parameter model that was not only state-of-the-art on language generation, but introduced the phenomenon of in-context learning.

In summary, the field has moved from treating text as unordered bags of words to using deeply contextual, knowledge-rich representations. Transformers like BERT and GPT capture the nuances of language – from syntax to semantics to world knowledge – far better than any prior approach. In practical terms, when dealing with task descriptions and AI capability descriptions, these models allow us to encode each description into a form where matching can be as simple as a vector comparison or as complex as an inference query to a language model.

2.3.3 LLM-Based Methods for Text Pair Classification and Matching

Text pair classification – determining the relationship or compatibility between two pieces of text – is a core problem in NLP, encompassing tasks like paraphrase identification, natural language inference, question-answer matching, and indeed matching task descriptions to capability descriptions. Over time, methodologies for text matching have evolved dramatically. Early approaches relied on lexical similarity metrics and manual features: for example, measuring word overlap, edit distance, or simple vector space similarity between two texts. While such metrics (like Jaccard similarity or cosine similarity on TF–IDF vectors) can catch obvious matches, they fail on examples with low lexical overlap or requiring understanding of meaning (e.g. “swap two numbers” vs. “exchange the values of variables” might describe the same task with no shared keywords).

The advent of deep learning brought new approaches to text pair modeling, initially with task-specific neural architectures. One line of work used dual-encoder or Siamese networks, where each text is processed by the same neural network to produce a vector, and then a similarity (or distance) between the vectors is computed. For instance, Mueller and Thyagarajan (2016) presented a Siamese LSTM network that reads two sentences and learns to predict their semantic similarity score. During training, the network adjusts such that pairs with higher similarity produce embedding vectors that are closer together (in Euclidean or cosine sense). This approach improved over raw lexical similarity by learning to map semantically related sentences nearer in embedding space even if their wording differs. Another influential early model is the Deep Structured Semantic Model (DSSM) by Huang et al. (2013), which used a feed-forward deep network to project a query and a document into a common vector space and then computed their dot-product to gauge relevance. However, Siamese setups sometimes struggle to capture fine-grained interaction between two texts, because they compress each text separately without explicit cross-reference. To address this, researchers developed interaction-focused models. These architectures consider the two texts together and allow richer interplay before making a judgment. For example, in answer selection or NLI, models like match-LSTM S. Wang and Jiang (2016) and ESIM (Q. Chen et al., 2017) used attention mechanisms to align words or phrases between the two sequences (e.g. aligning entities or events in a premise-hypothesis pair) and then aggregated this alignment information through LSTMs to decide if the sentences entail or contradict each other. Such models achieved state-of-the-art results pre-2018, but they often needed complex architectures and careful training from scratch on each dataset. A comprehensive survey by Jiang and Cai (2024) covers many of these architectures, noting that while they improved performance, they also highlighted the need for large labeled datasets and significant computational effort for each new task.

The latest development in text pair classification is the rise of using extremely large language models (LLMs) in a few-shot or zero-shot capacity. Models like GPT-3 and its successors can perform text pair tasks without explicit training by interpreting them as natural language queries. For example, one can prompt GPT-3 with: “Task: Translate a document from English to French. Capability: An AI system that converts spoken English audio to text. Question: Does the capability match the task?” and GPT-3 will output a likely answer (“No, because the capability is speech-to-text, not translation”). This is essentially zero-shot classification using the LLM’s knowledge and reasoning abilities. Brown et al. (2020) demonstrated that GPT-3 could achieve strong performance on tasks like answering whether two questions are the same, given just a few example pairs in the prompt (few-shot learning) or even just an instruction. However, early zero- and few-shot methods (e.g., Wei et al. (2022)) offered broad generalization but struggled with nuanced linguistic phenomena, while task-description prompts (P. Liu et al., 2021) and chain-of-thought prompting (Kojima et al., 2022) improved reasoning yet remained limited by context-window size. Retrieval-augmented approaches—including kNN-LMs that select nearest-neighbor examples for prompt inclusion (Gao et al., 2020; Khandelwal et al., 2020)—address token constraints but often lack an explicit diagnostic process. X. Sun et al. (2023) synthesize these ideas in the CARP framework, first extracting surface clues, then inducing a structured reasoning chain, and finally incorporating kNN retrieval from a fine-tuned classifier to ground decisions in task-specific evidence. By unifying clue extraction, reasoning, and retrieval, CARP achieves state-of-the-art performance on multiple benchmarks and demonstrates strong low-resource and domain-adaptation capabilities, charting a path for more interpretable and effective LLM-based classification.

In conclusion, large language model-based methods, whether via fine-tuned bi-encoders, cross-encoders, or prompt-driven LLMs, provide state-of-the-art solutions for comparing and matching text pairs. They bring together robust semantic encoding (thanks to transformer architectures and pre-training on huge corpora) and flexible interaction modeling (through attention mechanisms that compare every element of one text with every element of the other). Combined with strategies for efficiency and leveraging human guidance (like instruction tuning), the current generation of models can handle the task–capability matching problem with unprecedented accuracy and minimal manual feature engineering.

CHAPTER 3

LEVERAGING GCN FOR OCCUPATION AUTOMATION RISK PREDICTION

This chapter introduces a graph-based framework for assessing occupational automation risk, addressing limitations in prior studies that relied on coarse occupation-level data or static assumptions. By modeling the occupation–skill landscape as a heterogeneous bipartite graph, the chapter applies Graph Convolutional Networks (GCNs) to allow risk to diffuse along shared-skill edges, thereby uncovering spillover pathways that traditional indices overlook.

3.1 Overview

The rapid advances in artificial intelligence and robotics have renewed concerns about large-scale technological unemployment and labor market disruption. Historical evidence shows that automation waves often result in significant occupational displacement, creating not only individual insecurity but also broader social and policy challenges. Recent projections suggest that nearly half of existing occupations could be at risk of automation within the next decade. This study investigates how automation risk can be predicted with greater accuracy by exploiting both granular task–skill data and the networked structure of occupations. The guiding research question is: how can we identify which occupations are most at risk of automation by considering not only their internal attributes, but also their interdependencies within the broader labor market ecosystem?

In this chapter, automation is used as a deliberately broad concept, referring to the use of technologies to perform work tasks with reduced or no direct human intervention. This definition follows the tradition established in the automation and labour-economics literature and is not restricted to cutting-edge artificial intelligence. Building on this framework, the empirical analysis in this paper uses tasks and skills as core features to examine how automation exposure varies across occupations. Within this terminology, artificial intelligence is treated as a distinct but nested subset of automation: while all AI-enabled systems that perform work constitute automation, not all automation relies on AI. This distinction provides a conceptual bridge between traditional automation measures and subsequent AI-focused analyses, ensuring consistency across chapters while allowing AI-specific mechanisms to be examined separately where appropriate (D. H. Autor, 2015; Brynjolfsson & Mitchell, 2017; Frey & Osborne, 2013).

The study proposed a framework called Automated Occupation Classification with Graph Convolutional Networks (AOC-GCN). The proposed approach overcomes three key challenges: (1) the lack of granularity in conventional labor datasets, which obscures subtle differences between occupations; (2) the prevalence of non-numerical, text-based task and skill descriptions, which require careful embedding to preserve semantic meaning; and (3) the complex relational structure of the occupation–skill system, which cannot be fully captured by linear models. To tackle these challenges, we apply Natural Language Processing (NLP) techniques (Word2Vec, Doc2Vec) to embed task and skill text into dense numerical vectors, construct a heterogeneous occupation–skill bipartite graph to encode both job-level features and cross-occupation connections, and employ a semi-supervised Graph Convolutional Network (GCN) to allow automation risk signals to diffuse along shared-skill edges.

The contributions of this study are threefold. First, we introduce NLP-based embeddings into the domain of labor analytics, offering richer semantic representations of tasks and skills than conventional numerical ratings. Second, we pioneer the use of graph-based neural architectures for automation risk prediction, showing how relational modeling captures both local features and global dependencies across occupations. Third, we validate the proposed framework against government statistical data, demonstrating that AOC-GCN significantly outperforms baseline models. Beyond technical improvements, this framework highlights how automation risk propagates structurally through occupations, offering policymakers a more reliable tool to anticipate vulnerabilities and design targeted interventions.

By advancing a structured, graph-based approach, this study establishes a method-

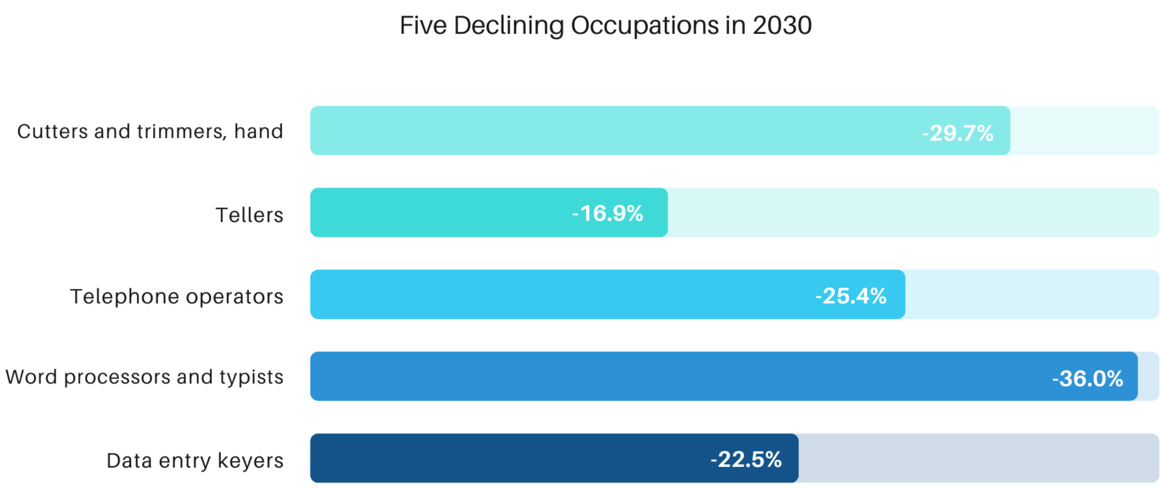


Figure 3.1: BLS Projected Declining Occupations in 2030

ological blueprint for modeling occupational exposure as a dynamic, networked process. It provides new evidence that task- and skill-aware models can enhance our understanding of AI-driven labor market risks and better inform workforce policy and individual career planning.

3.2 Preliminaries about AOC-GCN

Generally, GCN-based models (Hamilton et al., 2017; Kipf & Welling, 2017; Veličković et al., 2018) consist of multiple propagation layers and all the nodes are updated simultaneously in each propagation layer. A message-passing step in a Graph Convolutional Network (GCN) can be described in two phases: *aggregation* and *update*. Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ denote a graph with node set $\mathbf{v} \in \mathbf{V}$, edge set $(\mathbf{v}, \mathbf{v}') \in \mathbf{E}$, and initial node representation $z^0_{\mathbf{v}} \in \mathbb{R}^{p_0}$, where p_0 is the input feature dimension. The embedding of node $\mathbf{v}$ at the k-th layer is written as $z^k_{\mathbf{v}} \in \mathbb{R}^{p_k}$. For a GCN with K layers, the update at layer k $(k = 1, 2, \ldots, K)$ is given by:

$$z^k_{\mathbf{N}(\mathbf{v})} = \phi\Big(M^k \cdot \textsc{Agg}\big(\{z^{k-1}_{\mathbf{v}'} \mid \mathbf{v}' \in \mathbf{N}(\mathbf{v})\}\big)\Big), \tag{3.1}$$

$$z^k_{\mathbf{v}} = \text{COMBINE}\Big(z^{k-1}_{\mathbf{v}},\ z^k_{\mathbf{N}(\mathbf{v})}\Big), \tag{3.2}$$

where $\mathbf{N}(\mathbf{v})$ is the neighborhood of $\mathbf{v}$, AGG aggregates information from neighbors, M^k is a trainable weight matrix, and ϕ is a nonlinear activation (e.g., ReLU). The vector $z^k_{\mathbf{N}(\mathbf{v})}$ represents the aggregated neighbor features, and UPD fuses them with the previous embedding of node $\mathbf{v}$.

In heterogeneous graphs, both nodes and edges may belong to multiple categories, leading to embeddings of different dimensionalities. For an edge **e** connecting two dissimilar nodes, its representation from the prior layer is concatenated with the embeddings of the two incident nodes. Similarly, when computing node updates, the model aggregates not only the information from neighboring nodes but also attributes of connecting edges. The edge-level aggregation at layer k can be expressed as:

$$z_{\mathbf{e}}^{k} = \phi\Big(M_{\mathbf{e}}^{k} \cdot \mathrm{AGG}_{\mathbf{e}}^{k}\big(z_{\mathbf{e}}^{k-1}, z_{\mathbf{p(e)}}^{k-1}, z_{\mathbf{q(e)}}^{k-1}\big)\Big), \tag{3.3}$$

with

$$\mathrm{AGG}_{\mathbf{e}}^{k}(z_{\mathbf{e}}^{k-1}, z_{\mathbf{p(e)}}^{k-1}, z_{\mathbf{q(e)}}^{k-1}) = \mathrm{concat}\big(z_{\mathbf{e}}^{k-1}, z_{\mathbf{p(e)}}^{k-1}, z_{\mathbf{q(e)}}^{k-1}\big), \tag{3.4}$$

where $z_{\mathbf{p(e)}}^{k-1}$ and $z_{\mathbf{q(e)}}^{k-1}$ denote the embeddings of the two node types connected by edge **e**.

For nodes of type **p** and **q**, the aggregated neighbor representations at layer k are:

$$\begin{aligned} z_{\mathbf{N(p)}}^{k} &= \phi\Big(M_{\mathbf{p}}^{k} \cdot \mathrm{AGG}_{\mathbf{p}}^{k}(\mathrm{concat}(z_{\mathbf{q}}^{k-1}, z_{\mathbf{e}}^{k-1}),\ \forall \mathbf{e} = (\mathbf{p},\mathbf{q}) \in \mathbf{E}(\mathbf{p}))\Big), \\ z_{\mathbf{N(q)}}^{k} &= \phi\Big(M_{\mathbf{q}}^{k} \cdot \mathrm{AGG}_{\mathbf{q}}^{k}(\mathrm{concat}(z_{\mathbf{p}}^{k-1}, z_{\mathbf{e}}^{k-1}),\ \forall \mathbf{e} = (\mathbf{p},\mathbf{q}) \in \mathbf{E}(\mathbf{q}))\Big), \end{aligned} \tag{3.5}$$

where node types **p** and **q** each maintain their own parameter matrices $(M_{\mathbf{p}}^{k}, M_{\mathbf{q}}^{k})$ and aggregation functions $(\mathrm{AGG}_{\mathbf{p}}^{k}, \mathrm{AGG}_{\mathbf{q}}^{k})$.

Finally, the hidden states of **p**-type and **q**-type nodes are updated as:

$$\begin{aligned} z_{\mathbf{p}}^{k} &= \mathrm{concat}(U_{\mathbf{p}}^{k} \cdot z_{\mathbf{p}}^{k-1},\ z_{\mathbf{N(p)}}^{k}), \\ z_{\mathbf{q}}^{k} &= \mathrm{concat}(U_{\mathbf{q}}^{k} \cdot z_{\mathbf{q}}^{k-1},\ z_{\mathbf{N(q)}}^{k}), \end{aligned} \tag{3.6}$$

where $U_{\mathbf{p}}^{k}$ and $U_{\mathbf{q}}^{k}$ are trainable transformation matrices, and $z_{\mathbf{p}}^{k}$, $z_{\mathbf{q}}^{k}$ denote the updated embeddings of nodes of types **p** and **q** at layer k.

3.3 Methodology

Our objective is to assess the automation risk for occupations defined by the Standard Occupational Classification (SOC). The attributes for this task consist of task statements and skills extracted from the O*NET database. Each task statement is unique to a given occupation, whereas skill data form a shared set of attributes that may overlap across occupations. Based on this structure, we construct a bipartite graph $\mathbf{G} = (\mathbf{O}, \mathbf{S}, \mathbf{E})$, where **O** is the set of occupation nodes, **S** is the set of skill nodes, and **E** is the set of edges between them. An edge $\mathbf{e} \in \mathbf{E}$ between an occupation $\mathbf{o} \in \mathbf{O}$ and a skill $\mathbf{s} \in \mathbf{S}$ exists if **o** requires **s**. The task statements are used as feature representations of occupation nodes. Accordingly, the problem is formulated as a node classification task defined on an undirected bipartite graph comprising two types of attributed nodes.

The occupation–skill graph is bipartite with two node categories and a single edge type. Since edges in this setting carry no attributes, employing a full heterogeneous GCN would be excessive. Instead, we adopt a simplified approach by treating the bipartite structure as a homogeneous graph, aligning the feature dimensions of both node types. Under this setting, the propagation rule at layer k is given by:

$$z_{\mathbf{v}}^{k} = \text{COMBINE}\Big(z_{\mathbf{v}}^{k-1},\ \phi\big(M^{k} \cdot \text{AGG}(\{z_{\mathbf{v}'}^{k-1} \mid \mathbf{v}' \in \mathbf{N}(\mathbf{v})\})\big)\Big), \tag{3.7}$$

where $\mathbf{v} \in \mathbf{O} \cup \mathbf{S}$, $\mathbf{N}(\mathbf{v})$ denotes the neighborhood of node $\mathbf{v}$, $z_{\mathbf{v}}^{k}$ represents the embedding of node $\mathbf{v}$ at the k-th layer, M^{k} is the trainable weight matrix for layer k, ϕ is a nonlinear activation function, and AGG and UPD denote the aggregation and update operators, respectively.

3.3.1 Nodes Feature Representation

The text data of task statements and skills must be transformed into embeddings before being used as node features. We employ two embedding techniques, Word2vec (Mikolov et al., 2013) and Doc2vec (Le & Mikolov, 2014), to construct feature representations for skill and occupation nodes. First, we initialize Word2vec with pretrained GloVe vectors and further train it on the task statements to capture domain-specific semantics and contextual information. The Word2vec model produces a word embedding table containing vectors for each vocabulary term. Since skill data consist of single words or short phrases, each skill node is represented by either the corresponding word vector or the sum of vectors for multi-word phrases. Next, we initialize a Doc2vec model with the Word2vec embeddings and train it on the task statements to generate document-level representations for occupation nodes. Formally, the initial embeddings are given by:

$$z_{\mathbf{s}}^{0} = \text{Fuse}(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n), \tag{3.8}$$

$$z_{\mathbf{o}}^{0} = \text{Doc2vec}(t), \tag{3.9}$$

where $z_{\mathbf{s}}^{0}$ and $z_{\mathbf{o}}^{0}$ denote the initial embeddings of a skill node $\mathbf{s}$ and an occupation node $\mathbf{o}$, respectively. Here, $\mathbf{w}_i$ represents the i-th word vector from Word2vec associated with a skill, and t is the task statement document corresponding to an occupation. The parameters of Word2vec and Doc2vec are jointly optimized with the remaining components of the model during training.

3.3.2 Semi-supervised Automated Occupation Classification Model

The final embeddings of AOC-GCN are obtained by concatenating the node representations learned from the occupation–skill graph. Formally, we have

$$y = \text{softmax}\big(f(\text{concat}(z_{\mathbf{o}}, z_{\mathbf{s}}))\big), \tag{3.10}$$

where $z_{\mathbf{o}}$ and $z_{\mathbf{s}}$ denote the embeddings of occupation nodes $\mathbf{O}$ and skill nodes $\mathbf{S}$, respectively, as produced by the proposed GCN model. The function $f(\cdot)$ projects the concatenated embedding into a lower-dimensional space prior to applying softmax, which serves as the classifier to distinguish between automated and non-automated occupations. This approach is semi-supervised, since the entire graph is updated during training while supervision is available only for a subset of labeled nodes. The overall workflow of the proposed method is illustrated in Fig. 3.2.

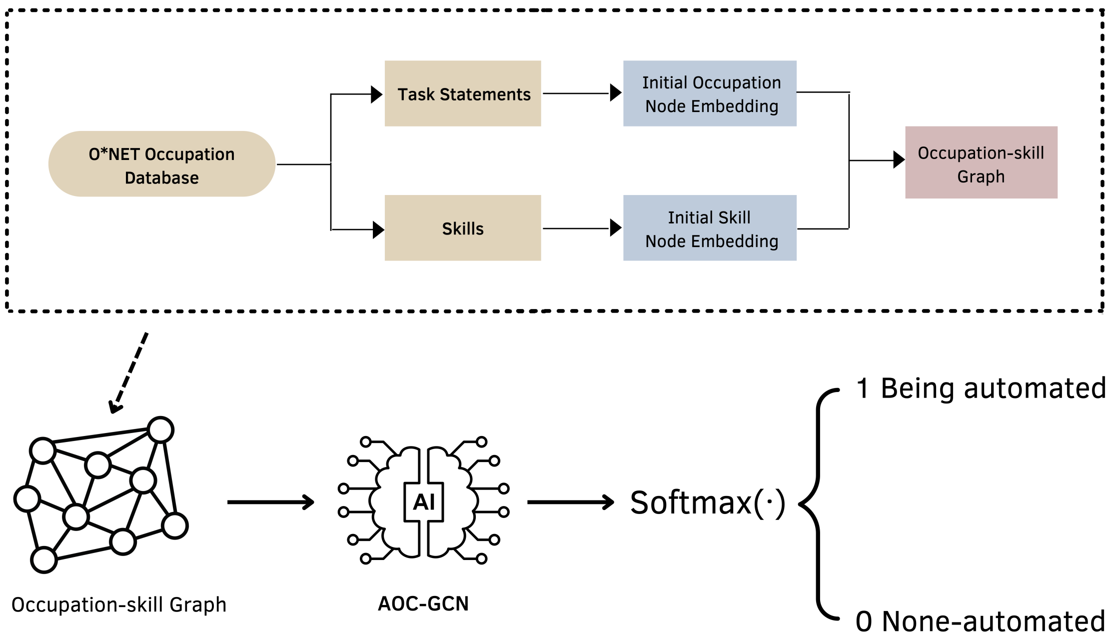


Figure 3.2: Pipeline of the proposed AOC-GCN system.

3.4 Experiments

Experiments are conducted on occupation lists, task statements and skills in Standard Classification Code (SOC) from O*NET database. The label information utilized in the experiments is generated from expert-based assessments on O*NET occupations conducted by

(Frey & Osborne, 2013; Vermeulen et al., 2018) with automated occupations being labeled as 1 and non-automated occupations being labeled as 0. However, the original label information has defects that the number of two classes are unbalanced, and some occupations are too close e.g., web developer and software programmer. We preprocess the data and finally select 112 labeled occupations with 56 automated occupations and 56 non-automated occupations which also have distance between each other. Then, all the data will form an undirected graph structure with 910 occupation nodes, 135 skill nodes and 13222 edges. The dataset is randomly partitioned into training, validation, and test sets with a ratio of 8:1:1. Given the prediction and ground truth, we evaluate the model's performance using accuracy, precision, recall and F1 score.

We compare our model with traditional classification models. Specifically, we design features of each occupation by concatenate the embeddings for both task statements and skills. Then we use zero padding to make sure all the occupations have identical feature dimension sizes before putting into baseline models. The chosen baseline models are as following:

• **Decision Tree**. DT splits the dataset as a tree based on a set of rules and conditions which is a supervised learning algorithm that can be used for both classification and regression (Quinlan, 1986).

• **Random Forest**. Random Forest is an ensemble method which combines the output of multiple decision trees to reach a single result (Ho, 1995).

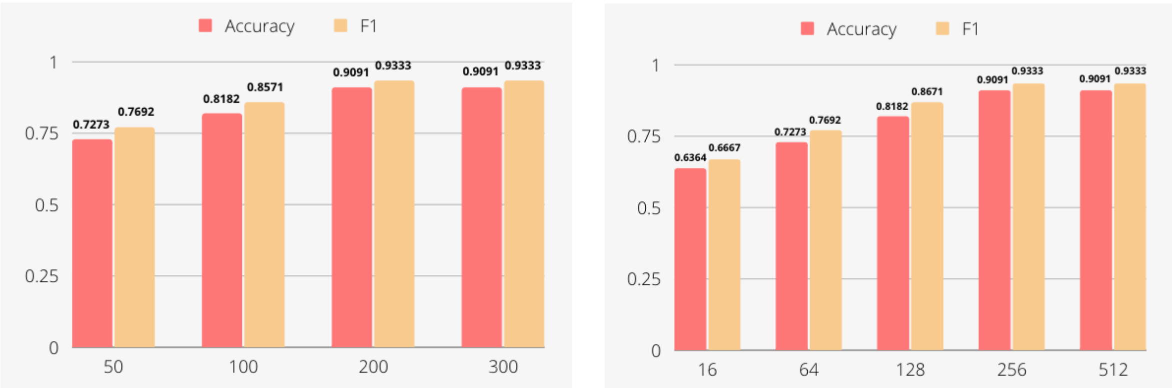
• **Adaptive Boost**. AdaBoost (Adaptive Boosting) is a boosting technique that aims at combining multiple weak classifiers to build one strong classifier and is widely used in both classification and regression problem (Freund & Schapire, 1996).

• **Light Gradient Boosting Machine**. LightGBM is a gradient boosting framework based on decision trees to increase the efficiency of the model and reduce memory usage (Ke et al., 2017).

We evaluate the results of AOC-GCN model on O*NET dataset with above metrics and compare them with baseline models. The results of the comparison are shown in Table 3.1 below. We find that our AOC-GCN method performs significantly better than other baseline models, since our model not only uses external node features, but also capture the graph structural information which is the interactions between nodes. Besides, baseline models using raw features only already achieve good results, the GCN's combination of both semantic and structural information of relations successfully attains the best performance. Furthermore, our proposed model ensures high model precision to avoid disturbing non-automated occupations.

Table 3.1: Preprocessed results of O*NET dataset

Method	Accuracy	Precision	Recall	F1
Decision Tree	0.7143	0.6667	0.8571	0.7500
Random Forest	0.7500	0.7692	0.7143	0.7407
LightGBM	0.7500	0.6842	0.9286	0.7879
AdaBoost	0.8571	0.8125	0.9286	0.8667
AOC-GCN	**0.9091**	**1.0000**	**0.8750**	**0.9333**



((a)) Model performance on different initial node embeddings size

((b)) Model performance on different GCN layer size

Figure 3.3: Parameters Effect On Model Performance

Next, we explore the effects on the model performance of different parameters. The key parameters used here are the dimensions of initial node embedding sizes and the dimension sizes of GCN layer. The dimension sizes of node embeddings range from 50 to 300 and the dimension sizes of GCN layer range from 16 to 512.

We can see from the Fig 3.3 that with the increase of sizes of initial node embeddings from 50 to 200, both the accuracy and F1 scores rise and reach the maximum value. When the size is greater than 200, both of the values keep stable. The same pattern can be found in GCN layer dimension sizes, when the size increases from 16 to 256, both of the accuracy and F1 increase. When the size is greater than 256, the two scores keep unchanged. Therefore, we can conclude that the model gets the best performance when the initial node embeddings size is greater than 200 and the GCN layer dimension size is greater than 256.

Next, we visualize the embeddings generated from the GCN layer using t-Distributed Stochastic Neighbor Embedding (t-SNE) (Hinton & Roweis, 2003) which is a feature reduction technique mitigating the effects of the "Curse of Dimensionality". We also conduct an unsupervised K-means clustering on the dataset and visualize the result using a dimensionality reduction algorithm Principal component analysis (PCA) (Hotelling, 1933).

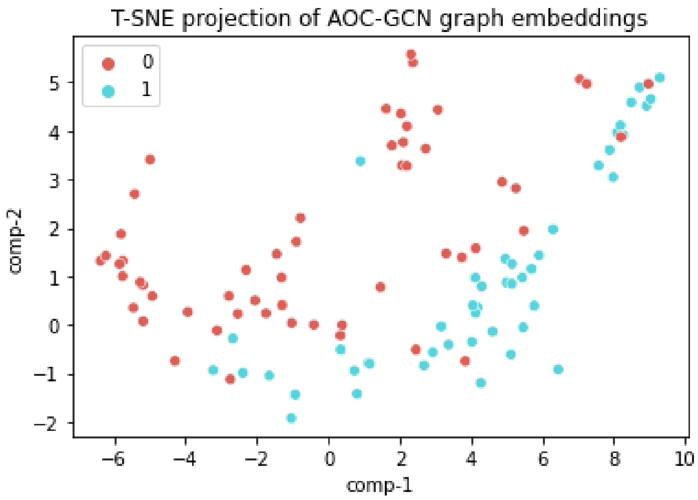


Figure 3.4: t-SNE projection of GCN embeddings

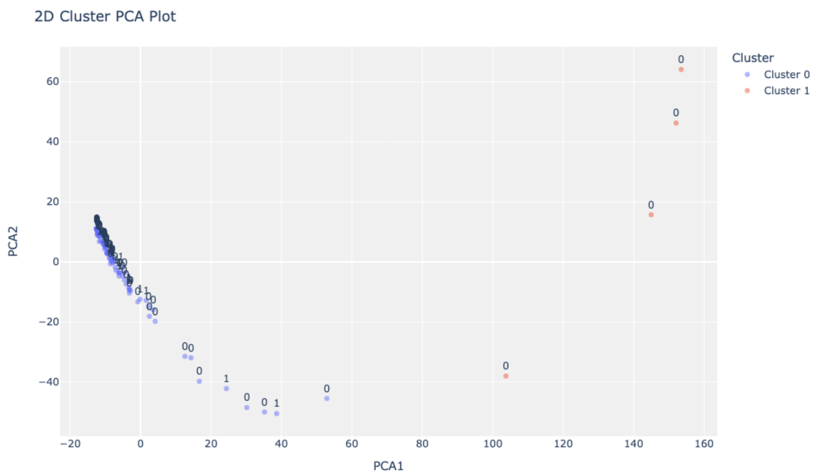


Figure 3.5: PCA projection of K-means clustering

Through comparison, we can find out the difference between applying unsupervised learning and semi-supervised learning. The results are showing in Fig 3.4 and 3.5.

We discover that there is a distinct boundary between label 1 and label 0 which are colored in blue and red in t-SNE projection of GCN embeddings in Fig 3.4. However, there is no distinguishable separation of two K-means clusterings colored in red and blue from PCA projection Fig 3.5, in which labels are annotated in the graph as 1 and 0. It means that GCN semi-supervised method achieved much better performance compared to the unsupervised training by using only a small portion of labelled training data.

Table 3.2: Top 10 Occupations in High Risk of Being Automated

Occupation	**Risk**
File Clerks	0.7002
Real Estate Brokers	0.6968
Word Processors and Typists	0.6943
Payroll and Timekeeping Clerks	0.6941
Data Entry Keyers	0.6940
Transportation Engineers	0.6938
Credit Analysts	0.6938
Insurance Appraisers, Auto Damage	0.6938
Insurance Underwriters	0.6938
Tax Examiners and Collectors, and Revenue Agents	0.6938

Our results show that the maximum probability of automated risk is 0.7002 which is the “File Clerks”. We follow the definition in (Frey & Osborne, 2013) and set 69% as the cut-off point so that if the probability is greater than 69% then this occupation has a risk of being automated. Then we obtain 230 occupations and a percentage of 25.3% among all occupations are at risk of being automated. This result is lower than the experiments conducted by (Brzeski & Burk, 2015; Frey & Osborne, 2013; Pajarinen et al., 2015) which are about 50% occupations are at risk and is also higher than (Arntz et al., 2017; Vermeulen et al., 2018) which conducted a job-level estimation of 9% automated risk. Considering the defects in occupation-level and job-level automated risk estimations, our result is quite reasonable to reach a median value between these two estimations since it predicts the occupation-level risk using granular task and skills data.

To further validate our results, we compare them with “U.S. Bureau of Labor Statistics (BLS) Occupations with the largest job declines, 2020 and projected 2030”. It projected out 29 occupations that have declining employments in 2030. The outcome is visualized in Figure 3.6.

We assume that automated risk greater than 50% would cause employment declining and we can find that 76% of the BLS projected declining occupations have automated risks higher than 50% except for “customer service representatives", “computer programmers", “inspectors, testers, sorters, samplers and weighters" which have an automated risk probability of 0.1%, 8.15% and 7.5% that are obviously automated safe but are projected declining in 2030. This discrepancy can be caused by rules used in expert assessment of the ground truth data. According to Frey and Osborne (2013), experts assess the risks based on three attributes of an occupation which are “perception and manipulation", "creative intelligence tasks" and “social intelligence tasks". i.e., “customer service representatives"

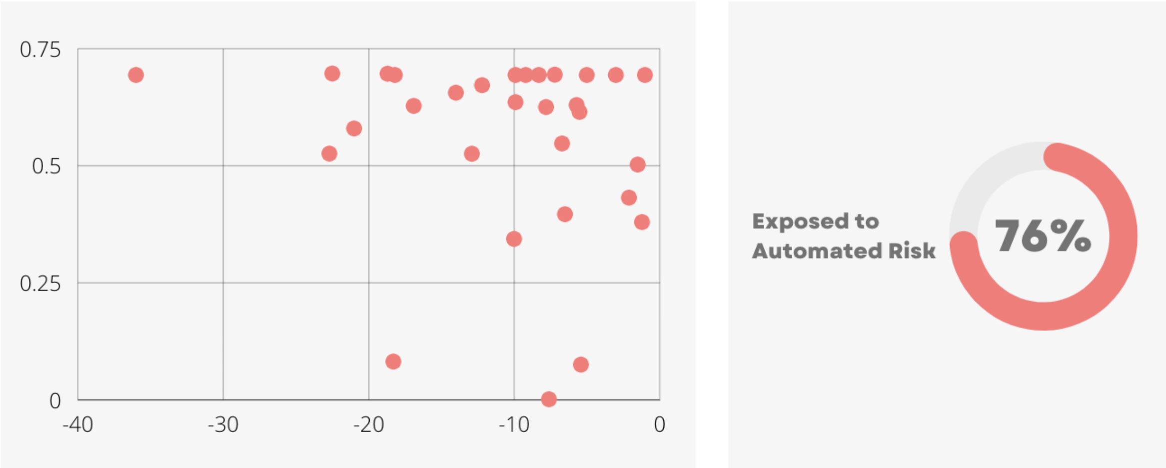


Figure 3.6: BLS Projected Declining Occupations in 2030

have tasks including persuading and dispute resolution which requires higher social intelligence and perception, therefore, they are not easy to be automated according to the analysis in Frey and Osborne (2013). BLS projection made predictions based on employment change in time series. And employment decline may be caused by various reasons e.g., supply and demand change (Vermeulen et al., 2018), other than automation technologies which mean the declining employment of “customer service representatives" can be caused by the demand change, i.e. some the customers’ basic demands can be satisfied by AI dialogue agents. That is to say, expert assessment based on occupations’ attributes only can be biased since some occupations may still decline even if they have tasks and skills that automation technologies cannot perform because of the supply and demand change. Overall, the performance and evaluation demonstrate the effectiveness of our AOC-GCN method in occupation’s automated risk prediction.

3.5 Summary

In this paper, we applied the AOC-GCN model to the Occupation-skill graph to identify the automated risk of a specific occupation. The initial node embeddings are generated from Word2vec and Doc2vec. GCN learns the graph structure and interactions between occupations. Extensive experiments and comparisons with real-world projections demonstrate the effectiveness of our method. We acknowledge several limitations in our method and the results presented in this paper. Firstly, data limitations hindered the performance of our model. O*NET is the only globally admitted database in labour market, however, it only classified 1016 occupations, and this data size limited the performance of classification using machine-learning method. Moreover, each occupation is an aggregation of several jobs which still obfuscates the difference between jobs. Secondly, lack of ground truth data is also a barrier that inhibits predictions of automated risk. Currently, most of the ground truths are expert assessments based on occupation's attributes which are inevitably subjective to some extent. Expert assessments based on occupations' attributes can be biased without considering the supply and demand change. Moreover, there is no job-level or skill-level ground truth data which impedes predictions on a more granular level. Thirdly, predicting on static information cannot capture the technological change and labor trends since the employment trends and changing demand for specific tasks and skills might change faster than static O*NET data can capture. Therefore, our future work will focus on automated risk identifications on job-level which requires more granular ground truth data such as which skills or tasks decide a specific occupation being automated (Dawson et al., 2021). Moreover, the study will be conducted on real-world datasets that capture technological change, employment trends and supply and demand change.

CHAPTER 4

Semantic Modeling of Job Tasks with BERT for Automation Risk Prediction

This chapter presents my second research work which shifts automation-risk analysis from occupations to tasks, arguing that the semantics of task descriptions reveal whether they are susceptible to substitution, complementarity, or negligibility. This study introduces a task-level framework that combines semantic annotation, data augmentation, and fine-tuned BERT models to classify whether tasks are more likely to be substituted, complemented, or unaffected by automation. By aggregating task-level predictions, the approach uncovers nuanced patterns of exposure across occupations and industries, showing how task-centric, context-aware modeling offers a sharper lens on AI's labor market impact and provides actionable insights for policy and workforce reskilling.

4.1 Overview

This study investigates the automatability of work at the level of individual tasks, moving beyond traditional occupation-level assessments that often mask important heterogeneity. While earlier research has offered influential estimates of automation risk, it has typically relied on coarse, job-level data and overlooked the rich semantic signals embedded in task descriptions. To address this gap, we develop a BERT-based classifier that leverages contextual and semantic information from textual task statements to predict whether tasks are susceptible to substitution, complementarity, or negligibility with respect to AI.

Our approach draws on three major sources of task descriptions—O*NET Task Statements, ESCO Skills, and Australian Labour Market Insights—each annotated into the three categories of automation susceptibility. We enrich this dataset through GPT-4 paraphrase augmentation to mitigate class imbalance, and fine-tune BERT to capture subtle linguistic cues that reveal the nature of a task. The model substantially outperforms traditional machine learning, recurrent neural networks, and alternative transformer architectures, demonstrating the value of context-aware language models in labor market analytics.

Beyond predictive performance, we aggregate task-level results to derive occupation- and industry-level insights. This enables a more granular understanding of where automation risks are concentrated, highlighting both substitution vulnerabilities and areas of complementarity where AI is more likely to augment human work. In doing so, the study provides not only methodological advances but also practical implications for workers, firms, and policymakers facing the challenges of rapid technological change.

The contributions of this work can be summarized as follows:

- Introduces a task-level framework for predicting automation risk, offering a more nuanced view than occupation-level analyses.
- Demonstrates that contextual and semantic features of task statements can be effectively leveraged using BERT to classify tasks by their susceptibility to automation.
- Shows that the proposed method outperforms a range of baseline models, underscoring the strength of transformer-based architectures in this domain.
- Provides a curated and annotated dataset combining O*NET, ESCO, and Australian sources, supporting robustness and cross-contextual validity of the model.

4.2 BERT-Based Semantic Modeling of Job Task Risks

This study develops a task-level framework for predicting automatability by integrating expert-labeled datasets, data augmentation, and a fine-tuned BERT-based classifier. The methodology is designed to move beyond coarse occupation-level proxies by directly leveraging the semantic and contextual information contained in task statements. Figure 4.1 provides an overview of the full pipeline.

In this study, automation is defined broadly as the potential for technologies to perform or partially perform work tasks with reduced human involvement, rather than as the adoption of specific AI systems. The notion of automatability used here follows the task-based tradition in the labour economics literature, where exposure is determined by the characteristics of tasks—such as routine content, cognitive requirements, and contextual dependence—rather than by the presence of particular technologies at a given point in time. Accordingly, the expert labels integrated into our framework reflect general susceptibility to computerisation and task substitution, encompassing both traditional rule-based automation and more advanced algorithmic methods. While recent advances in artificial intelligence, including large language models, expand the range of tasks that may be automated, they are treated in this paper as part of a broader automation continuum rather than as a distinct or exclusive focus. Thus, the task-level predictions produced by the model should be interpreted as measuring exposure to broad automation potential, with AI understood as an important but nested subset of automation technologies whose specific effects can be analysed separately when required.

The framework in this study proceeds in three stages. First, we construct a comprehensive dataset of task statements by combining three major public sources: O*NET, ESCO, and Australian Labour Market Insights. Each task is annotated into one of three categories — Substitution, Complementarity, or Negligibility—based on expert consensus, reflecting its susceptibility to automation. To address class imbalance and enrich task variability, we apply GPT-4 paraphrase augmentation, generating diverse yet semantically consistent variants of the task descriptions.

Second, we employ BERT to encode task statements into dense contextual embeddings. BERT's bidirectional transformer architecture enables the model to capture subtle semantic nuances in language, allowing it to differentiate between tasks that appear similar at the surface but differ in meaning and automatability. These embeddings are fed into a classification layer that assigns probabilities across the three automation categories.

Finally, the model is fine-tuned on the annotated dataset to adapt pre-trained representations to the domain of labor tasks. This process optimizes classification accuracy while maintaining interpretability through the use of attention mechanisms, which highlight the words and phrases most influential in model predictions. Aggregated predictions are then extended from the task level to occupations and industries, producing a more granular and dynamic map of automation risk.

This methodology thus combines expert knowledge, data augmentation, and state-of-the-art language modeling to provide a robust and scalable framework for measuring task-level exposure to AI and automation technologies.

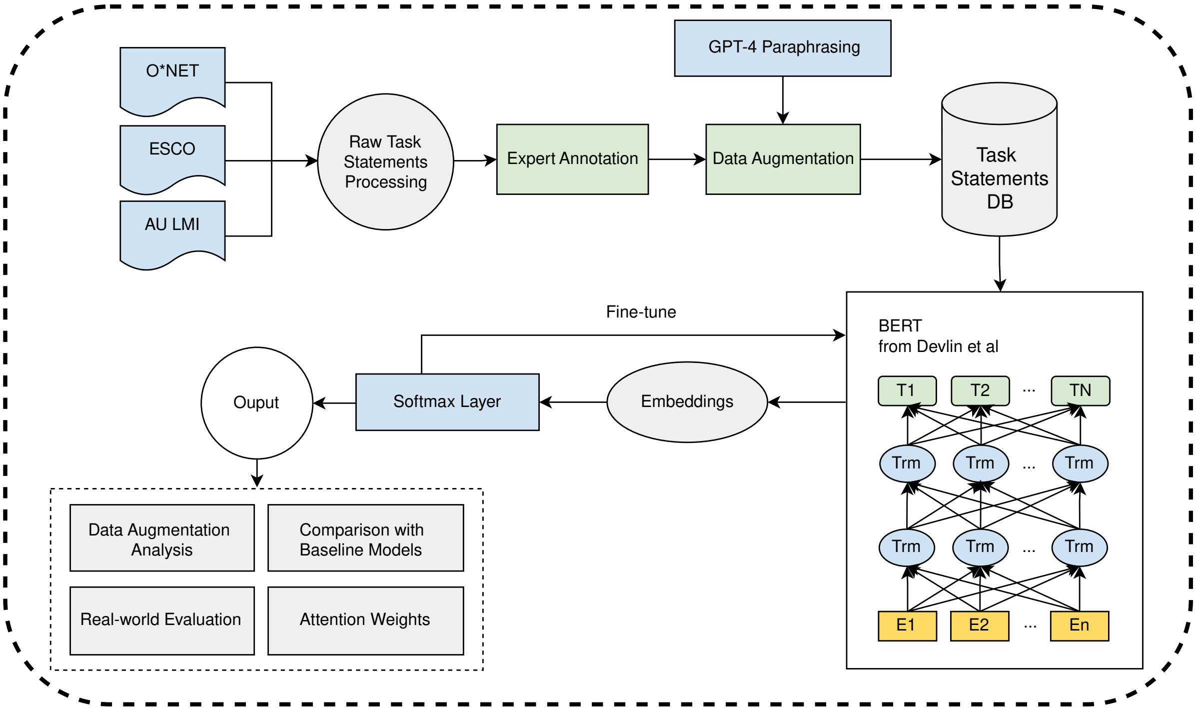


Figure 4.1: The framework of the proposed method.

4.3 Methodology

4.3.1 Dataset Creation

We used three public datasets: ONET Task Statements, ESCO skills, and AU Labor Market Insights Task statements. They provide a wide range of occupation-specific task statements with varying geographical and vocational coverage. We randomly sampled tasks: 5,060 from ONET, 4,783 from ESCO, and 3,356 from AU Labor Market Insights. This was done to ensure diversity across occupations, sectors, and regions for generalizable findings and to optimize computational resources and model training efficiency due to the high-dimensionality of text data and the large dataset sizes.

4.3.1.1 Expert Annotation

Our annotation process employs a voting mechanism involving five experts. For each task and each class, if a class receives more than three votes, it is designated as the final label for that task. This approach ensures a consensus among experts and enhances the reliability of the labels. Our expert annotation process labeled each task as “Substitution" (fully automatable), “Complementarity" (partial automation with human involvement), or “Negligibility" (unlikely to be automated). This provides a categorical understanding of task-level automation susceptibility. We base our process on the Skill-Biased Technological Change (SBTC) (Card & Dinardo, 2002) and Routine-Biased Technological Change (RBTC) (D. Autor et al., 2003) hypotheses, which explain technological change's impact on labor markets. SBTC postulates that technology favors high-skill tasks and replaces lower-skilled routine jobs (Kristal, 2020), while RBTC asserts that technology replaces routine tasks, regardless of skill level (Buyst et al., 2018). We also incorporate six major automation bottlenecks : Complex Problem Solving, Social Interaction and Emotional Intelligence, Fine Motor Skills and Dexterity, Creative and Artistic Abilities, Contextual Understanding, and Ethical Decision Making (Chui et al., 2016; Nedelkoska & Quintini, 2018). These bottlenecks highlight areas where humans still outperform machines.

4.3.1.2 Data Augmentation

The initial exploration of the labeled data revealed a class imbalance issue, a common problem in many real-world classification tasks (Krawczyk, 2016). This imbalance can lead to biased models, overfitting to the majority class and neglecting minority ones. Conventional data augmentation methods suitable for image data aren't applicable to text data due to its sequential and semantic nature (X. Zhang et al., 2015). Apart from class imbalance issue, effectively increasing the volume of training data is particularly beneficial for deep learning models like BERT that perform better with larger datasets.

We addressed this using GPT-4 (Achiam et al., 2023) to paraphrase sentences, creating new task statements with identical meanings but different wording. GPT-4's ability to generate high-quality and varied paraphrases ensures the augmented data remains relevant and increases the classifier's robustness to different task expressions. Specifically, each statement was fed into GPT-4 for paraphrasing, followed by automated checks and expert reviews to validate the paraphrases for semantic similarity to the original statements:

Original Sentence:

Generate reports utilizing visual aids such as charts, graphs, and narratives by examining and documenting test data.

Paraphrased Sentence:

Create reports that *incorporate* visual *elements* like *diagrams, plots,* and descriptive narratives by *scrutinizing* and *recording* the results of tests.

4.3.2 BERT-based Classifier

BERT stands out with its multi-layer bidirectional Transformer encoder, which enables it to capture the deep contextual information embedded in text data (Devlin et al., 2018). Mathematically, this is achieved by applying the attention mechanism, where the output embeddings E for a given input token is a weighted sum of all input token embeddings:

$$E = \sum Attention\_Scores \times Token\_Embeddings \tag{4.1}$$

For our classification task, we fine-tuned a pre-trained BERT model on our task statement data. Each task statement was fed into BERT, which transformed the text into high-dimensional embeddings.

The embeddings were then passed through a softmax function to generate the final class probabilities. If we denote the output of our model before the softmax layer as z, the softmax function can be expressed as:

$$Softmax(z_i) = \frac{e^{z_i}}{\sum e^{z_j}} \tag{4.2}$$

where z_i is the output corresponding to the i-th class and the sum in the denominator runs over all possible classes.

The training process involved adjusting the model's parameters to minimize the discrepancy between the predicted and actual class labels, typically quantified using the cross-entropy loss function:

$$L = -\sum y\_true \times \log(y\_pred) \tag{4.3}$$

where *y_true* is the true label and *y_pred* is the predicted probability.

4.3.3 Attention Mechanism

A fundamental feature of BERT's ability to capture deep contextual information is its use of the attention mechanism, specifically, the scaled dot-product attention as introduced in the original Transformer model (Vaswani et al., 2017). This attention mechanism allows the model to weigh the importance of different words in a sentence, providing a powerful tool for understanding language semantics.

Mathematically, the attention mechanism can be described as mapping a query and a set of key-value pairs to an output. Given a query (Q), keys (K), and values (V), the output (O) of the attention mechanism is calculated as a weighted sum of the values, where the weight assigned to each value is determined by the query's compatibility with the corresponding key:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.4}$$

where the denominator $\sqrt{d_k}$ is used for scaling, with d_k being the dimension of the key vectors. The softmax function ensures the attention scores are normalized to lie between 0 and 1, thus can be interpreted as probabilities. This results in words that are more important to the meaning of the sentence receiving a higher attention score, while less important words receive a lower score.

The use of attention weights allows us to visualize and interpret the model's decision-making process. By examining these weights, we can understand which parts of the input sentence are most influential in determining the output of the model.

4.4 Experiments

The datasets were partitioned into training, evaluation, and test subsets following an 8:1:1 ratio, providing a comprehensive and balanced basis for our experimental studies. And both the original and augmented datasets would be passed through the model to evaluate and quantify the influence of data augmentation.

The table 4.1 presented below provides a visual representation of both the initial and augmented datasets. Specifically, during the training phase using the original dataset, we observed that the AU LMI dataset exhibited superior performance compared to the other two datasets. Consequently, during the process of data augmentation, we prioritized augmenting the ONET and ESCO datasets to a greater extent, while applying relatively fewer augmentations to the AU LMI dataset. This decision was based on the recognition that ONET and ESCO present greater challenges in terms of machine learning, warranting additional exposure to augmented instances for effective learning.

We assessed our model and baselines using precision, recall, and F1 score. Precision evaluates the ratio of correct positive predictions to all positive predictions, indicating the model's false-positive avoidance. Recall measures the fraction of true positives from all actual positives, reflecting the model's ability to recognize all relevant instances. The F1 score, defined as the weighted average of precision and recall, provides a balanced measure of both; a high value reflects strong performance in terms of both precision and recall.

Table 4.1: The overview of dataset

		O*NET	ESCO	AU LMI
Substitution	Original	1,594	1,435	998
	Augmented	**3,188**	**3,157**	**1,796**
Complementarity	Original	2,519	2,272	1,776
	Augmented	**3,023**	**3,181**	**1,776**
Negligibility	Original	947	1,076	582
	Augmented	**3,030**	**3,228**	**1,764**

In our multi-class problem, these metrics were computed for each class, considering the class in question as positive and the rest as negative. The average of per-class metrics provided a single performance measure across all classes.

In our work, we make a comprehensive comparison of our proposed BERT-based model with a range of baseline models, encompassing both traditional machine learning approaches, neural network architectures, and other transformer models. These baselines span different approaches to text classification, allowing us to assess the relative merits of our approach in a broad context.

For traditional classifiers, we employ Logistic Regression (Cessie & Houwelingen, 1992), Random Forest (Breiman, 2001), and Support Vector Machines (SVM) (Cortes & Vapnik, 1995). These models, with their varying theoretical underpinnings, provide a solid foundation against which to compare our more complex neural model. They represent different forms of linear and non-linear decision boundaries and incorporate different forms of regularization and ensemble learning.

In the realm of neural networks, we utilize Bi-directional Long Short-Term Memory (BiLSTM) (Graves & Schmidhuber, 2005), Gated Recurrent Units (GRU) (Cho et al., 2014), and one-dimensional Convolutional Neural Networks (Conv1D) (Yoon & McDonnell, 2020). These architectures demonstrate the potential of deep learning for text classification, and their comparison with our model allows us to quantify the value of pre-training and transformer architecture in this context.

Finally, we include three additional transformer models in our baselines: ALBERT (Lan et al., 2020), ELECTRA (Clark et al., 2020), and DistilBERT (Sanh et al., 2019). As close relatives of BERT, they provide a stringent test of the specific benefits of our BERT-based approach. By contrasting with these models, we can highlight the unique strengths of our chosen methodology.

Table 4.2: The results of proposed model and baselines.

Model	O*NET			ESCO			AU LMI			O*NET + ESCO + AU LMI		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
LR	0.6091	0.61	0.609	0.5915	0.5915	0.5911	0.657	0.6602	0.6566	0.5902	0.5922	0.5908
SVM	0.6087	0.6072	0.6046	0.5738	0.5719	0.5707	0.6462	0.6473	0.6452	0.5837	0.5867	0.5811
RF	0.5922	0.593	0.5921	0.5657	0.5656	0.5652	0.6449	0.6483	0.6399	0.5889	0.5909	0.5891
BiLSTM	0.7487	0.7562	0.7532	0.7553	0.7558	0.7445	0.8107	0.8002	0.8135	0.7676	0.7701	0.7665
GRU	0.7665	0.7562	0.7512	0.7587	0.7612	0.7611	0.8235	0.8202	0.8135	0.7609	0.7624	0.7611
CONV1D	0.7695	0.7354	0.7444	0.7641	0.7622	0.7521	0.8224	0.8102	0.8008	0.7766	0.7798	0.7723
ALBERT	0.7401	0.6989	0.6988	0.7687	0.7723	0.7605	0.7997	0.8033	0.7946	0.7253	0.7114	0.6801
ELECTRA	0.7544	0.7621	0.7602	0.7384	0.7381	0.7285	0.8145	0.8093	0.8122	0.7459	0.7419	0.7425
BERT	**0.7698**	**0.7643**	**0.7594**	**0.7701**	**0.7723**	**0.7698**	**0.8241**	**0.8225**	**0.8198**	**0.7981**	**0.7955**	**0.7944**

The table 4.2 summarizes the results for proposed BERT and other baseline models of different datasets. We evaluate them with precision, recall and F1-score.

As we can see from the results, the GRU model demonstrates the highest performance on the O*NET dataset, with Precision, Recall, and F1-Score at 0.7712, 0.7676, and 0.7712, respectively. BiLSTM exhibits superior performance on the ESCO dataset, with Precision, Recall, and F1-Score at 0.7973, 0.7898, and 0.7876, respectively. The same model outperforms others on the AU LMI dataset, with all three metrics around 0.8995.

When all datasets are combined, our proposed BERT-based classifier emerges as the most effective model. The precision, recall, and F1-score are 0.7981, 0.7955, and 0.7944, respectively, which are the highest among all models. These results indicate the robustness of BERT in handling diverse and large-scale datasets and its superiority over traditional, neural network, and other transformer models for this multi-class classification task. We will use BERT as the best performance model for follow-up experiments and analysis.

Now, we examine the impact of different augmentation levels on our BERT model's performance. The data augmentation methods include no augmentation (Original), augmenting data to match the class with the largest instances (Balanced), and incremental augmentation of the original data by 1.5, 2, 2.5, 3, 4, and 5 times. The results are shown below as Fig. 4.2:

A careful review of the results indicates that augmentation improves the precision of the model across all three datasets (ONET, ESCO, and AU LMI). For the ONET and ESCO datasets, the precision increases modestly from “Balanced" to “2 Times" and then plateaus slightly at “2.5 Times". However, a dramatic rise is seen at “4 Times", implying potential over-augmentation, which could lead to an overfitted model.

Therefore, we choose to augment our data between “Balanced" and “2 Times" for the O*NET and ESCO datasets. As for the AU LMI dataset, which already shows good perfor-

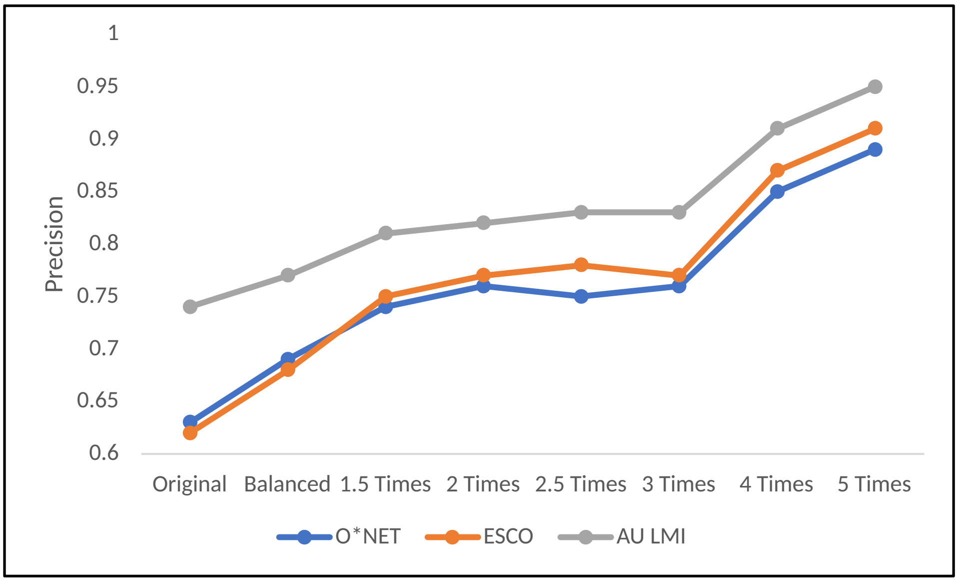


Figure 4.2: The Comparison of different augmentation.

mance, we only balance the classes.

To ascertain the robustness of our proposed model for real-world applications, we conducted an experiment with varying training set sizes. The dataset was shuffled, and training sets were created with different proportions of the original dataset, namely 80%, 70%, 60%, 50%, 40%, 30%, and 20%. The remaining data in each scenario was used for performance evaluation. The results are presented at Fig. 4.3:

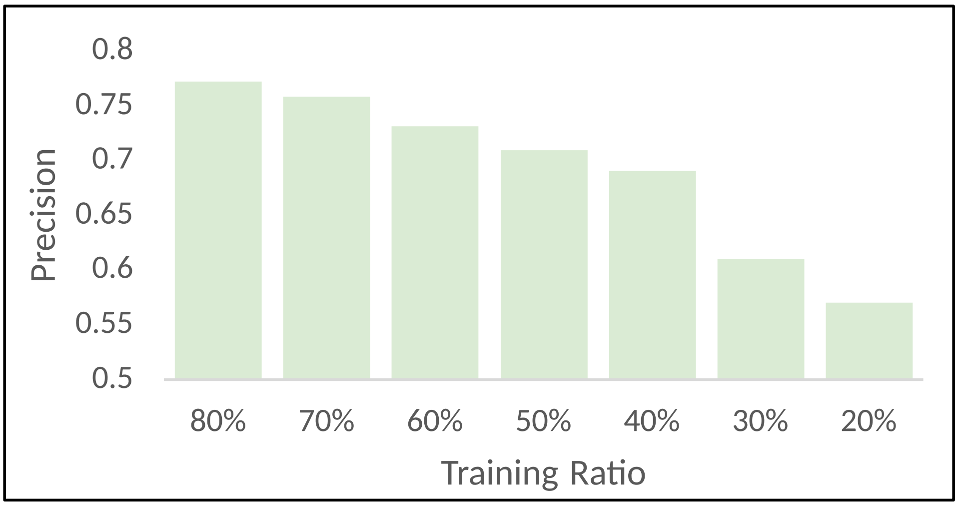


Figure 4.3: The performance of model at different data split.

As can be observed, the proposed BERT model's performance is relatively stable, and it gets better performance with the increase of training data ratio. Besides, our model with only 20% training data can catch up with the baselines who use 80% data for training, which validates the effectiveness of BERT in the task of predicting the automatability at task level.

Although BERT-based classifiers are now relatively established in NLP, their performance in this study reflects suitability to the task rather than architectural novelty. Task statements are short, semantically dense, and context-dependent, requiring nuanced language understanding that is not well captured by surface features or sequential models alone. BERT's bidirectional transformer architecture is effective in modeling such contextual semantics and in handling linguistic variation across heterogeneous task datasets. Empirically, BERT's advantage over alternative methods is modest but consistent. As shown in Table II, while recurrent and convolutional models perform competitively on individual datasets, BERT exhibits the most stable performance when datasets are combined and under reduced training data, indicating stronger generalization and robustness rather than a large absolute performance gain.

For the visualization of the model's attention, we have employed the technique of Wordcloud. For class Substitution, the prominent terms include “using system", “machinery operate", “data record", “routine perform", and “trucks load". These tasks are generally routine and predictable, and hence, are more prone to automation. On the other hand, for class Complementarity and class Negligibility, the key terms hint at tasks that require a higher level of human judgment or interaction like “information provide", “educational program", “research conduct", “medical procedures", “human expertise", and “children care". These tasks correlate with recognized automation bottlenecks, corroborating the model's predictions. The results can be found at Fig. 4.4:

Having established the optimal performance of our BERT model through a series of experiments, we moved to a practical application, leveraging the model for inference on real-world datasets. We selected O*NET task statements, a rich and diverse dataset that encapsulates a broad variety of professional tasks, totalling 19,530 individual tasks. The results showed that among the total tasks, 6664 tasks were labeled as “Substitution", 10,678 tasks were “Complementarity" and 2188 tasks were “Negligibility" which could be visualized as Fig. 4.5:

We mapped individual task automatability measures to 974 distinct occupations using the O*NET task statement to occupation mapping. This allowed us to assess automatability at an occupation level by aggregating tasks and quantifying task type distributions. The results, summarized in Table 4.6, show the top 10 occupations with the highest substitution and negligibility, indicating the most and least likely automated occupations, respectively. As can be seen from the results, occupations with high automation susceptibility are pri-

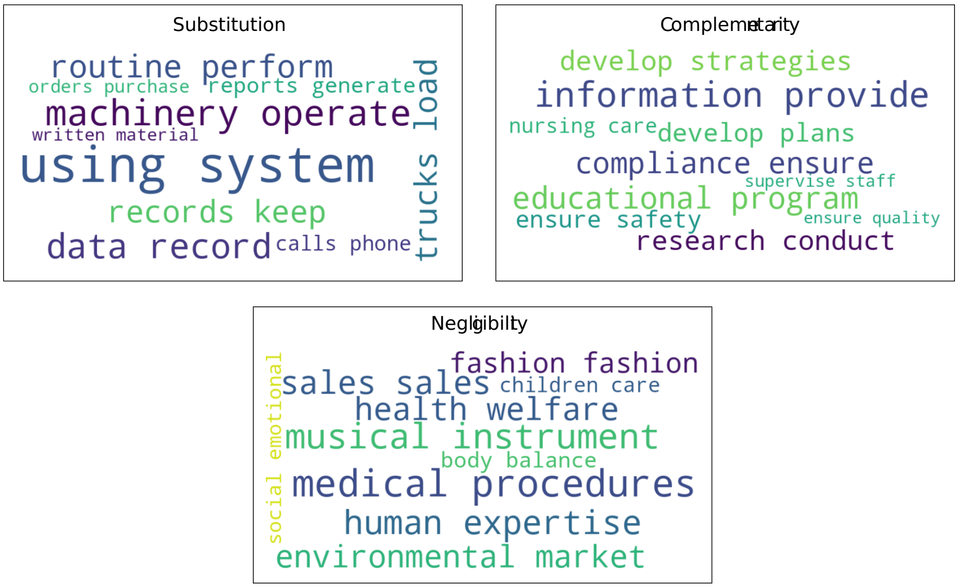


Figure 4.4: The Wordclouds for different categories.

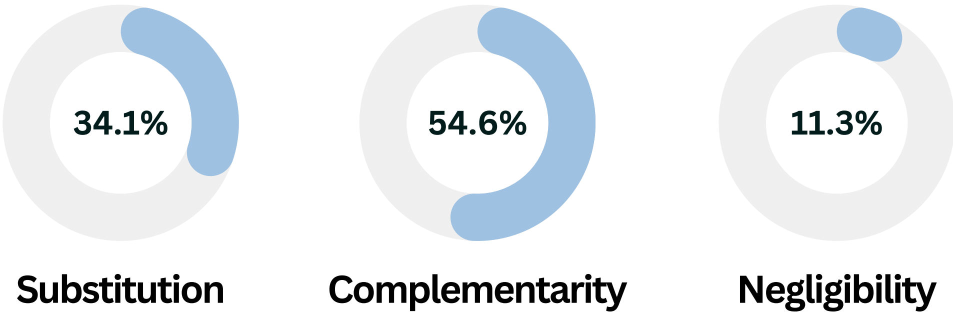


Figure 4.5: The Distribution of O*NET Task Automatability.

marily manual labor, repetitive tasks, or data-intensive roles, such as "Dishwashers" and "Packaging and Filling Machine Operators and Tenders". Conversely, roles requiring human interaction, creativity, specialized skills, or unpredictable environments, like "Athletes and Sports Competitors" and "Respiratory Therapists", showed lower automatability.

Our research findings indicate that out of 974 ONET occupations, 244 display a substitution score exceeding 50%. This suggests that approximately 25.1% of occupations within the ONET database are at substantial risk of automation. This outcome is aligned with the findings put forth in D. Xu et al. (2022), thus adding robustness to our estimations. Furthermore, these results are nested between the estimations reported in Frey and Osborne (2013) which places 47% of occupations at risk and Arntz et al. (2016) which estimates the figure at a comparatively lower 9%. Considering the limitations often associated with automated risk evaluations at the occupation and job level, the median value that our research

TOP 10 OCCUPATIONS WITH HIGHEST AUTOMATABILITIES			
	SUBSTITUTION	COMPLEMENTARITY	NEGLIGIBILITY
Dishwashers	1	0	0
Fabric Menders, Except Garment	1	0	0
Packaging and Filling Machine Operators and Tenders	1	0	0
Pile-Driver Operators	1	0	0
Agricultural Equipment Operators	0.94	0	0.06
Packers and Packagers, Hand	0.92	0.08	0
Tire Builders	0.91	0.09	0
Word Processors and Typists	0.89	0.11	0
Data Entry Keyers	0.89	0.11	0
Roof Bolters, Mining	0.89	0.11	0
TOP 10 OCCUPATIONS WITH LOWEST AUTOMATABILITIES			
	SUBSTITUTION	COMPLEMENTARITY	NEGLIGIBILITY
Athletes and Sports Competitors	0.11	0	0.89
Sales Agents, Financial Services	0.13	0.12	0.75
Respiratory Therapists	0.32	0.05	0.63
Real Estate Brokers	0.26	0.16	0.58
Clergy	0	0.43	0.57
Choreographers	0.06	0.38	0.56
Chiropractors	0.18	0.27	0.55
Producers	0.08	0.38	0.54
Zoologists and Wildlife Biologists	0.27	0.2	0.53
Nuclear Equipment Operation Technicians	0.47	0	0.53

Figure 4.6: Top 10 occupations that have highest and lowest automatabilities.

arrives at seems reasonable. This is due to our methodology which leverages granular task statement data to predict occupation-level risks.

Additionally, our results indicate that a majority of occupations, specifically 603 out of 974 which is 61.8%, face a high risk of complementarities. Conversely, a relatively smaller fraction, constituting 128 occupations or 13.1%, appear to be comparatively safe from impending automation over the forthcoming decades. These conclusions underscore the nuanced complexities underpinning automation risks and the need for more granular data analysis in this domain. Building on our earlier findings of automatability at the occupation level, we bridge the occupation-industry gap utilizing O*NET's detailed occupation-industry mapping. Our results shown as Fig. 4.7 indicate a diverse spectrum of automation vulnerability across sectors. On one end, industries such as "Accommodation and Food Services", "Administrative and Support Services", "Retail Trade", "Mining, Quarrying, and Oil and Gas Extraction", and "Manufacturing" emerge as the sectors most susceptible to automation. These industries comprise occupations with high automatability scores, revealing their high likelihood of experiencing significant changes due to automation.

On the other hand, industries such as "Educational Services", "Arts, Entertainment and Recreation", "Other Services (Except Public Administration)", "Real Estate and Rental and Leasing", and "Health Care and Social Assistance" stand on the less vulnerable end of the automation spectrum. Occupations within these sectors possess low automatability scores, indicating a lower likelihood of their roles being fully automated, largely due to the complexity of tasks or the high level of human judgment, interaction, and creativity required.

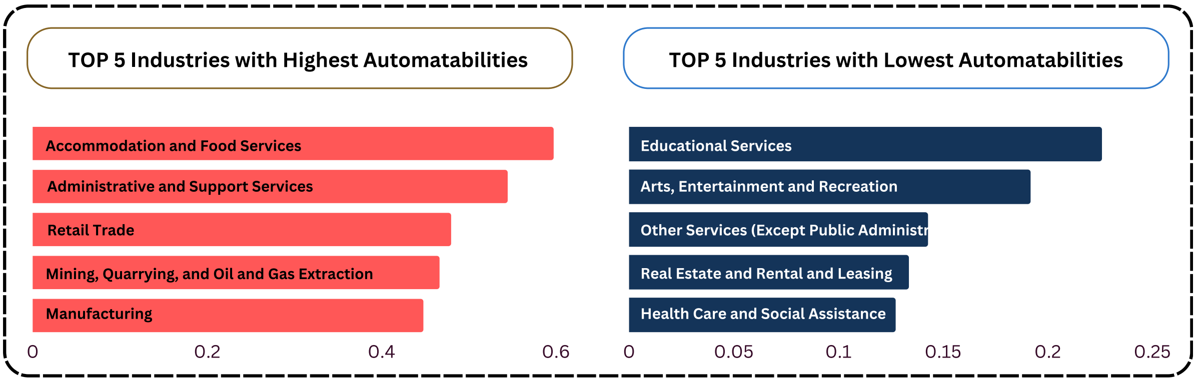


Figure 4.7: Industries with highest and lowest automatabilities.

To validate the results, we leveraged the insights from the McKinsey & Company (2020), and identified 4 sectors showing siginifican susceptibility to automation which are “Accommodation and Food Services", “Manufacturing", “Retail Trade", and “Transportation and Warehousing". These findings align well with our own results.

4.5 Summary

This study presents a unique approach to predict task-level automatability with a BERT-based classifier, utilizing three diverse, public datasets and expert annotations. Rigorous experiments demonstrated our model's efficacy.

In practical application, we applied our model to real-world datasets, providing a comprehensive perspective of occupational automatability. Our findings indicate that approximately 25.1% of occupations within the O*NET database are at substantial risk of automation. Furthermore, our results reveal a diverse spectrum of automation vulnerability across sectors, with industries such as “Accommodation and Food Services", “Administrative and Support Services", “Retail Trade", “Mining, Quarrying, and Oil and Gas Extraction", and “Manufacturing" emerging as the sectors most susceptible to automation.

These findings have significant implications for workers and policymakers. For workers, understanding the susceptibility of their tasks to automation can help them make informed decisions about their career paths and upskilling opportunities. For policymakers, these insights can guide the development of policies and initiatives aimed at managing the transition to an increasingly automated workforce. This could include strategies for retraining and reskilling workers, as well as measures to support industries and regions that are particularly vulnerable to automation.

In conclusion, our research provides a robust and effective approach to predicting task-level automatability, offering valuable insights for policymakers, educators, and workers. Understanding how automation may affect tasks, occupations, and industries enables more effective preparation for the future of work.

CHAPTER 5

ONTOLOGY-ANCHORED AND CAPABILITY-ALIGNED FRAMEWORK FOR AI EXPOSURE ASSESSMENT

This chapter presents the third work, which explores the assessment of task and occupational exposure to artificial intelligence through a novel methodological framework. Building upon insights and methodologies established in earlier chapters, this paper introduces an ontology-anchored and capability-aligned approach that leverages advanced large language models (LLMs) to align detailed occupational tasks with dynamically evolving AI capabilities. By doing so, the work contributes significantly to our understanding of how AI is likely to influence labor markets, identifies occupations and tasks most exposed to AI, and offers practical tools for policymakers and industry stakeholders. The findings presented herein reflect a synthesis of empirical rigor, innovative methodological integration, and responsiveness to contemporary advancements in artificial intelligence.

5.1 Overview

This chapter builds directly on the conceptual and methodological foundations established in the previous two chapters, while extending them in scope and perspective. The earlier chapters focused on explicit automation risks at the task and occupation levels, distinguishing between substitution and complementarity effects using supervised learning frameworks (AOC-GCN and the BERT-based classifier) grounded in expert labels and predefined task features. These approaches were designed to answer whether and how specific tasks are likely to be replaced or augmented by automation, providing interpretable and empirically validated risk estimates. In contrast, the present chapter adopts a more agnostic and forward-looking perspective by shifting from categorical risk labels toward a continuous exposure framework that directly links job tasks to evolving AI capabilities. Rather than relying on cognitive ability taxonomies, fixed rubrics, or human labeling of automation outcomes, this chapter aligns fine-grained job tasks with AI tasks through semantic matching and weights them by dynamic measures of AI progress and research momentum. Conceptually, this represents a progression from outcome-oriented risk prediction (replacement vs. complementarity) to capability-driven exposure measurement, allowing the framework to capture multiple channels of technological impact—including substitution, augmentation, recombination, and latent future effects—within a unified index. Together, the three chapters form an integrated research trajectory: the first two establish task-level automation risks under explicit assumptions, while this chapter generalizes the analysis by grounding exposure in observable AI capability trajectories, thereby unifying and extending the thesis's approach to measuring AI's labour-market effects.

This study develops a principled framework for measuring AI exposure at the task and occupation level that is both ontology-anchored and capability-aligned. While prior research has provided valuable insights into automation risk, most approaches have relied on coarse occupational categories, static rubrics, or outdated benchmarks, resulting in inconsistent and incomplete estimates of how advances in AI affect work. In particular, existing measures often overlook the fine-grained structure of job tasks, fail to capture recent breakthroughs in AI, and struggle to integrate heterogeneous performance metrics across domains such as vision, language, and control.

To address these challenges, we build on O*NET's detailed task statements and the Intelligence Task Ontology (ITO), a curated hierarchy of AI tasks and benchmarks. Each job task is decomposed into its *action* and *purpose*, then semantically aligned with AI capabilities using large language model–assisted matching. This produces a structured job task–AI matrix that captures nuanced correspondences beyond keyword overlap. On top of these alignments, we construct a multi-dimensional exposure index that incorporates three key elements: task significance within occupations, the slope of AI performance improvement (Gain) as a measure of technological momentum, and research Popularity as a proxy for attention and resources. Together, these form the Gain–Popularity Score (GPS), which enables exposure measures to reflect both the current frontier and the dynamics of AI progress.

The framework contributes in three main ways. First, it provides transparency by tracing occupation-level exposure back to specific job task–AI links. Second, it delivers a more forward-looking measure by integrating dynamic progress indicators rather than relying solely on historical data. Third, it introduces a simulation capacity, allowing researchers to explore counterfactual scenarios such as accelerated breakthroughs in individual subfields. Empirical results validate the framework against expert annotations and confirm that it captures complementary dimensions of AI exposure overlooked by prior indices. Distinctive patterns emerge, with higher exposure in education-related and skilled trade occupations, and lower exposure in tasks requiring interpersonal judgment or dexterity.

Overall, this study advances the measurement of AI's labor market impact by uniting semantic task decomposition, ontology-based alignment, and dynamic capability tracking. The resulting AI Exposure Index provides researchers, policymakers, and practitioners with an interpretable and simulation-ready tool for monitoring how technological change intersects with workforce tasks, informing reskilling strategies and long-term planning in an era of rapid AI progress.

5.2 An Ontology-anchored and Capability-aligned Framework

We fill prior efforts gaps by presenting a new framework for measuring occupational AI exposure that is ontology-anchored and capability-aligned. In contrast to earlier approaches that relied on coarse ability categories, static rubrics, or historical judgments, our method grounds exposure assessment in the Intelligence Task Ontology (ITO) (Blagec et al., 2022)—a large-scale, expert-curated hierarchy that organizes AI tasks, sub-tasks, benchmarks, and performance metrics—and aligns these with the fine-grained job task statements provided by O*NET (National Center for O*NET Development, 2023)—a comprehensive taxonomy of U.S. occupations that provides detailed breakdowns of occupation titles, job task statements, and task significance. Each job task is then decomposed into two elements—the *action* performed and its intended *purpose*—which creates a structured representation for alignment with AI capabilities. Using a combination of rule-based procedures and large language model–assisted semantic reasoning, we match these action–purpose pairs with the most relevant AI tasks in the ontology. This process allows us to capture nuanced correspondences beyond surface-level keyword overlap. For example, decomposing "transcribing meeting minutes" into action: transcribe speech and purpose: produce written text aligns it with speech recognition and transcription tasks. The outcome is a fine-grained job task–AI matrix in which each link indicates the degree to which AI can perform a given task. We then aggregate these links into task and occupation-level exposure indices. Crucially, our aggregation is multi-dimensional: in addition to task significance, we incorporate measures of technological progress which is the slope of improvement over time to capture the pace of advancement while mitigating metric-specific distortions, and research popularity as a proxy for the attention and resources devoted to each capability. We term this composite measure as Gain–Popularity Score (GPS). By anchoring AI exposure in a formal ontology and incorporating these dynamic dimensions, our framework produces measures that are interpretable, precise, and forward-looking. Beyond quantifying exposure, it also supports simulation exercises that explore counterfactual scenarios of AI progress, as we demonstrate through case studies on post secondary teaching occupations and a robotics shock.

5.3 Methodology

To construct a consistent and interpretable measure of AI exposure to artificial intelligence (AI), we develop an ontology-anchored and capability-aligned framework. Our approach builds on recent advances in linking technological progress to labor market tasks (Eloundou et al., 2023; Felten et al., 2019; Tolan et al., 2021), while addressing key limitations of earlier studies. Specifically, we adopt a three-stage methodology: (1) curating and structuring AI capabilities within an ontology that reflects both technological advances and research popularity; (2) aligning job tasks with AI tasks through a rule-based procedure supported by large language models (LLMs); and (3) aggregating task-level alignment scores into an occupation-level exposure index. Together, these components provide granularity at the task level and comparability across occupations, enabling a systematic assessment of AI exposure in the workforce.

Fig 5.1 provides an overview of the proposed methodology and illustrates how the three main components of the framework correspond to Sections 3.1–3.3. Section 3.1 (Data Collection) supplies the two foundational inputs shown on the left of the figure: job task data from the O*NET database, including task statements and task significance measures, and AI task data from the Intelligence Task Ontology (ITO), including task descriptions, benchmark results, and research metadata. These inputs feed into the Gain and Popularity Construction block, where benchmark performance improvements and research activity are combined to compute the AI Task Gain–Popularity Score (GPS), detailed in Section 3.3.1. Section 3.2 (Job Task–AI Alignment Pipeline) corresponds to the central alignment component of the framework, where job task statements are decomposed into action–purpose representations and semantically matched to AI tasks using LLM-based few-shot prompting, producing aligned job–AI task pairs. Finally, Section 3.3 (AI Exposure Index Construction) maps to the integration and aggregation components on the right of Figure 1: alignment scores are combined with AI Task GPS to generate task-level exposure scores, which are then aggregated using task significance weights to produce occupation-level exposure indices. Together, these stages form a coherent pipeline that links raw task descriptions and dynamic AI capabilities to interpretable measures of job and occupational exposure, ensuring consistency between the framework visualization and the methodological subsections.

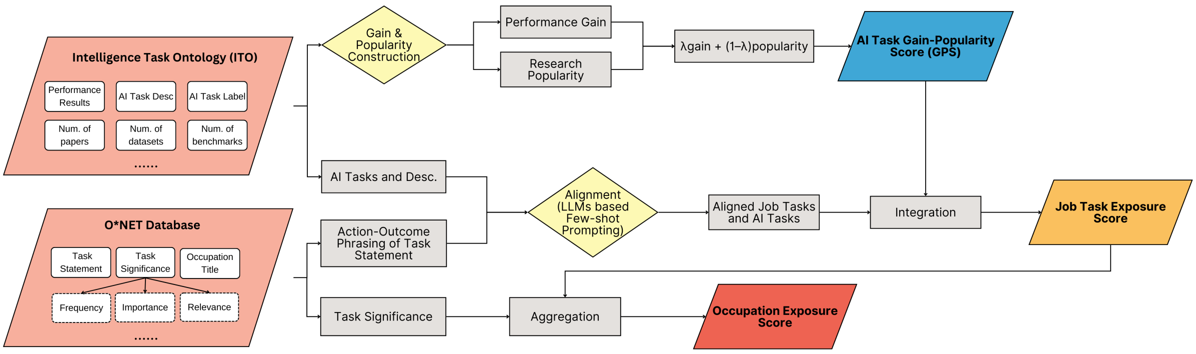


Figure 5.1: Framework of the methodology.

5.3.1 Data Collection

Our analysis draws on the public O*NET 28.1 database (National Center for O*NET Development, 2023), which provides detailed information for 1,016 Standard Occupational Classification (SOC) occupations. Each occupation entry includes its SOC code, title, and major group affiliation, enabling analyses at multiple levels of granularity.

We focus on *Task Statements*, the occupation-specific units of work expressed as concise action phrases (e.g., "*Calibrate sensors to ensure accurate readings*"). Each Task Statement is evaluated on three incumbent-based rating scales that capture its role in everyday work: *frequency* (0–7), *importance* (0–5), and *relevance* (0–5). Together, these ratings allow us to assess both the intensity of task performance and its centrality within an occupation.

For the analytic sample, we retain only tasks with complete ratings and occupations with at least ten valid Task Statements. This yields 18,947 rated Task Statements covering 1,009 occupations, each linked to its SOC code and major group, providing a rich basis for both occupation-level and group-level analysis.

To characterise the landscape of AI tasks and capabilities, we draw on the Intelligence Task Ontology (ITO) (Blagec et al., 2022). ITO is a manually curated, ontology-based knowledge graph that integrates benchmark results from *Papers with Code* with a hierarchical taxonomy of AI tasks, datasets, models, and evaluation metrics. The ontology is encoded in RDF/OWL, enabling SPARQL queries and automated inference, and is organised into 16 top-level task classes (e.g., *Natural Language Processing*, *Vision*, *Graph Processing*) that expand into more than 1,000 fine-grained AI tasks.

The dataset snapshot used in this study spans research published between 2000 and 2022, comprising 7,649 peer-reviewed or preprint papers and 26,495 benchmark results. Each benchmark entry links an AI task to its dataset, model or system, and an associated performance metric (from a curated set of 1,995 measures such as *accuracy* and F_1 *score*). Table 5.1 summarises key characteristics of the ITO dataset.[1]

For analysis, we aggregate results at the AI task level and construct *gain–popularity scores (GPS)*, which capture changes in research popularity and performance over time. Although the ITO defines 16 broad top-level process classes and more than 1,000 fine-grained tasks, this taxonomy is either too coarse (e.g., *Natural Language Processing*) or too fragmented (*e.g., 3D Point Cloud Classification*) for systematic alignment with job tasks. Prior work shows that very broad categories can mask important variation across subfields, while overly granular benchmarks often fail to map meaningfully to job tasks (Martínez-Plumed et al., 2020; Tolan et al., 2021). To balance interpretability with effectiveness, we collaborated with domain experts to reorganize the fine-grained ITO tasks into 107 intermediate 'parent tasks' and removed overly general entries such as reinforcement learning[2]. For each parent task's capability, we take the maximum score among its constituent fine-grained. This aligns with the common practice of using peak benchmark performance as a proxy for frontier capability (Felten et al., 2019), while also recognizing that progress in many AI tasks is interdependent. These parent-level performance trajectories provide the basis for the GPS used to construct our AI exposure index. Throughout this paper, we use the term *AI task* to refer to these reorganized parent tasks.

We incorporate employment and wage information from the Occupational Employment and Wage Statistics (OEWS) series (formerly OES), released annually by the U.S. Bureau of Labor Statistics. The dataset spans eleven public-use waves under a consistent NAICS/SOC scheme, covering reference years 2012–2022 (U.S. Bureau of Labor Statistics, 2024). For each occupation–year cell, the OEWS reports total employment, mean and median hourly wages, and long-term employment projections. To integrate OEWS with the

[1]All resources, including the OWL file, SPARQL endpoints, and example Jupyter notebooks, are publicly available at https://github.com/OpenBioLink/ITO.

[2]The complete set of resources for this paper is available in the code repository https://github.com/listenerxdw/ai-exposure.git

Table 5.1: Descriptive statistics of ITO (v2.0) used in this study

Metric	**Count**
Total RDF triples (edges)	685 560
Classes (all)	9 037
of which AI process classes	1 100
Individuals (nodes)	50 826
Parent process classes	16
Data properties (performance metrics)	1 995
Benchmark datasets	3 633
Benchmark results	26 495
Distinct papers	7 649
Publication years covered	2000–2022

O*NET task data, we follow the BLS-recommended *SOC crosswalk* [3].

These measures are used in two ways. First, employment counts and wage levels serve as *control variables* in all empirical specifications, allowing us to account for baseline labour-market conditions when estimating the relationship between AI exposure and outcomes. Second, as detailed in the validation analysis, observed wage and employment patterns provide an external benchmark to assess whether our exposure index aligns with economic signals in the labour market.

5.3.2 Job Task–AI Alignment Pipeline

Our goal is to automatically match each free-text job task statement to the most relevant AI task. This matching relies on the fact that job tasks and AI tasks can both be represented in terms of an *action* and a *purpose*. Job task statements, however, are often written in natural language with varying levels of detail, making it difficult to directly compare them to AI task categories. We therefore decompose each job task into two components: the *action (what is done)* and the *purpose (why it is done)* (Table 5.2). In contrast, AI tasks do not require decomposition: their titles already denote the action or purpose (e.g., *machine translation*, *image classification*) and their ontology-provided descriptions clearly state the purpose. After decomposing both job tasks and AI tasks into actions and purposes, we apply semantic matching to assess the degree of alignment. Each comparison is assigned one of three judgment categories: *Fully matched*, *Partially matched*, or *Mismatched*. A *Fully matched*

[3] https://www.bls.gov/emp/documentation/crosswalks.htm

case occurs when the AI task captures either the action or the purpose of the job task in a comprehensive way. A *Partially matched* case arises when a job task decomposes into multiple actions and purposes, but the AI task corresponds to only one component. A *Mismatched* case indicates that the job task and the AI task are semantically different, with no meaningful overlap in action or purpose. This structured scheme allows us to measure alignment consistently and to distinguish between different coverage of job tasks by AI capabilities. This approach is consistent with prior frameworks that emphasize decomposing occupations or tasks into structured elements to improve comparability and alignment (J. Liu et al., 2020; L. Zhang et al., 2021). For example, J. Liu et al. (2020) decomposes occupation titles into *functions* and *responsibilities* to support standardized crosswalks. To operationalize this alignment, we design a multi-stage pipeline that combines semantic retrieval with large language model (LLM) reasoning.

Table 5.2: Sample alignment of job tasks with decomposition and AI tasks

Job Task	**Decomposition**	**AI Task (Title/Description)**
Analyze sales data to forecast demand	**Action:** analyze sales data **Purpose:** forecast demand	**Time-series forecasting**: Predict future values in temporal data using statistical or ML models.
Write emails to respond customer support	**Action:** Write emails **Purpose:** resolve customer issues	**Natural language generation**: Generate coherent, context-aware text responses.
Classify medical images for diagnosis	**Action:** classify medical images **Purpose:** support diagnosis	**Computer vision**: Detect and classify objects in images, e.g., tumors in scans.

Our approach frames the alignment problem as a text classification task using large language models (LLMs) with in-context demonstrations (Brown et al., 2020; Luo et al., 2024; Rubin et al., 2022). To establish reliable training examples, we collaborated with domain experts to annotate 300 job–AI task pairs, producing a gold-standard dataset with labels of *Fully matched*, *Partially matched*, or *Mismatched*. Each pairs a decomposed job task (action and purpose) with a candidate AI task and its description. Then for new job-AI task pair, we first encode the texts into a semantic vector using Sentence-BERT (Reimers & Gurevych, 2019) and compute cosine similarity with the embeddings of all annotated pairs. The top-k most similar examples (typically $k = 3$–5) are then retrieved and used as demonstrations for in-context prompting, allowing the LLM to classify the alignment by reasoning over comparable, previously solved cases.

Using the retrieved examples, we construct a tailored prompt that incorporates: (1) the new job task's action and purpose, (2) a short list of demonstration pairs with their associated match labels, and (3) an instruction to classify the alignment with AI tasks. The prompt is designed to follow a structured reasoning process inspired by chain-of-thought prompting (Brown et al., 2020; Wei et al., 2022). The model is asked to extract diagnostic clues from the action–purpose pair, reason about their alignment with the candidate AI task descriptions, and then assign a classification label: *Fully matched*, *Partially matched*, or *Mismatched*. By including solved demonstrations in the prompt, the model is guided toward consistent reasoning and classification. Figure 5.2 illustrates the contrast between standard prompting approaches and our tailored pipeline.

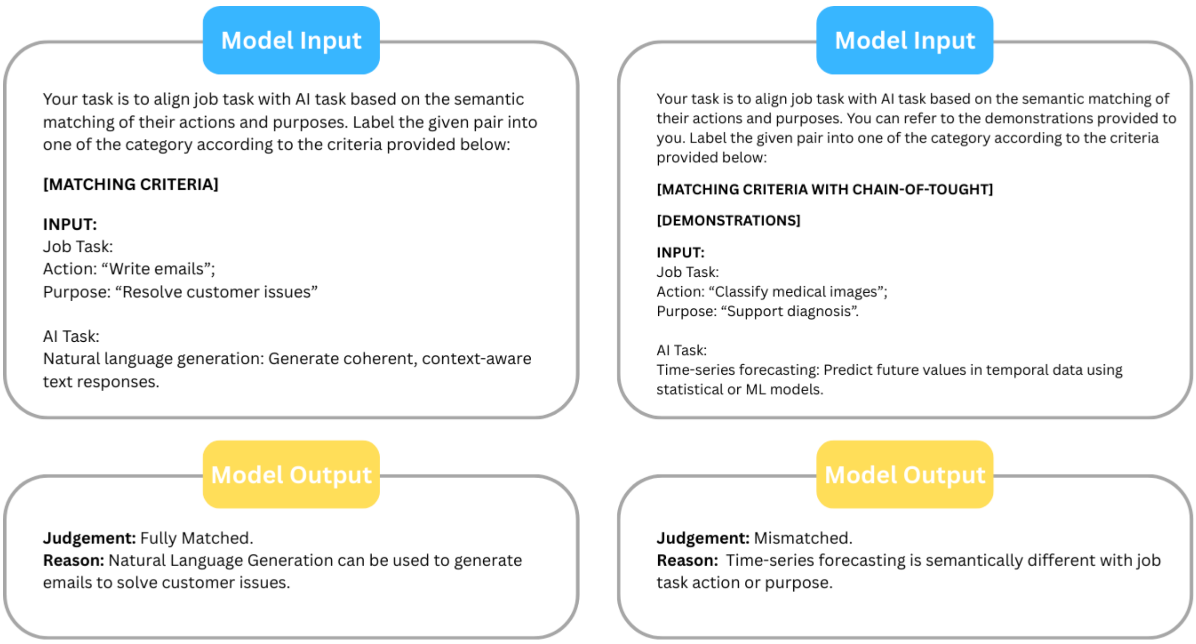


Figure 5.2: Standard prompts vs. tailored prompts. Tailored prompts integrate demonstrations and a structured chain-of-thought process.

Once the pipeline is established, the system operates without further human labeling: for each new job task, it retrieves comparable examples and queries the LLM, making the procedure effectively few-shot. This enables thousands of tasks to be aligned quickly and consistently, supporting large-scale analysis of workforce job tasks. The pipeline integrates recent advances in semantic retrieval and LLM prompting, while also accounting for evidence that model performance is sensitive to instruction phrasing (X. Sun et al., 2023). By combining structured task decomposition, semantic search, ontology grounding, and chain-of-thought reasoning, our method provides a robust and interpretable mechanism for aligning job tasks with AI tasks, yielding match labels that scale across occupations.

5.3.3 AI Exposure Index Construction

We now describe the construction of the AI Exposure Index, which proceeds in three stages. First, we derive a *gain–popularity score* (GPS) for each AI task, combining evidence of performance improvements on benchmarks with indicators of task complexity and research popularity. Second, we map AI tasks to job task statements and compute task-level exposure as an alignment-weighted average of the corresponding AI task GPS. Finally, we aggregate these values into occupation-level scores that capture the overall extent to which the core tasks of an occupation are aligned with areas of active AI progress. This staged approach enables us to link frontier AI capabilities with the detailed structure of workforce, yielding a normalized exposure score that is both interpretable and comparable across occupations.

We aim to construct a *gain–popularity score* (GPS) that enables systematic comparison of performance progress across diverse AI tasks. However, this effort faces several challenges. Benchmark datasets differ widely in their evaluation metrics (e.g., *Accuracy*, F_1, *BLEU*, and *Perplexity*)-each with distinct scales and interpretations. Some metrics improve as their values increase (e.g., *Accuracy*), while others improve as they decrease (e.g., *MAE*), creating polarity inconsistencies. Improvements near performance ceilings also require disproportionately greater innovation but appear numerically small, which can understate true progress. Finally, benchmarks differ in inherent difficulty: some tasks are evaluated using only a single, relatively simple dataset, while others are tested on a broad range of datasets and metrics, making raw performance gains non-comparable across tasks. These challenges have been highlighted in prior efforts to link AI progress with labor outcomes, such as (Felten et al., 2019; Tolan et al., 2021).

To address these issues, we normalize all benchmark metrics to a $[0,1]$ scale, inverting those where "lower is better" so that higher values always represent improvement. We then apply a logarithmic compression,

$$f(y) = \log_{10}(1+9y),$$

which reduce the influence of near-ceiling gains. Instead of relying on raw benchmark levels—which can be distorted by initial conditions, metric idiosyncrasies, or ceiling effects—we focus on the *slope of improvement over time*. For each task–dataset–metric series, we retain the first and last reported performance values and compute a yearly gain as the slope of the compressed score trajectory. Specifically, for the first and last measurements, $(y_{\text{first}}, t_{\text{first}})$ and $(y_{\text{last}}, t_{\text{last}})$, we compute the performance change:

$$\Delta y = y_{\text{last}} - y_{\text{first}}, \qquad \Delta t = t_{\text{last}} - t_{\text{first}}$$

then we apply logarithmic compression and fit a linear trend in the scaled metric over time, yielding the yearly gain (slope) for that metric:

$$\rho = \frac{f(y_{\text{last}}) - f(y_{\text{first}})}{\Delta t}$$

Aggregating these gains across all benchmarks k for task a yields a weighted average:

$$g_a = \frac{\sum_k \rho_k}{\sum_k \Delta t_k}$$

which gives a normalized gain $g_a \in [0,1]$.

This approach, also used by Felten et al. (2019), provides a more robust signal of progress by emphasizing the rate of change rather than absolute values; also, Hernandez and Brown (2020) quantify progress via rates of change (doubling times/slopes) rather than absolute values. Together, these justify our use of slopes to compare across tasks and metrics.

To account for variation in benchmark difficulty, we weight each task's gain by a *complexity factor*, defined as the logarithmic average of the number of distinct datasets and metrics used to evaluate the task. This adjustment credits tasks tested across richer and more diverse benchmarks with greater robustness, while those based on narrow or single-metric evaluations contribute less. Formally, let $n_{d,a}$ and $n_{m,a}$ denote the number of distinct datasets and metrics associated with AI task a, and define

$$c_a \;=\; \tfrac{1}{2}\big[\ln\!\big(1+n_{d,a}\big) \;+\; \ln\!\big(1+n_{m,a}\big)\big].$$

If g_a denotes the normalized, polarity-adjusted performance gain for task a, the complexity-adjusted gain is given by

$$h_a = g_a \cdot c_a$$

which is then rescaled to the unit interval $[0,1]$ across all task denoted as $\tilde{h}_a$.

Next, to capture research popularity, we integrate publication counts linked to each task, log-transformed and rescaled for comparability. Let $n_{p,a}$ be the number of AI papers (or models) linked to task a. Define $p_a = \ln\!\big(1+n_{p,a}\big)$ and normalize via min–max to $[0,1]$:

$$\tilde{p}_a \;=\; \frac{p_a - \min_{a'\in\mathscr{A}} p_{a'}}{\max_{a'\in\mathscr{A}} p_{a'} \;-\; \min_{a'\in\mathscr{A}} p_{a'}} \;\in\; [0,1].$$

Finally, combine the two signals with an adjustable weight $\lambda \in [0,1]$ (default $\lambda = 0.5$ in our experiments) to obtain the Gain–Popularity score (GPS):

$$\text{GPS}_a \;=\; \lambda\,\tilde{h}_a \;+\; (1-\lambda)\,\tilde{p}_a \qquad \in [0,1]. \tag{5.1}$$

Intuitively, $\tilde{h}_a$ captures performance progress adjusted for task complexity, while $\tilde{p}_a$ captures research popularity; their convex mix yields a stable, comparable exposure signal across AI tasks.

We next construct task-level exposure scores by combining the results of the alignment pipeline with AI Task GPS. Each O*NET task statement j may align with one or more AI tasks. For each matched pair, we assign a *matching weight* ω_{ja} based on the alignment category: $\omega_{ja} = 1$ if the match is *Full*, and $\omega_{ja} = 0.5$ if it is *Partial*. When a job task aligns with multiple AI tasks, we compute the weighted average of their matched AI Tasks GPS values:

$$\text{Exposure}_{task} = \frac{\sum_{a \in \mathscr{A}_j} \omega_{ja} \, \text{GPS}_a}{\sum_{a \in \mathscr{A}_j} \omega_{ja}},$$

where $\mathscr{A}_j$ is the set of AI tasks aligned to task j. This ensures that task-level exposure reflects both the strength of the alignment and the research momentum of the corresponding AI tasks.

For occupation-level exposure scores, we begin by assessing how central each job task is within its occupation. To do so, we combine the O*NET *frequency*, *importance*, and *relevance* ratings for each task into a single composite measure that we denote as α_j, the *Significance* of task j. Each rating is normalized to the unit interval and then averaged, so that tasks performed more often, considered more important, and more representative of the occupation receive higher significance scores.

The occupational exposure contributed by task j is then obtained by multiplying its significance α_j with its task-level exposure Exposure_{task}. This ensures that tasks strongly associated with active areas of AI progress, and central to the occupation, weigh more heavily in the overall exposure.

Finally, the occupation-level exposure is computed by aggregating across all tasks in occupation o:

$$E_o = \sum_{j \in o} \alpha_j \cdot \text{Exposure}_{task},$$

and normalizing by the maximum observed value across all occupations for comparability:

$$\text{Exposure}_{occu} = \frac{E_o}{\max_{o'} E_{o'}} \in [0, 1].$$

In summary, occupation-level exposure scores combine three ingredients: (1) the gain–popularity momentum of AI tasks, (2) the alignment strength between AI tasks and job tasks, and (3) the significance of those tasks within occupations. This construction produces an interpretable, normalized index that highlights which occupations are most exposed to current trajectories of AI research and capability development.

5.4 Experiments

Results and Patterns We selected GPT-4o as the large language model for this study. From our experiments, we generated a matrix linking AI tasks to job tasks with corresponding alignment scores. In total, 106 AI tasks aligned with at least one job task, resulting in 5,314 matches out of 16,939 job tasks. These covered 849 of 1009 occupations, each with measurable exposure values. Table 5.3 reports descriptive statistics for the main measures across the AI task, job task, and occupational levels. The AI task measures capture different dimensions of technological progress: Gain reflects performance improvements in benchmark tasks, Popularity proxies research attention, Breadth indicates the number of job tasks connected to a given AI task, and Depth captures the average significance of the affected job tasks. On the labor side, Job Task Significance measures how central a task is within occupations, while Job Task Breadth denotes the variety of AI tasks mapped to a single job task. Finally, Job Task Exposure and Occupation Exposure provide aggregated measures of exposure, weighted by AI task impact and task significance.

Table 5.3: Descriptive statistics of key measures

	Mean	Std	Min	25th Perc.	Median	75th Perc.	Max
AI Task Gain	0.38	0.25	0.00	0.19	0.37	0.59	1.00
AI Task Popularity	0.47	0.25	0.00	0.32	0.47	0.60	1.00
AI Task Breadth	0.09	0.16	0.00	0.01	0.03	0.13	1.00
AI Task Depth	0.42	0.20	0.00	0.29	0.41	0.51	1.00
Job Task Significance	0.67	0.14	0.03	0.59	0.69	0.78	1.00
Job Task Breadth	0.05	0.09	0.00	0.00	0.00	0.08	1.00
Job Task Exposure	0.46	0.23	0.00	0.25	0.48	0.63	1.00
Occupation Exposure	0.25	0.18	0.00	0.12	0.20	0.34	1.00

Overall, the statistics show notable variation across measures. AI Task Gain and Popularity have similar means (0.38 and 0.47, respectively), but the wide ranges and standard deviations indicate heterogeneous progress across AI domains. Breadth measures (AI Task Breadth and Job Task Breadth) are strongly right-skewed, with most tasks linking to very few counterparts and a small number exhibiting extensive connectivity. Job Task Significance is consistently high (mean = 0.67), underscoring that many tasks included in the mapping are central to occupational functioning. Exposure scores decrease as we move up the aggregation ladder, from job tasks (mean = 0.46) to occupations (mean = 0.25), reflecting the smoothing effect of aggregation.

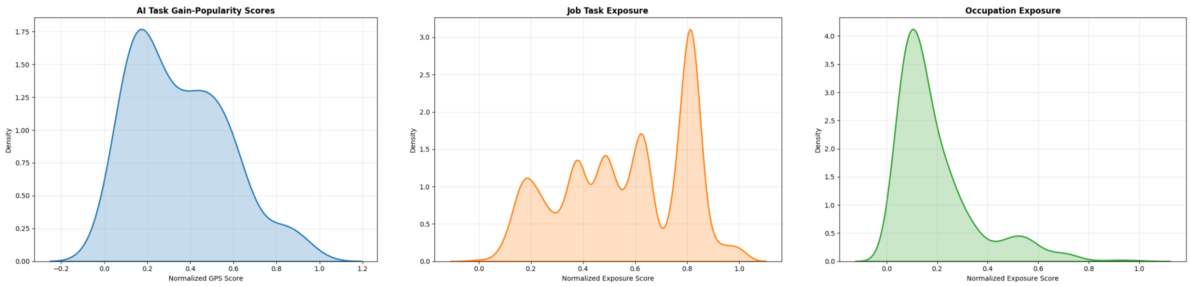


Figure 5.3: Distributions of AI task GPS scores (left), job task exposure (middle), and occupational exposure (right). The comparison shows how heterogeneity at the AI task level smooths as exposure is aggregated to job tasks and occupations.

Figure 5.3 illustrates the distributional patterns across levels. AI task GPS scores (left panel) are right-skewed, peaking around 0.2–0.3. Most AI tasks therefore exhibit moderate impact, while a few—such as those in natural language processing or computer vision—reach high combined scores of gain and popularity.

The distribution of job task exposure (middle panel) is multimodal, with several peaks across the range. This pattern reflects clustering of tasks into distinct exposure groups, shaped by the accumulation of multiple AI task influences on a single activity. Tasks directly aligned with well-established AI capabilities (e.g., classification, translation, recognition) appear at the high-exposure end, while tasks requiring embodied interaction or complex social reasoning remain less exposed.

At the occupational level (right panel), the distribution is smoother and more concentrated toward the lower end. Most occupations exhibit modest exposure, with relatively few showing high values. This attenuation underscores how heterogeneity observed at the AI task level becomes progressively diluted when aggregated through job tasks to occupations, offering a layered view of how AI capabilities diffuse across the labor market.

To understand how advances in AI capabilities translate into potential labor-market relevance, we examine three dimensions of AI task characteristics: performance gain, research popularity, and their joint impact captured by the Gain–Popularity Score (GPS). Figure 5.4 presents the joint distribution of gain and popularity scores. The Pearson correlation is $r = 0.63$, indicating a strong positive association: AI domains that show rapid technical improvement also tend to attract sustained research attention. The quadrant-based categorization highlights that while many tasks achieve balanced progress and popularity, others remain underexplored despite significant technical advances, or conversely, popular but with limited measurable progress.

Figure 5.5 presents the top-10 and bottom-10 AI tasks ranked by normalized GPS. The

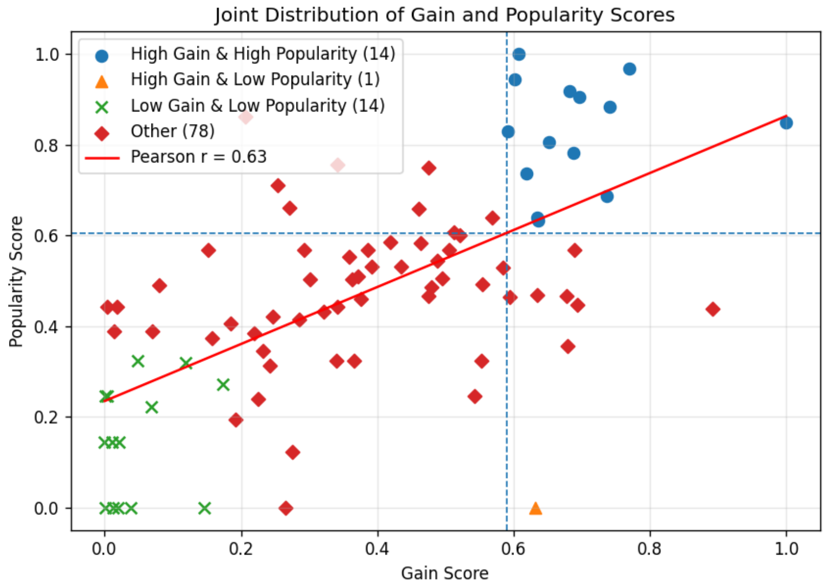
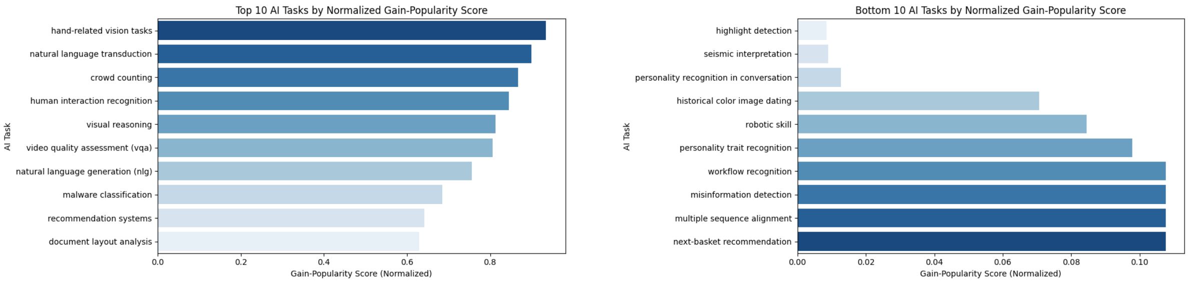


Figure 5.4: Joint distribution of Gain-Popularity scores, with data points categorized into quartile-based quadrants. The red line shows the Pearson correlation between gain and popularity.

highest-scoring tasks reflect areas where technical breakthroughs align with substantial research momentum. *Hand-related vision tasks* (0.93) lead the ranking, driven by progress in gesture recognition, hand pose estimation, and robotics/augmented reality applications. *Natural language transduction* (0.90) and *crowd counting* (0.87) follow closely, while other vision-related tasks—such as visual reasoning, video quality assessment, and human interaction recognition—dominate the upper tier, underscoring computer vision as a highly active research domain. *Natural language generation* (0.76) reflects the rapid rise of large language models, while applied domains including *malware classification* (0.68) and *recommendation systems* (0.64) highlight strong industrial relevance and continued algorithmic refinement. By contrast, the lowest-ranked tasks represent either narrow applications or areas with limited momentum. *Highlight detection* and *seismic interpretation* (both $\approx$ 0.009) show minimal benchmarking and niche applicability. *Personality recognition in conversation* (0.013) reflects an emerging yet technically challenging domain. Other low scorers, such as *historical color image dating*, *workflow recognition*, and *multiple sequence alignment*, correspond either to highly specialized domains or to fields where research progress has plateaued. Interestingly, tasks with clear societal or commercial significance—such as *misinformation detection* and *next-basket recommendation*—also appear in the bottom tier, suggesting unresolved technical barriers or insufficient research focus relative to their potential impact.



((a)) Top-10 AI tasks with highest GPS ((b)) Bottom-10 AI tasks with lowest GPS

Figure 5.5: Top and bottom AI tasks ranked by GPS values.

Beyond overall impact, it is important to consider how widely and how centrally AI tasks map onto human work. Figure 5.6 combines the three measures: GPS (x-axis), task depth (y-axis), and breadth (bubble size and color). The plot reveals several noteworthy patterns. First, GPS values are widely dispersed, with most tasks concentrated in the lower to mid-range but a limited set of domains achieving exceptionally high impact. Second, the scatter does not reveal a strong linear relationship between GPS and task depth, that is to say, high-impact AI tasks are distributed across both low- and high-depth regions though the most influential ones tend to align with mid- to high-significance job tasks. Third, task breadth is highly skewed: while the majority of AI capabilities influence only a small number of job tasks, a select group of domains, such as *visual reasoning* and *general imaging tasks*, exhibit broad applicability, thereby amplifying their labor-market relevance.

The joint consideration of breadth, depth, and GPS highlights a set of "transformative" AI domains. These include *visual reasoning*, *object detection and understanding*, *robotic grasping*, *autonomous driving*, and *dialog systems*, which stand out as simultaneously broad, central to occupations, and rapidly advancing. The dominance of vision and language tasks in this category reflects both the abundance of benchmark data and the maturity of evaluation ecosystems in these fields, whereas other domains remain less developed and underrepresented.

Appendix Tables 5.4 and 5.5 report the ten job tasks with the highest and lowest AI exposure scores, respectively, illustrating the uneven distribution of exposure across the task spectrum. High-exposure tasks are concentrated in activities involving information processing, translation, transcription, and structured communication. Examples include verifying technical translations, transcribing proceedings, compiling records, and collecting survey data. Many of these tasks are tightly linked to natural language processing capabilities (*checking translations*, *compiling information for translation*, *instructing clients in communication techniques*), reflecting the rapid progress of language technologies. Other

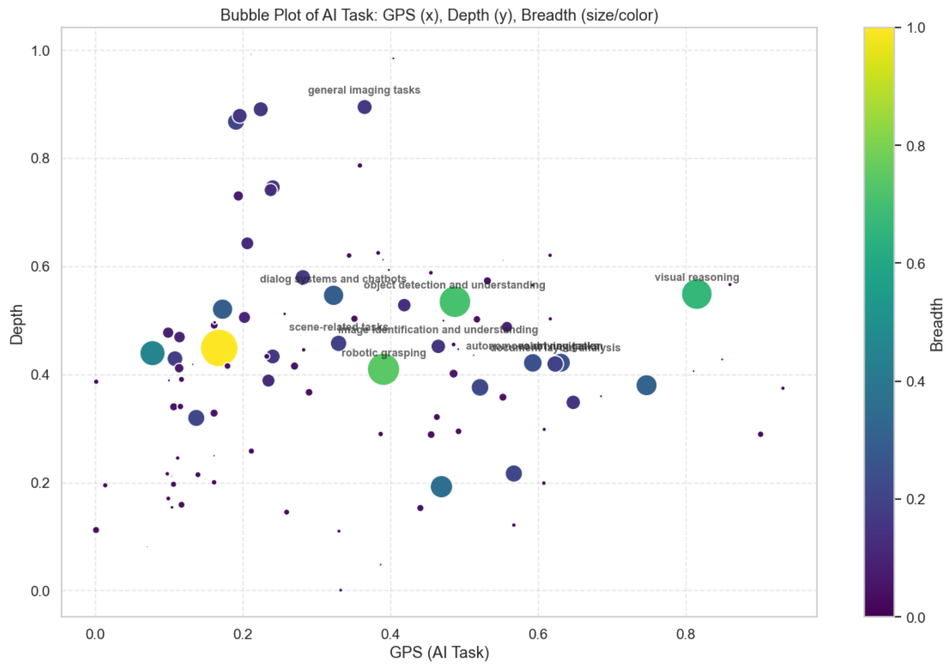


Figure 5.6: Bubble plot of AI tasks by GPS (x-axis), depth (y-axis), and breadth (color and bubble size). The plot illustrates the skewed distribution of breadth, wide variation in depth, and the presence of both narrow–deep and broad–high-intensity tasks.

highly exposed tasks, such as *keeping transaction records* or *taking inmate counts*, align with AI's comparative advantage in systematic, rule-based monitoring. Notably, tasks involving communication with external stakeholders (*communicating with investors, contacting corporate representatives or community leaders*) also rank highly, suggesting that AI is increasingly applied to outward-facing activities that support professional interaction.

By contrast, the least exposed tasks are concentrated in the geosciences, particularly in oil, gas, and mineral exploration. These include conducting geological surveys, preparing maps, measuring geological characteristics, interpreting core samples, and evaluating the economic viability of drilling sites. Such activities typically require professional, on-site physical presence in field environments, together with specialized instrumentation and the integration of tacit expertise that cannot be easily codified or benchmarked. Even where computational analysis is applied, the scarcity of standardized datasets and the complexity of subsurface phenomena limit AI's current applicability. The clustering of bottom-ranked tasks in extractive industries underscores how data scarcity, physical constraints, and reliance on domain-specific expertise insulate some activities from AI exposure.

Figure 5.7 plots job task exposure against task importance, with color denoting task breadth. Exposure values are widely dispersed, ranging from near zero to one, while impor-

Table 5.4: Top 10 Job Tasks by Exposure

Rank	Job Task
1	Check translations of technical terms and terminology to ensure that they are accurate and remain consistent throughout translation revisions.
2	Compile information on content and context of information to be translated and on intended audience.
3	Transcribe recorded proceedings in accordance with established formats.
4	Collect and analyze data, such as studying old records, tallying the number of outpatients entering each day or week, or participating in federal, state, or local population surveys as a Census Enumerator.
5	Keep records of transactions and of the number of customers entering an establishment.
6	Take, receive, or check periodic inmate counts.
7	Observe group interactions and role affiliations to collect data, identify problems, evaluate progress, and determine the need for additional change.
8	Communicate with stockholders or other investors to provide information or to raise capital.
9	Contact corporate representatives, government officials, or community leaders to increase awareness of organizational causes, activities, or needs.
10	Instruct clients in techniques for more effective communication, such as sign language, lip reading, or voice improvement.

tance scores are more tightly concentrated between 0.5 and 0.8, reflecting that most tasks are rated as moderately to highly important across occupations. The scatter indicates no strong linear relationship between exposure and importance: highly exposed tasks are not exclusively the most important, and low-exposure tasks are not confined to marginal activities. This suggests that AI affects both central and peripheral tasks. Task breadth is highly skewed, with the majority of tasks linked to only a few AI capabilities, while a small subset shows broad connections. Interestingly, these broad tasks cluster around moderate exposure values, implying that being associated with many AI capabilities does not necessarily imply extreme exposure.

Table 5.5: Bottom 10 Job Tasks by Exposure

Rank	Job Task
1	Participate in geological, geophysical, geochemical, hydrographic, or oceanographic surveys, prospecting field trips, exploratory drilling, well logging, or underground mine survey programs.
2	Prepare geological maps, cross-sectional diagrams, charts, or reports concerning mineral extraction, land use, or resource management, using results of fieldwork or laboratory research.
3	Measure geological characteristics used in prospecting for oil or gas, using measuring instruments.
4	Evaluate and interpret core samples and cuttings, and other geological data used in prospecting for oil or gas.
5	Assess the environmental impacts of development projects on subsurface materials.
6	Assess costs and estimate the production capabilities and economic value of oil and gas wells, to evaluate the economic viability of potential drilling sites.
7	Develop plans for oil and gas field drilling, and for product recovery and treatment.
8	Take samples to assess the amount and quality of oil, the depth at which resources lie, and the equipment needed to properly extract them.
9	Select or develop mineral location, extraction, and production methods, based on factors such as safety, cost, and deposit characteristics.
10	Examine maps, deposits, drilling locations, or mines to determine the location, size, accessibility, contents, value, and potential profitability of mineral, oil, and gas deposits.

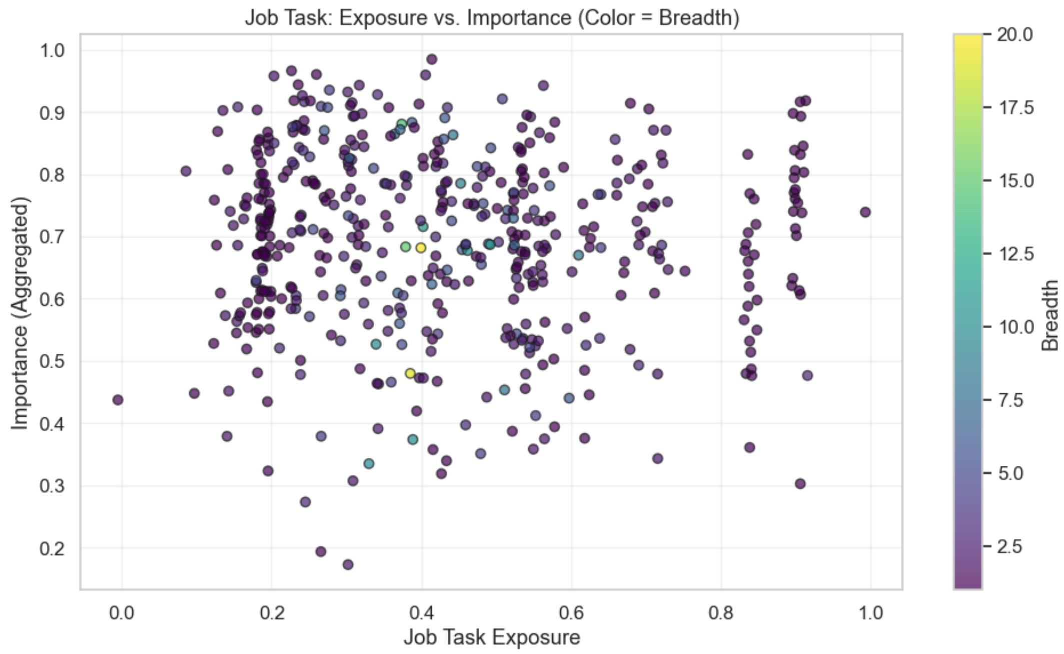


Figure 5.7: Job task exposure against task importance, with color indicating task breadth.

Finally, Appendix Figure 5.8 compares AI exposure across four task categories from D. Autor et al. (2003): routine cognitive, non-routine cognitive, routine manual, and non-routine manual. All categories exhibit wide internal variation, indicating that category membership alone does not determine exposure. Routine and non-routine cognitive tasks display similar patterns, with medians around 0.4–0.5 and broad interquartile ranges, suggesting that both routine and non-routine information-processing tasks are broadly susceptible to AI. Non-routine manual tasks stand out with relatively high median exposure and wide dispersion, reflecting that while some physical tasks remain insulated, others involving navigation or visual inspection are increasingly complemented by AI technologies. Taken together, these findings imply that the traditional routine versus non-routine distinction is less predictive of AI exposure than previously assumed, highlighting the need for more granular, task-level analysis.

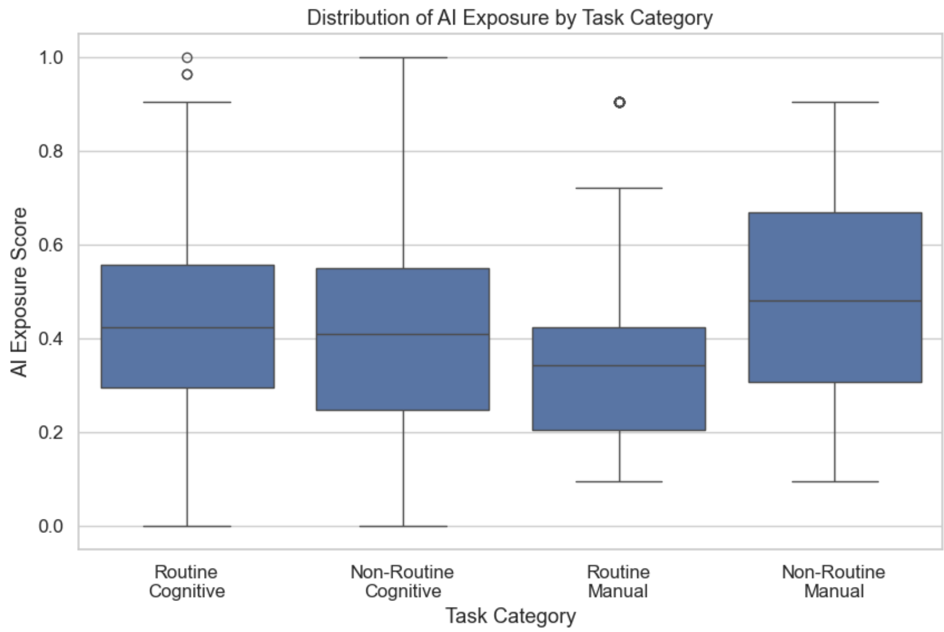


Figure 5.8: Job Task Exposure vs. Category

We next examine AI exposure at the level of entire occupations, highlighting the most and least affected roles according to our exposure index. Tables 5.6 and 5.7 report the top- and bottom-ranked occupations.

The list reveals a diverse set of jobs, spanning trades, technical services, and academic roles. At the very top are *transportation security screeners* and *explosives workers*, whose reliance on visual inspection and detection tasks aligns closely with advances in computer vision. Other high-exposure occupations include *operating engineers*, *payroll clerks*, and

Table 5.6: Top 20 occupations by AI–exposure score

#	**Occupation (O*NET title)**	**Exposure**
1	Transportation Security Screeners	1.00
2	Explosives Workers, Ordnance Handling Experts, and Blasters	0.92
3	Operating Engineers and Other Construction Equipment Operators	0.85
4	Payroll and Timekeeping Clerks	0.83
5	Helpers–Production Workers	0.82
6	Drywall and Ceiling Tile Installers	0.80
7	Film and Video Editors	0.79
8	Purchasing Agents, Except Wholesale, Retail, and Farm Products	0.79
9	Automotive Body and Related Repairers	0.78
10	Interpreters and Translators	0.77
11	Art, Drama, and Music Teachers, Postsecondary	0.76
12	Biological Science Teachers, Postsecondary	0.75
13	Library Science Teachers, Postsecondary	0.74
14	Psychology Teachers, Postsecondary	0.74
15	Nursing Instructors and Teachers, Postsecondary	0.72
16	Heavy and Tractor-Trailer Truck Drivers	0.72
17	Inspectors, Testers, Sorters, Samplers, and Weighers	0.70
18	Rail Yard Engineers, Dinkey Operators, and Hostlers	0.68
19	Engineering Teachers, Postsecondary	0.67
20	English Language and Literature Teachers, Postsecondary	0.67

helpers–production workers, reflecting the spread of AI capabilities across both manual and clerical domains. A notable cluster of postsecondary teaching roles also appears in the top twenty (e.g., *art and music teachers*, *biology teachers*, *library science teachers*), indicating that higher education is not insulated from AI's reach. This pattern underscores that exposure is driven less by education level and more by the specific task content. For example, large language models (LLMs) are increasingly relevant for teaching-related activities such as preparing instructional material or grading, while benchmark progress in perception tasks maps to screening, monitoring, and inspection duties.

In contrast, the bottom of the distribution is populated by occupations that rely heavily on dexterity, context sensitivity, or interpersonal interaction, such as *hand sewers*, *animal trainers*, and *social workers*, where empathy, persuasion, or fine motor skills remain difficult to automate. High-stakes judgment roles (e.g., *compliance managers*, *treasurers*) and creative professions (*fashion designers*) also register low exposure, despite the availability of AI drafting or generative tools (Frey & Osborne, 2013).

It is important to emphasize that the index reflects *technical potential* rather than real-

Table 5.7: Bottom 20 occupations by AI–exposure score

#	**Occupation (O*NET title)**	**Exposure**
1	Sewers, Hand	0.00
2	Marketing Managers	0.00
3	Actuaries	0.00
4	Property, Real Estate, & Community Association Managers	0.01
5	Animal Trainers	0.01
6	Child, Family, and School Social Workers	0.01
7	Fashion Designers	0.02
8	Survey Researchers	0.02
9	Hearing Aid Specialists	0.02
10	Administrative Services Managers	0.02
11	Lodging Managers	0.02
12	Automotive and Watercraft Service Attendants	0.02
13	Postmasters and Mail Superintendents	0.03
14	Cleaning, Washing, and Metal Pickling Equipment Operators and Tenders	0.03
15	Biofuels/Biodiesel Technology and Product Development Managers	0.03
16	Treasurers and Controllers	0.03
17	Human Resources Assistants, Except Payroll and Timekeeping	0.03
18	Aircraft Cargo Handling Supervisors	0.03
19	Animal Breeders	0.03
20	Compliance Managers	0.04

ized adoption. Actual labor-market effects will depend on implementation costs, organizational adaptation, and regulatory constraints.

Figure 5.9 reports mean exposure scores across major occupational groups, revealing substantial variation. *Educational Instruction and Library* stands out with the highest average exposure, exceeding 0.40, consistent with the strong alignment of AI with language-intensive teaching and knowledge-delivery tasks. Above-average exposure also appears in *Arts, Design, Entertainment, Sports, and Media*, driven by generative content tools, and in *Protective Service* and *Computer and Mathematical* occupations, where advances in vision, language, and algorithmic capabilities are highly relevant. Skilled domains such as *Farming, Fishing, and Forestry* and *Architecture and Engineering* also rank above the median, reflecting the spread of AI into applied technical fields. At the lower end of the spectrum, *Building and Grounds Cleaning and Maintenance*, *Legal*, and *Community and Social Service* record the lowest mean exposure scores, all below 0.15. These groups involve embodied physical work or discretionary, judgment-heavy activities that remain hard to automate. Healthcare support, personal care, and management-related roles occupy the middle of the distribution, highlighting the continued importance of empathy, interpersonal skills, and organizational decision-making.

Taken together, the results underscore two dynamics. First, education emerges as an unexpectedly exposed sector, reflecting the overlap between LLM capabilities and instructional tasks. Second, exposure is not confined to white-collar cognitive work but extends into skilled trades and technical services, while domains reliant on social intelligence, physical dexterity in unstructured environments, or managerial discretion remain relatively insulated.

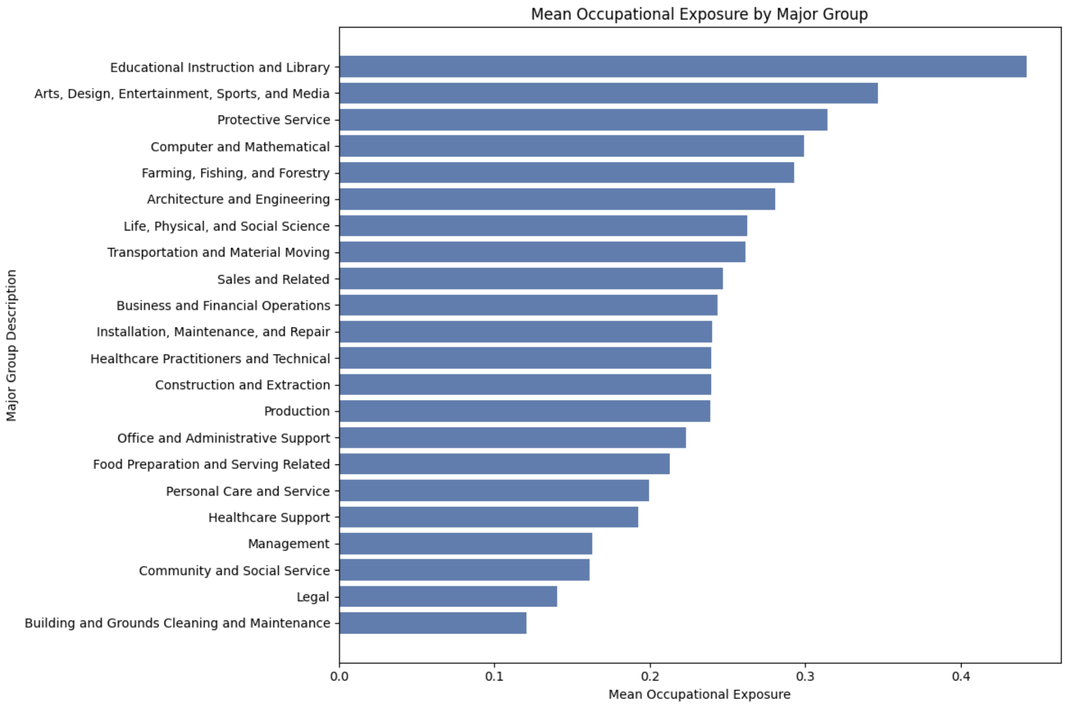


Figure 5.9: Occupational Exposure by Major Group(first 2 digits of SOC Code)

Validation To assess the reliability of our alignment framework, we conducted a series of validation exercises. The validation checks serve three purposes: first, to ensure that the human reference set itself is consistent; second, to benchmark the performance of GPT-4o against the reconciled gold standard; and third, to compare our occupation-level exposure scores with prior measures from the literature and to validate the framework through a focused experiment on education-related occupations. Collectively, these exercises provide evidence for the robustness of our methodology and the credibility of the resulting exposure scores.

Our first validation step examined the consistency of human judgments. Three expert annotators independently labeled a shared set of job task–AI task alignments, after which disagreements were reconciled into a gold standard. Inter-annotator agreement provides a measure of how reproducible the alignment process is across human experts.

Across the three annotator pairs, Cohen's κ values fall within a narrow band of [0.88, 0.90], and the multi-coder Krippendorff's α for nominal data is $\alpha = 0.86$ (Table 5.8). According to the Landis–Koch scale, these figures correspond to *"almost perfect"* reliability and are well above the commonly cited threshold of $\kappa/\alpha \geq 0.80$ for semantic labeling tasks. In practical terms, fewer than 12% of category assignments remained in disagreement after the two-round *Delphi* reconciliation, setting a strong empirical benchmark against which automated methods can be evaluated.

Next, we compared GPT-4o's annotations to this reconciled gold standard. This exercise evaluates whether the model can approximate expert-level agreement in assigning job tasks to AI capabilities. Without any task-specific fine-tuning, GPT-4o achieves an accuracy of 0.86 and a macro-F1 of 0.84 on the 300 held-out task pairs. Chance-corrected measures support these results: Cohen's $\kappa = 0.84$ and Krippendorff's $\alpha = 0.83$ place the model at the lower bound of the expert–expert reliability range. In effect, GPT-4o agrees with the reconciled reference labels nearly as closely as individual experts agree with each other. Most remaining disagreements occur in borderline PARTIAL ↔ NO-MATCH cases, typically driven by vague or multipurpose job-task statements. The small κ gap (approx. 0.06) between the model and the human upper bound suggests that costly manual annotation is only required to seed a demonstration set; once established, the LLM can propagate labels across large corpora with high fidelity. Most residual errors stem from underspecified *purpose* clauses in job-task statements, pointing to a clear improvement path: enrich purpose extraction or supplement prompts with short clarifying snippets. Given inherent task ambiguity, $\kappa \approx 0.90$ represents a practical ceiling, so gains beyond the current 0.84 would likely involve disproportionate effort.

We next examine how our occupation-level exposure score relates to existing indices in the literature. Specifically, we compute Spearman rank correlations with nine widely cited benchmarks: (1) Felten et al. (2019) AI-ability linkage, which maps technical progress to occupational abilities; (2) Webb (2019) text-based patent proximity scores, including overall AI, software, and robotics variants; (3) Brynjolfsson and Mitchell (2017) Suitability-for-ML rubric, which classifies tasks by their learnability; (4) Frey and Osborne (2013) automation-risk probabilities, based on expert task assessments; (5) Tolan et al. (2021) cognitive-benchmark based AI exposure and D. Xu et al. (2022, 2023) susceptibility and automatability indices

Panel A. Inter-annotator agreement (E1–E3)		
Statistic	**Value**	**N (pairs)**
Cohen's κ (E1 vs. E2)	0.88	300
Cohen's κ (E1 vs. E3)	0.90	300
Cohen's κ (E2 vs. E3)	0.88	300
Krippendorff's α (nominal)	0.86	300
Panel B. GPT-4o vs. Gold standard		
Metric	**Value**	**N (pairs)**
Accuracy	0.86	300
Macro-F1	0.84	300
Cohen's κ (GPT-4o vs. Gold)	0.84	300
Krippendorff's α (nominal)	0.83	300

Table 5.8: Validation results. Panel A reports inter-annotator agreement for the expert coders (E1–E3). Panel B reports GPT-4o performance against the reconciled gold standard.

that refine occupation-level risk measures. Rank-based measures are preferred over Pearson coefficients because several of the reference distributions are skewed and the underlying associations are unlikely to be linear. The results are summarized in Figure 5.10. For completeness, detailed scatter and kernel-density plots are provided in Appendix 5.11, where they illustrate the underlying distributional patterns.

The correlations are positive across all pairs but generally modest in magnitude. The strongest link is with Webb (2019)'s software score ($\rho = 0.25$), followed by the Suitability-for-ML index of Brynjolfsson and Mitchell (2017) ($\rho = 0.22$). In contrast, robotics-oriented or automation-risk measures (Webb, 2019; D. Xu et al., 2022) yield coefficients close to zero and fail to reach conventional significance thresholds. This pattern underscores that different indices capture distinct notions of technological impact: robotics scores emphasize physical automation, substitution, and exposure to routine tasks, while patent- or software-based measures highlight complementarity with digital technologies. Our pathway aligns more closely with the software/ML family than with robotics-risk metrics. Finally, the scatter plots and kernel densities in Appendix 5.11 confirm that all distributions are heavily skewed toward the lower end (most occupations cluster below 0.5), with diffuse point clouds in the off-diagonal panels. Taken together, the correlation analysis suggests that our index provides complementary information: it is broadly consistent with software/ML exposure measures, yet largely orthogonal to traditional automation-risk indices. This indicates that the proposed score captures a novel dimension of occupational exposure to AI.

Table 5.9 reports multivariate OLS estimates where the dependent variable is our AI

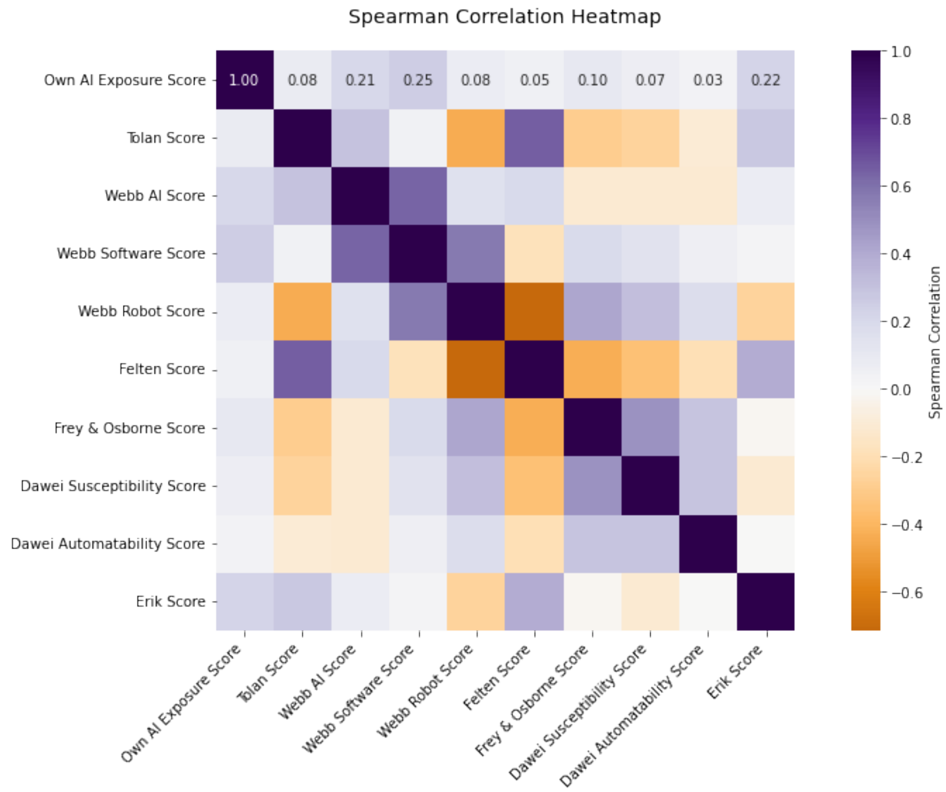


Figure 5.10: Correlation Heat maps (Spearman Ranks) against existing methods

exposure score. The explanatory block includes the nine external indices, augmented with two labour market controls. Robust HC3 standard errors are reported.

To construct the labour–market controls, we fitted for each SOC occupation the regression $\ln(tot\_emp)_{ot} = \beta_{0o} + \beta_{1o} t + \varepsilon_{ot}$ using BLS employment aggregates for $t = 2012, \ldots, 2022$. The estimated slope $\hat{\beta}_{1o}$ serves as the annualised log employment growth rate. In parallel, we obtained 2012 and 2022 real hourly means (after CPI deflation) and computed the compound annual growth rate $\left(\frac{w_{2022}}{w_{2012}}\right)^{1/10} - 1$, which is used as our wage control. These variables proxy for demand-side dynamics that could otherwise confound the association between technological exposure and occupational salience.

The regression results highlight that the strongest and most robust predictors of our index are those tied to software and machine-learning research. Webb (2019)'s software exposure and the Suitability-for-ML index of Brynjolfsson and Mitchell (2017) enter with large, highly significant coefficients ($\hat{\beta} = 0.284$ and 0.357, both $p < 0.01$), consistent with the bivariate correlations. By contrast, the robotics score is negative, economically negligible, and statistically insignificant, confirming that our measure is not simply a reformulation of classic robot exposure. Taken together, these variables contribute modest ex-

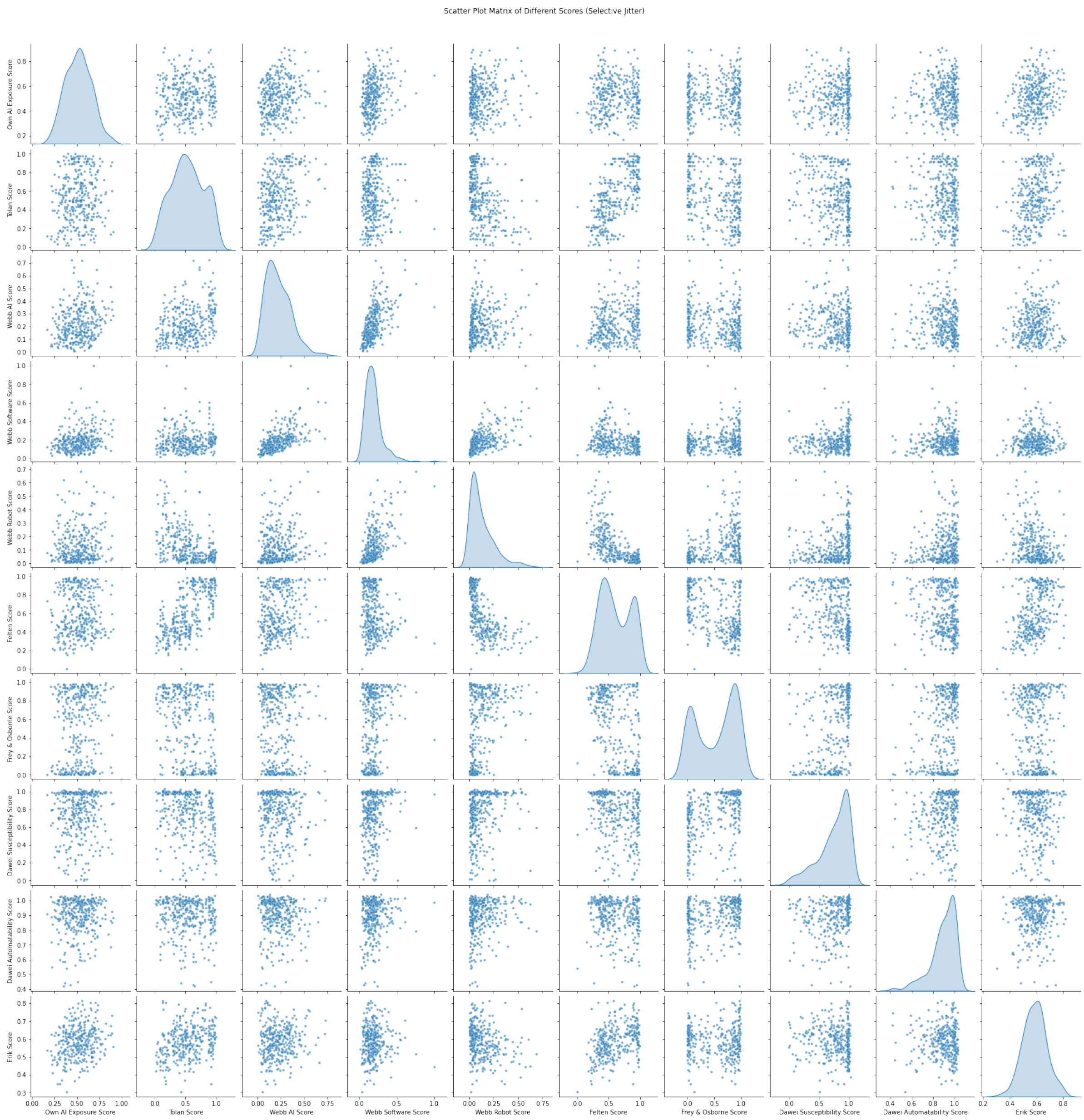


Figure 5.11: Correlation scatter and kernel-density maps against existing methods.

Variable	**Coef.**	**Std. Err.**
Cognitive-based Exposure (Tolan)	0.0049*	0.0035
AI Score (Webb)	0.0470**	0.0261
Software Score (Webb)	0.2839***	0.090
Robot Score (Webb)	−0.0225	0.105
AIOE (Felten)	0.0024*	0.0018
Automation Score (Frey & Osborne)	0.0244	0.028
Susceptibilities (Dawei)	0.0005	0.033
Automatabilities (Dawei)	0.0252*	0.0213
Suitability for ML (Erik)	0.3570***	0.088
Log Employment Growth Rate	−0.2227*	0.1898
Real Wage Growth Rate	−0.0852	0.125
Constant	0.2899**	0.155
Observations	540	
R^2	0.267	

Table 5.9: OLS regression results of various prior efforts and AI exposure measures. Robust (HC3) standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

planatory power—their joint effect passes an F-test only at the 10 percent level—indicating partial but limited overlap. Standard labour–market controls for employment and wage growth remain individually insignificant. The coefficient on log employment growth is negative (−0.22), implying that faster-growing occupations are weakly associated with lower exposure scores, although the effect is not statistically reliable. Including both controls marginally improves model fit (lower AIC) without altering the technology coefficients, suggesting that omitted demand-side dynamics are unlikely to bias the central estimates. Overall, the model explains roughly one quarter of the cross-occupational variance (R^2 = 0.27)—a reasonable fit given the conceptual heterogeneity of the regressors and the deliberately parsimonious specification. The results confirm that our AI–exposure score (1) is statistically coherent with leading measures of software and ML intensity, (2) complements rather than duplicates robot-centric automation-risk indices, and (3) remains robust to the inclusion of labour–market controls.

We observed that education-related occupations receive relatively highest exposure scores in our index. To check whether this is unique to our measure or shows up in other approaches, we compared how teachers rank across nine well-known exposure frameworks. We then explored the gap at the task level using regression and permutation tests.

Table 5.10 shows how education jobs are distributed across the exposure percentiles under different methods. Our index, along with Tolan et al. (2021), Webb (2019)'s AI score, and

Felten et al. (2019), places around two-thirds of education occupations in the top quartile (75th–100th percentile). In contrast, methods built around software engineering or robotics benchmarks (e.g., Webb Software, Webb Robot, Frey–Osborne, and Dawei Susceptibility) push teachers toward the middle of the distribution, with many not appearing until the upper quartile. The Suitability-for-ML rubric of Brynjolfsson and Mitchell (2017) sits in between, with teachers showing moderate exposure. This split reflects the nature of teaching work. Most core tasks are language- and cognition-heavy, which aligns well with large language model capabilities. Benchmarks that emphasize physical automation or programming tasks do not capture these activities as strongly.

Table 5.10: Share of education-related occupations appearing in each percentile bucket across alternative exposure measures

Method / Source	25th (%)	50th (%)	75th (%)	100th (%)
AI Exposure (Proposed)	62.50	73.21	89.29	100.00
Cogitive-based Exposure (Tolan)	64.62	96.92	96.92	100.00
AI Score (Webb)	66.10	84.75	96.61	100.00
AIOE (Felten)	63.33	96.67	100.00	100.00
Software Score (Webb)	0.00	5.08	30.51	100.00
Robot Score (Webb)	0.00	6.78	23.73	100.00
Automation Score (Frey & Osborne)	4.55	13.64	50.00	100.00
Susceptibilities (Dawei)	0.00	14.52	51.61	100.00
Automatability (Dawei)	4.00	20.00	30.00	100.00
Suitability for ML (Erik)	8.47	33.90	84.75	100.00

Appendix Figure 5.12 compares kernel-density violins for 17,238 non-teacher tasks and 412 teacher tasks. The teacher distribution is clearly shifted upward: the median task-level exposure score rises from 0.40 for other tasks to 0.63 for teacher tasks ($\Delta = 0.23$), with a large concentration of teacher tasks between 0.6 and 0.9. Formal statistical tests confirm this visual gap. A cluster-robust OLS regression at the task level shows that being a teacher task is associated with an increase of about 0.04 in the exposure score ($\beta = 0.040 \pm 0.012$, $p = 0.001$), corresponding to a medium effect size of Cohen's $d \approx 0.46$. The effect remains when tasks are weighted by their O*NET importance ratings ($\beta = 0.038$, $p = 0.003$) or when analysed using median regression. Randomisation checks reinforce these findings. A 5,000-draw permutation test that randomly re-labeled the teacher flag across tasks never produced a gap as large as the observed one (teachers rank about 190 positions higher among ~849 occupations), yielding $p < 0.0001$. The permutation-based p-value remains below 0.0002 even with 5,000 draws. These results confirm that the observed elevation of teach-

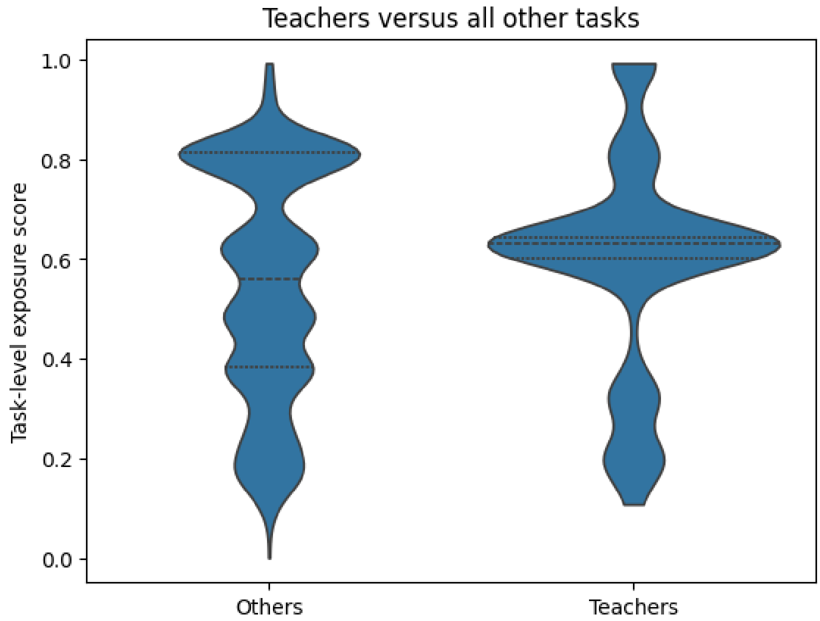


Figure 5.12: Distribution of exposure scores for teacher-related tasks (violin plot).

ing roles is highly unlikely to be due to chance. Results are summarised in Table 5.11.

Table 5.11: Teachers' tasks are systematically more exposed to LLM capabilities

	Coef.	S.E.	z	p-value	95% CI	
Intercept	0.5478	0.0090	60.41	<0.001	0.530	0.566
is_teacher	0.0404	0.0106	3.80	<0.001	0.020	0.061
Observations			5 314 tasks			
Clusters (occupations)			849			
Cohen's d (teacher vs others)			0.42			
Permutation test ($n = 5000$)			$\hat{p} < 0.0002$			

Notes. The dependent variable is the task-level exposure score (0−1), one observation per O*NET task statement. The key regressor *is_teacher* equals 1 for tasks belonging to Education–Training–Library occupations (SOC 25-xxxx).

Case Study In this section, we use case studies to demonstrate our framework's interpretability and simulation capacity. To demonstrate how our framework operates in practice, we conduct a focused case study on postsecondary teachers—a group of education-related occupations that consistently emerge as highly exposed in our analyses. This group includes 37 occupations out of the 849 in our database, spanning diverse fields such as biological sciences, psychology, engineering, and the arts. Across these 37 teaching occupations, we identify 236 unique job tasks, drawn from the 5,314 tasks in the full dataset.

These tasks are linked to 77 distinct AI tasks, covering nearly three-quarters of the AI task ontology (77 out of 106). The calculated exposure values for postsecondary teachers range from 0.10 to 0.75, with a mean of 0.54 and a median of 0.55, indicating consistently high exposure across the group. Table 5.12 lists the ten most frequently shared tasks and their associated AI capabilities. Core academic activities such as advising students, supervising research, grading, and writing grant proposals dominate the teacher task set. On the AI side, these tasks most often align with knowledge tracing, human skill assessment, dialog systems, and natural language generation. This mapping highlights the strong overlap between teaching responsibilities and recent advances in AI.

Table 5.12: Top 10 most frequently shared job tasks and AI tasks among postsecondary teachers.

Job Tasks (with counts)	**AI Tasks (with counts)**
Act as advisers to student organizations (34)	Knowledge tracing (43)
Compile bibliographies for outside reading etc. (34)	Human skill assessment (42)
Serve on academic or administrative committees etc. (34)	Dialog systems and chatbots (30)
Maintain office hours to advise and assist students etc. (34)	Workflow recognition (26)
Maintain student attendance records, grades, etc. (34)	Natural language generation (22)
Participate in student recruitment, registration, placement etc. (34)	Recommendation systems (21)
Collaborate with colleagues on teaching/research issues (33)	Action and activity recognition (16)
Compile, administer, and grade examinations etc. (33)	Document layout analysis (15)
Write grant proposals for external research funding (33)	Object detection and understanding (15)
Supervise teaching, internship, and research work etc.(33)	Visual reasoning (13)

Figure 5.13 visualizes these relationships by mapping overlapping job tasks to their corresponding AI tasks and exposure scores. For example, advising students and moderating curricula are closely linked to dialog systems (exposure = 0.32) and knowledge tracing (0.20), while research-related activities such as writing grant proposals are strongly associated with natural language generation (0.76). The figure also makes clear that exposure is not uniform: administrative and evaluative duties tend to show moderate alignment, while text-intensive scholarly tasks display markedly higher exposure.

Next, to illustrate the simulation capacity of our framework, we conduct a counterfactual simulation in which progress in a cluster of robotics-related AI tasks is assumed to accelerate sharply. Specifically, we impose a "robotics shock" by raising the GPS of key tasks: *robotic grasping* (0.38 → 0.80), *robotic skill* (0.08 → 0.50), *robot navigation* (0.62 → 0.90), and *robot task planning* (0.17 → 0.50)-based on the distribution of current AI Task GPS.

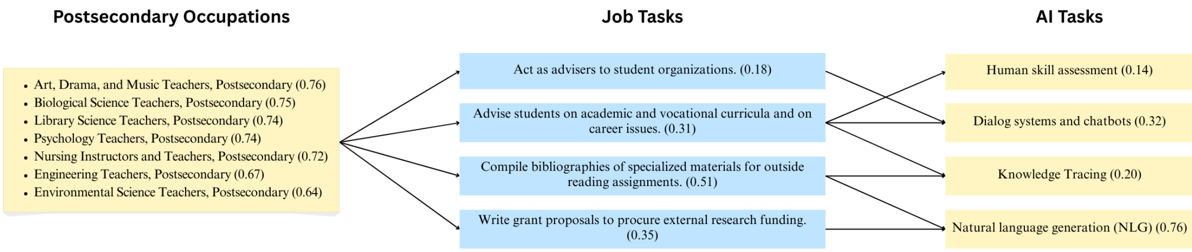


Figure 5.13: Case study mapping of postsecondary teacher occupations to overlapping job tasks and their corresponding AI tasks. The values in parentheses indicate the Exposure scores for job tasks and the GPS for AI tasks.

At the task level, the robotics shock leads to sharp increases in exposure for job tasks that involve physical handling, movement, and stocking (Appendix Table 5.13). The largest gains occur for tasks such as folding or removing products during processing (Δ = 0.46), loading and unloading trucks (Δ = 0.41), and maneuvering roofing units for installation (Δ = 0.38). These results highlight how robotics capabilities most strongly affect routine physical manipulation tasks, which are widespread across production, logistics, and construction settings. Aggregating to the occupation level, we observe that most occupations

Table 5.13: Sample job tasks with largest increases in exposure under the robotics shock scenario

Δ Exposure	**Job Tasks**
0.46	Fold products and product parts during processing; Remove products, machine attachments, or waste material from machines.
0.43	Obtain merchandise from bins or shelves; Pack bottles into cartons or crates, using machines.
0.41	Load and unload trucks, vans, or automobiles; Identify, pack, or transport hazardous or radioactive materials.
0.40	Move machinery and equipment, using hoists, dollies, rollers, and trucks; Perform general cleaning activities in kitchen and dining areas.
0.38	Maneuver completed roofing units into position for installation; Stock shelves with products.

remain close to their baseline exposure, but those with robotics-aligned tasks experience pronounced increases (Figure 5.14). Notable examples include *Automotive Body and Related Repairers*, *Drywall and Ceiling Tile Installers*, *Helpers–Production Workers*, and *Roofers*, all of which register substantial jumps in exposure. These results underscore that the effects of robotics are highly uneven, concentrated in occupations that involve physical construction, maintenance, and repair, while service-oriented and knowledge-intensive roles remain largely unaffected.

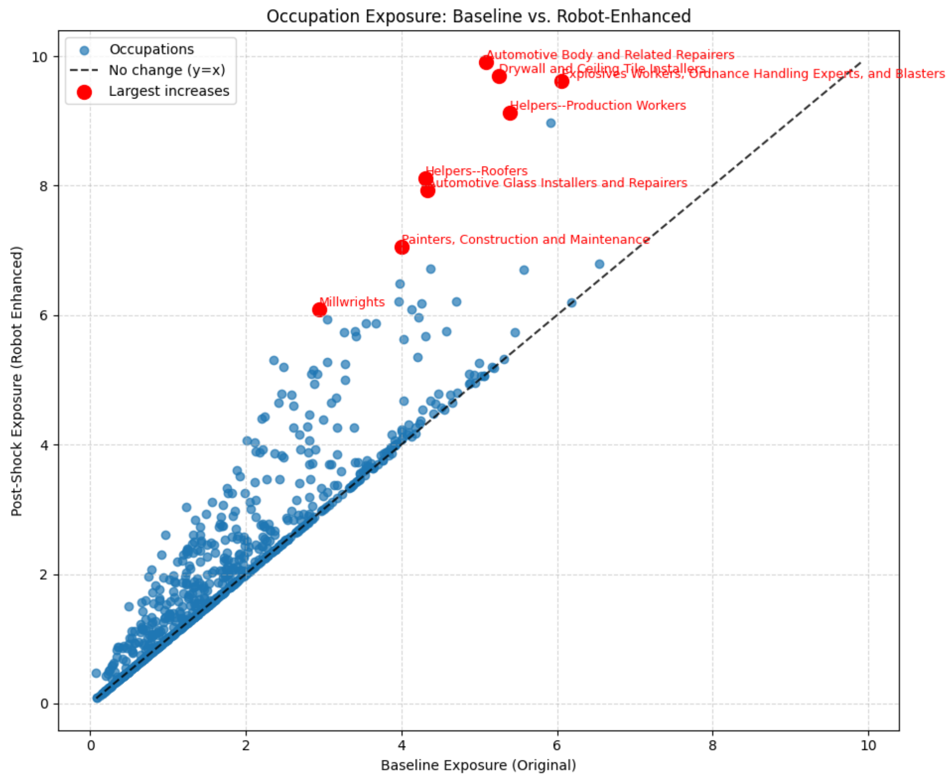


Figure 5.14: Occupation-level exposure before and after the robotics shock. Each point represents an occupation; the dashed line indicates no change. Highlighted in red are the occupations with the largest exposure increases.

Finally, we examine exposure aggregated by major occupational groups (Figure 5.15). The results reveal clear sectoral patterns. Groups such as *Production, Construction and Extraction*, and *Installation, Maintenance, and Repair* exhibit the largest increases, consistent with their heavy reliance on manual manipulation and equipment handling. *Transportation and Material Moving* and *Farming, Fishing, and Forestry* also show notable rises, reflecting robotics' potential to reshape logistics and agriculture. By contrast, groups such as *Legal*, *Management*, and *Community and Social Service* remain essentially unchanged. This divergence highlights how localized progress in robotics disproportionately amplifies exposure in manual domains, while leaving cognitive, legal, and social-interaction-intensive sectors largely unaffected.

Together, these two cases show how the framework can both interpret current exposure patterns and explore hypothetical trajectories of AI progress.

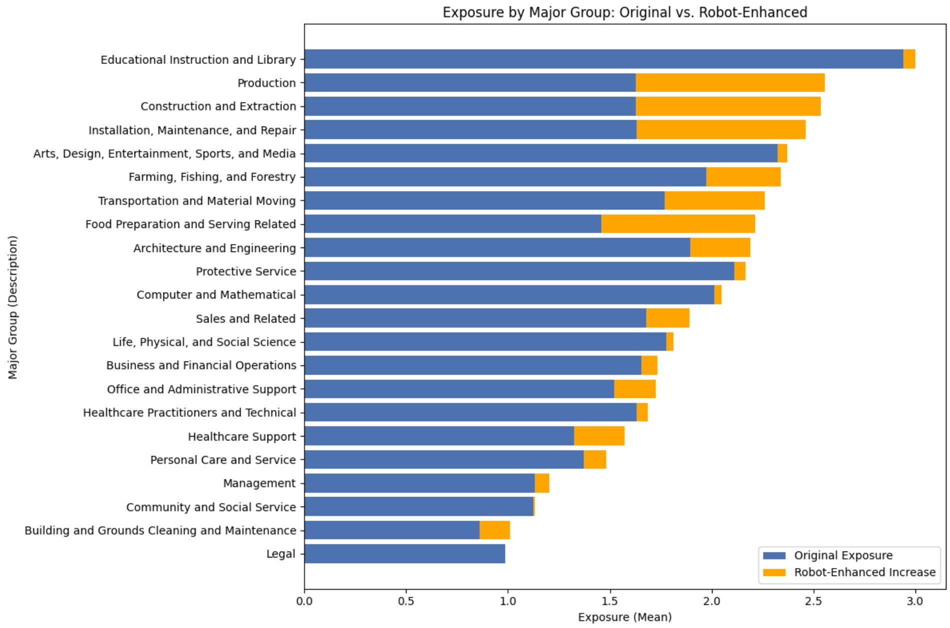


Figure 5.15: Exposure by major occupational group before and after the robotics shock. Bars show original exposure (blue) and the additional increase due to enhanced robotics capabilities (orange).

5.5 Summary

Our findings show that AI exposure is both broader and more nuanced than often assumed. Validation exercises confirm that the job task–AI alignment procedure is reliable: expert annotators achieved near-perfect agreement, and GPT-4o replicated these labels with accuracy close to human coders. At the occupational level, our index correlates moderately with prior software- and ML-based exposure measures, but shows little alignment with robotics- or automation-risk indices. Regression analysis further confirms that software and ML intensities are the dominant drivers of our exposure construct, while traditional robotics scores contribute little. Building on this validation, our results reveal systematic patterns that connect meaningfully to prior research. At the AI task level, we confirm earlier findings that progress is concentrated in language and vision domains (Felten et al., 2019; Tolan et al., 2021), with dialog systems, gesture recognition, and natural language generation emerging as especially high-impact. At the job-task level, we extend this literature by showing that exposure is uneven and multimodal: information-processing and text-based activities (e.g., translation, transcription, grading) are highly exposed, while geoscience fieldwork and subsurface assessment remain largely insulated. Importantly, exposure does not scale linearly with task importance or breadth, underscoring nuances that earlier occupation-level and ability-based studies could not capture. At the occupational level, we find moderate exposure on average, but with notable clusters—skilled trades and educational teachers—appearing near the top of the distribution, whereas jobs rooted in empathy, judgment, or dexterity (e.g., social workers, massage therapists) remain consistently low. Strikingly, our framework also places arts, design, entertainment, and media occupations among the most exposed (see Fig. 5.9), despite prior claims that such creative or socially intelligent work would remain largely unaffected by automation (Frey & Osborne, 2013). This suggests that the impact of AI is more widespread than previously thought: recent evidence shows that LLMs are demonstrating creative capabilities well beyond routine tasks, with outputs increasingly competitive in domains once assumed to require uniquely human creativity (L. Sun et al., 2025). Moreover, the results also confirm Eloundou et al. (2023)'s finding that LLMs can classify and evaluate tasks with human-comparable reliability, reinforcing the credibility of automated semantic matching.

Methodologically, we propose the first framework that directly aligns fine-grained job tasks with AI tasks, using an ontology-anchored semantic matching approach. By leveraging GPT-4o's reasoning to map task actions and purposes to a structured AI task ontology, our method provides interpretability and granularity that earlier indices lacked. Prior approaches relied on intermediate layers such as patent data, broad cognitive ability categories, or expert rubrics (Brynjolfsson & Mitchell, 2017; Tolan et al., 2021; Webb, 2019), whereas our pipeline establishes a direct and automated mapping. This offers a scalable way to capture nuanced correspondences that would be impractical to enumerate manually. The second contribution is the dynamic alignment with evolving AI capabilities. We weight job task–AI connections using technical progress and research popularity, which we term the Gain–Popularity Score (GPS), ensuring that the index reflects contemporary advances such as generative AI and computer vision models. Unlike earlier indices, which were static or backward-looking, our measure can adapt as benchmarks shift, allowing us to surface newly affected occupations. Finally, extensive validation demonstrates that the framework is both reliable and revealing. Expert annotators and GPT-4o reached near-human agreement, the index correlates most strongly with software/ML-based measures, and case study shows how the framework can explain current exposure patterns and simulate future AI progress to reveal shifts across tasks, occupations, and groups. In sum, our contribution lies in combining state-of-the-art NLP with a structured ontology and dynamic progress signals, yielding an interpretable, granular, and adaptive measure of AI's evolving impact on work.

Despite its contributions, our framework has limitations. First, it inherits biases from the data sources. We rely on O*NET task statements, which, while comprehensive, are generic and may miss within-occupation heterogeneity or new tasks that emerge in fast-changing jobs. Prior work shows that treating occupations as monolithic can give a misleading picture of automation risk (Arntz et al., 2016); our task-level focus improves but does not eliminate this issue. Similarly, the Intelligence Task Ontology and benchmark data may not fully capture emerging or hard-to-quantify AI capabilities, such as common-sense reasoning or tacit physical skills (Blagec et al., 2022). While the ITO provides one of the most comprehensive taxonomies of AI tasks, it is largely constructed from benchmarks hosted on Papers With Code. This reliance introduces gaps: some benchmarks are incomplete, sparsely populated, or lack consistent updates, which can create missing values in the ontology. New or fast-evolving AI capabilities may fall outside its current scope, and without regular updates, systematic blind spots could persist in our index.

In the future, we will refine and extend our framework in several ways. First, future research should draw on more comprehensive and dynamic datasets that record AI performance, popularity, and research investment. Such data would allow systematic gap analyses between where AI progress is concentrated and where human labor is most prevalent. Second, the simulation capacity of our framework can be deployed with real-world labor data to examine the effects of accelerated progress in specific AI subfields, or to evaluate how prioritizing research in areas would reshape exposure across occupations. Third, future indices should incorporate economic and institutional factors such as adoption costs, regulatory hurdles, and organizational readiness, as well as expert judgment on nuanced task requirements like ethics or aesthetics, to move from technical feasibility toward realistic impact. Taking these steps will ensure that AI exposure measures remain adaptive, realistic, and aligned with human-centric outcomes.

CHAPTER 6

AI EXPOSURE AND TALENT POTENTIAL: A LEVEL-AWARE GNN APPROACH

This chapter presents the fourth work, which focuses on modeling talent potential through career mobility prediction. We propose LADDER-GNN (Level-Aware Dynamic Development & Entity Representation Graph Neural Network), a framework that integrates hierarchical career features and dynamic trajectory modeling to predict three key aspects of workforce mobility: next job position, next employer, and employment duration. From these predictions, we derive a two-dimensional Career Projection Space that captures both the Ability to Thrive (upward mobility) and Willingness to Stay (organizational commitment), providing an interpretable lens for assessing talent potential. This work serves as a foundation for future research that explores AI's impact on workforce implications at the individual level—an underexplored yet crucial area. In particular, we aim to investigate how the AI Exposure Index interacts with measures of talent potential—for example, whether high-potential individuals are less exposed to automation risk, or whether they tend to cluster in highly exposed occupations where adaptability and continuous learning are essential for career advancement. By extending this line of inquiry, we move toward a unified perspective that links external technological pressures with internal workforce development, offering richer insights into the implications of AI for labor markets and talent management.

6.1 Overview

This work serves as a foundational extension of the thesis's overarching objective to understand AI's labour-market impacts at increasing levels of granularity. The preceding chapters focus on measuring AI and automation risk at the task, occupation, and job level, primarily addressing what types of work are exposed to technological change and through which mechanisms (e.g., substitution, complementarity, or capability alignment). This chapter shifts the analytical focus from jobs to individuals by modeling career mobility dynamics, capturing how workers move across positions, organizations, and hierarchical levels over time. Such a model is a necessary prerequisite for extending AI risk analysis to the individual level, as task- or occupation-level exposure alone cannot explain heterogeneous outcomes among workers with different career trajectories. By establishing a robust, interpretable framework for predicting individual career progression, this chapter provides the structural backbone for future work that integrates AI exposure indices with longitudinal career paths, enabling the assessment of AI-related risk, opportunity, and resilience at the individual level. In this sense, the chapter complements the earlier exposure-measurement studies by supplying the dynamic career model required to operationalize AI risk beyond static job characteristics.

Talent potential assessment has long been central to workforce planning and organizational success, as it shapes recruitment, development, and retention strategies. Traditional approaches—such as performance reviews, competency models, or assessment centers — provide useful insights but remain inherently static, offering only point-in-time snapshots of an individual's abilities. These methods fail to capture the dynamic, longitudinal nature of professional growth, which unfolds across different roles, organizations, and sectors. In today's increasingly fluid labor market, where careers span multiple institutions and involve complex trajectories, such static evaluations are insufficient. This limitation highlights the need for trajectory-based approaches that can reflect both career progression and adaptability, providing a more comprehensive measure of talent potential.

To address this challenge, we propose LADDER-GNN (Level-Aware Dynamic Development & Entity Representation GNN), a novel framework that explicitly incorporates hierarchical features across entities such as job positions, companies, universities, and academic majors. By modeling career trajectories as both sequences and graphs, LADDER-GNN distinguishes between lateral moves and upward mobility, integrates ranking signals (e.g., prestige or seniority), and accounts for real-world complexities including parallel experiences and career gaps. This design enables a more nuanced understanding of career development patterns and long-term talent dynamics. The framework involves a two-stage learning process: first, contrastive pretraining to derive unified career representations, and second, joint fine-tuning on three downstream tasks—*next job position*, *next employer*, and *employment duration*. These tasks serve as complementary proxies for talent potential, capturing different dimensions of mobility and stability. Experimental results show that LADDER-GNN consistently outperforms existing baselines, particularly in scenarios involving significant hierarchical transitions, which are critical signals of professional growth.

Beyond predictive accuracy, we introduce the concept of a Career Projection Space, a two-dimensional representation of talent potential along the axes of *Ability to Thrive* (upward mobility) and *Willingness to Stay* (organizational commitment). This interpretable framework allows organizations to segment their workforce into actionable categories—such as high-potential leaders, loyal but underutilized employees, or turnover-prone talent—and to design targeted interventions for planning, retention, and career development.

Finally, this work also lays the foundation for future research connecting talent potential with AI exposure. By combining the AI Exposure Index developed in earlier chapters with individual-level measures of potential, we can explore whether high-potential individuals are systematically less exposed to automation risk, or whether adaptability becomes a key differentiator in highly exposed occupations. In this way, LADDER-GNN not only advances career modeling but also provides a bridge toward understanding how technological change interacts with individual trajectories in the workforce.

6.2 Preliminary for LADDER-GNN

Career mobility prediction can be formulated as the problem of modeling and forecasting an individual's future career trajectory based on their historical employment and education records, including the possibility of parallel experiences such as working while studying or holding multiple jobs simultaneously. Let R_i denote an individual, whose observed career path is a temporal sequence of (possibly overlapping) sets of career events: $\mathcal{H}_i = \{\mathcal{E}_i^1, \mathcal{E}_i^2, \ldots, \mathcal{E}_i^T\}$, where $\mathcal{E}_i^t = \left\{\left(c_{i,j}^t, p_{i,j}^t, d_{i,j}^t, \ell_{c,j}^t, \ell_{p,j}^t\right)\right\}_{j=1}^{n_t}$. Here, $\mathcal{E}_i^t$ represents the set of parallel experiences active during time step t, where n_t is the number of simultaneous roles at that step; $c_{i,j}^t$ and $p_{i,j}^t$ denote the company/school and position/major for the j-th experience; $d_{i,j}^t$ is the duration of that experience; and $\ell_{c,j}^t$ and $\ell_{p,j}^t$ are the corresponding hierarchical level indicators. In addition to the sequence $\mathcal{H}_i$, we optionally include individual profile features $\mathcal{I}_i$ such as demographics, skills, and other relevant attributes. The goal is to learn a predictive function:

$$M : (\mathcal{H}_i, \mathcal{I}_i) \to (c_i^{T+1}, p_i^{T+1}, d_i^{T+1}, \ell_c^{T+1}, \ell_p^{T+1})$$

which forecasts the individual's next employer c_i^{T+1}, position p_i^{T+1}, expected duration d_i^{T+1}, and the associated hierarchical levels $\ell_c^{T+1}, \ell_p^{T+1}$, accounting for overlapping career experiences.

6.3 Methodology

To capture career mobility through a realistic, hierarchy-sensitive lens, we present LADDER-GNN, a level-ware model that learns rich representations of career trajectories for prediction in dynamic labor markets. In LADDER-GNN, we selectively adopt the sequence and graph modules from UniTRep (Zha, Sun, et al., 2024) as core representation learning components, then we integrate hierarchical level features into both graph and sequence modules and introduce mechanisms to model parallel experiences and career gaps. The learned representations are fine-tuned on three tasks—predicting the next job position, predicting the next employer, and estimating job duration—jointly optimized in a *multi-task learning* framework to enhance generalization. Through these extensions, LADDER-GNN offers a more expressive and comprehensive approach to modeling and predicting career mobility in labor markets (See Figure 6.1 for LADDER-GNN architecture).

6.3.1 Input Module

The Input Module serves as the foundation of the LADDER-GNN framework by jointly encoding heterogeneous career information into a unified feature space. As illustrated on the left side of Figure 6.1, individual career histories are first represented as hierarchical graphs, where nodes correspond to fine-grained positions and parent-level entities, and edges capture structural relations and transitions. Each node is associated with multiple attributes, including temporal information (start date, duration), hierarchical indicators (child- and parent-level features), and parallel contextual features. These raw attributes are transformed through feature embedding layers to obtain dense representations that preserve both semantic and temporal characteristics. In parallel, individual career trajectories are modeled as ordered sequences, enabling the framework to capture longitudinal dependencies through convolutional and recurrent layers. This dual input design—graph-structured representations for structural context and sequence-based representations for

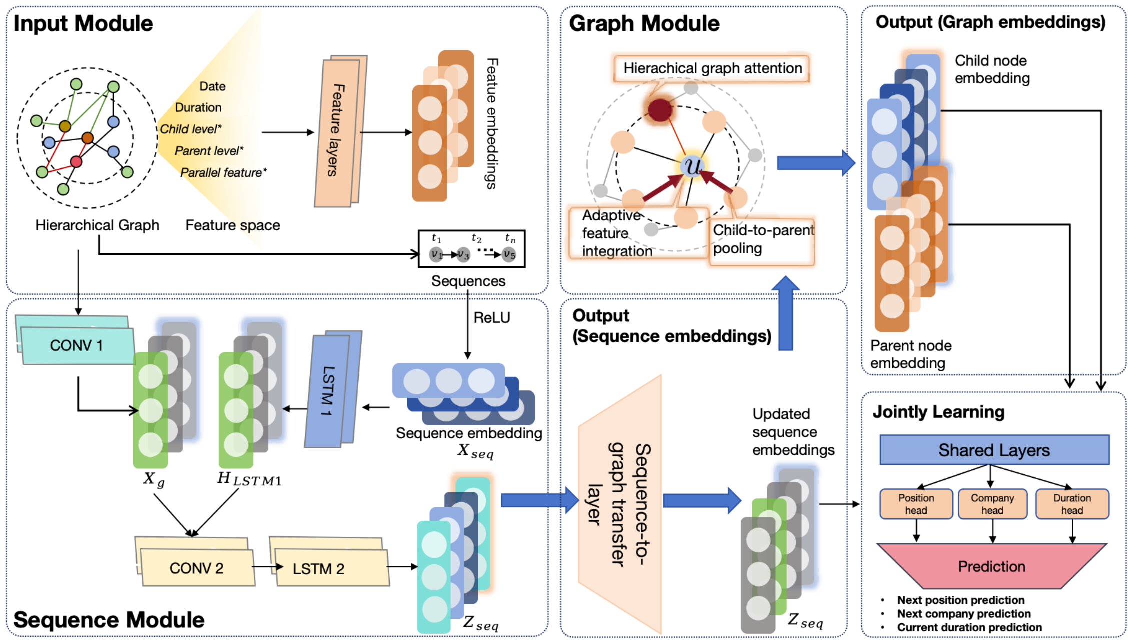


Figure 6.1: Overview of the *LADDER-GNN* framework.

temporal dynamics—ensures that both relational and sequential signals are explicitly retained before higher-level interaction and fusion, consistent with prior unified career representation learning frameworks

6.3.2 Sequence Module

The Sequence Module models temporal career trajectories by capturing entity evolution and contextual interactions. Given a set of individual trajectories $\mathscr{S} = s_1, s_2, \ldots, s_m$, where each s_i comprises a sequence of tuples in the form (v_j, t_j^s, t_j^e) representing an entity and its associated time span, the module generates temporal representations through a multi-stage process.

First, in the *historical information encoding* stage, raw sequences are transformed into enriched embeddings by incorporating three critical dimensions: (1) temporal dynamics are modeled via diachronic embeddings, where base entity embeddings evolve over time using a parametric time-aware function; (2) hierarchical features are integrated to reflect an entity's level—child-level features encode the hierarchy of job positions or majors, while parent-level features denote companies' or schools' ranking ; and (3) parallel experience features are introduced to represent overlapping job positions or concurrent study-work patterns, capturing both temporal overlap and intensity. These components are concatenated and projected into the initial sequence representation $\mathbf{X}_{\text{seq}}$ through a non-linear transformation. Next, to capture temporal dependencies and progression patterns, a recurrent neural network is applied to the sequence, producing intermediate hidden states denoted by:

$$\mathbf{H}_{\text{lstm}^1} = \text{LSTM}_1(\mathbf{X}_{\text{seq}}) \tag{6.1}$$

To enrich these representations with external labor market dynamics, a graph-to-sequence attention mechanism—denoted as HTGT_{g2s}—is used to inject global structural information from the market graph G into the sequence. This mechanism employs multi-head attention, where sequence entities (as queries) attend to relevant graph nodes (as keys and values):

$$\alpha_{ij} = \text{softmax}\left(\frac{(\mathbf{W}^Q \mathbf{h}_i^{seq})(\mathbf{W}^K \mathbf{h}_j^{graph})^T}{\sqrt{d_k}}\right)\mathbf{W}^V \mathbf{h}_j^{graph} \tag{6.2}$$

The attention weights α_{ij} determine how much information from each graph node j influences sequence entity i, allowing sequence representations to incorporate market-wide patterns. This results in context-aware embeddings $\mathbf{X}_{\text{g2s}}$. Finally, a second LSTM layer refines these enriched embeddings to produce the final trajectory-level representations $\mathbf{Z}_{\text{seq}} = \text{LSTM}_2(\mathbf{X}_{\text{g2s}})$. These output vectors encode both the individual's historical progression and the influence of broader market context, and are subsequently passed to the graph module through a historical embedding transfer mechanism.

After obtaining the sequence embeddings $\mathbf{Z}_{\text{seq}}$, the Sequence-to-Graph Transfer Layer aggregates entity representations across all their occurrences in different trajectories through an attention-based pooling operation, enabling information flow from individual trajectories to the structural graph representation:

$$\mathbf{h}_v = \sum_{i \in \mathcal{O}(v)} \alpha_{iv} \mathbf{h}_i^{seq} \tag{6.3}$$

where $\mathcal{O}(v)$ represents all occurrences of entity v across sequences, and α_{iv} are attention weights computed based on node types. The pooled embeddings update a global history tensor: $\mathbf{H}_{\text{history}}[v] = \mathbf{h}_v$, which maintains a persistent memory of temporally-informed entity representations. Later, when the graph module initializes node features, it draws upon these historical embeddings through an adaptive integration mechanism:

$$\mathbf{x}_v = \sigma(\beta_v) \cdot \mathbf{e}_v^{\text{base}} + (1 - \sigma(\beta_v)) \cdot \mathbf{H}_{\text{history}}[v] \tag{6.4}$$

where β_v is a learnable parameter that controls the mixing ratio between static embeddings and sequence-derived historical information. This mechanism ensures that structural patterns learned in the graph module are enriched with temporal dynamics captured in individual trajectories.

6.3.3 Graph Module

The Graph Module focuses on the hierarchical structure of the labor market by capturing relationships among entities across different levels. It operates on a hierarchical hypergraph $G = (V_c \cup V_p, E)$, where V_c and V_p represent child nodes (e.g., majors/jobs) and parent nodes (e.g., schools/companies), respectively. The module begins with an adaptive feature integration stage where node features blend static embeddings with sequence-derived historical information:

$$\mathbf{x}_v = \sigma(\beta_v) \cdot \mathbf{e}_v^{\text{base}} + (1 - \sigma(\beta_v)) \cdot \mathbf{H}_{\text{history}}[v] \tag{6.5}$$

This integration allows the model to control the influence of temporal dynamics on structural patterns using learnable parameters β_v that can vary by node type.

Central to the graph module is its hierarchical level-aware message passing mechanism implemented through Hierarchical Temporal Graph Transformer (HTGT) layers. During message passing, both child-level and parent-level features are incorporated into node representations. Child-level features are retrieved from pre-computed mappings, while parent-level features are dynamically calculated on-the-fly during message propagation using timestamp dependent lookups: $\mathbf{l}_{\text{parent}} = \text{ParentLevelEmb}(\text{get\_parent\_level}(v, t))$. This timestamp-specific approach captures the evolution of organizational prestige and influence over time. The HTGT layers employ multi-head attention to selectively aggregate information from neighboring nodes: $\mathbf{h}_i^{l+1} = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}^V [\mathbf{h}_j^l, \mathbf{e}_{ij}, \mathbf{l}_j^{\text{child}}, \mathbf{l}_j^{\text{parent}}]$. The attention coefficients α_{ij} are computed as:

$$\alpha_{ij} = \frac{\exp(e_{ij}/\sqrt{d})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik}/\sqrt{d})}, \quad \text{where } e_{ij} = (\mathbf{W}^Q \mathbf{h}_i)^\top (\mathbf{W}^K [\mathbf{h}_j, \mathbf{e}_{ij}, \mathbf{l}_j^{\text{child}}, \mathbf{l}_j^{\text{parent}}]) \tag{6.6}$$

This attention mechanism enables entities to focus on the most relevant connections in their neighborhood, capturing nuanced relationships between different types of nodes. After processing through multiple HTGT layers, the module employs a hierarchical pooling mechanism to derive parent node representations from their associated child nodes. This can use either attention-based pooling: $\mathbf{h}_p = \sum_{c \in \mathscr{C}(p)} \gamma_{cp} \mathbf{W}_{type(c)} \mathbf{h}_c$, where γ_{cp} are attention weights computed with type-specific projections, or average pooling that normalizes the contributions from all child nodes. The attention-based approach allows more influential child entities (e.g., prestigious majors or senior positions) to contribute more significantly to their parent representations.

The graph module ultimately produces two sets of embeddings: $\mathbf{Z}_{\text{child}}$ for majors and jobs, and $\mathbf{Z}_{\text{parent}}$ for schools and companies. These embeddings capture both the structural position of entities in the hierarchical graph and their temporal evolution informed by sequence dynamics.

6.3.4 Training Strategy and Jointly Learning

Our LADDER-GNN model employs a two-phase learning approach beginning with unsupervised pretraining. We use multi-objective contrastive learning that combines three complementary objectives: (1) Sequence-Node alignment that maximizes similarity between sequence embeddings and their target node representations, (2) Child Graph Link Prediction that preserves structural relationships between major/job entities, and (3) Parent Graph Link Prediction that captures connections between school/company entities. These objectives are combined into a weighted loss function $\mathscr{L} = \lambda_{\text{seq}}\mathscr{L}_{\text{seq}} + \lambda_{\text{child}}\mathscr{L}_{\text{child}} + \lambda_{\text{parent}}\mathscr{L}_{\text{parent}}$.

After pretraining, we employ a jointly learning strategy for fine-tuning across all downstream tasks simultaneously. We attach task-specific heads to the sequence embeddings:

$$P(v_{n+1}|s) = \text{softmax}(W_{\text{pos}}\mathbf{Z}_{\text{seq}}(s) + b_{\text{pos}}) \tag{6.7}$$

for next position prediction,

$$P(p_{n+1}|s) = \text{softmax}(W_{\text{comp}}\mathbf{Z}_{\text{seq}}(s) + b_{\text{comp}}) \tag{6.8}$$

for next company prediction, and

$$d_{n+1} = W_{\text{dur}}\mathbf{Z}_{\text{seq}}(s) + b_{\text{dur}} \tag{6.9}$$

for duration prediction. We optimize all tasks jointly using a weighted multi-task objective:

$$\mathscr{L}_{\text{finetune}} = \alpha_1\mathscr{L}_{\text{pos}} + \alpha_2\mathscr{L}_{\text{comp}} + \alpha_3\mathscr{L}_{\text{dur}} \tag{6.10}$$

where the weights are tuned to balance task contributions. This approach enables knowledge sharing across tasks while freezing the graph module to preserve structural information, leading to more robust representations and improved performance compared to task-specific fine-tuning.

6.3.5 Output Module

The Output Module produces two complementary embeddings that jointly characterize individual career trajectories and their structural context. As shown on the right side of Figure 6.1, the graph module employs hierarchical graph attention and child-to-parent pooling to generate graph embeddings for both fine-grained (child) and coarse-grained (parent) nodes, capturing cross-level structural dependencies within the career graph. In parallel, the sequence module outputs updated sequence embeddings that encode temporal career progression after interaction with graph-level representations through the sequence-to-graph transfer layer. These two embeddings—sequence embeddings reflecting individual temporal evolution and graph embeddings reflecting relational and hierarchical structure—are jointly optimized via shared layers and multi-head prediction tasks. This design allows the model to maintain a clear separation between temporal and structural signals while enabling controlled information exchange, supporting robust downstream predictions such as next position, next employer, and tenure duration. By jointly learning sequence and graph embeddings, the framework achieves a coherent representation of career mobility that integrates individual trajectories with broader labor-market structure, aligning with established cross-scale representation learning paradigms.

6.4 Experiments

In this study, we leverage a large-scale open professional network (OPN) dataset containing detailed educational and work histories of over 259,000 individuals. The dataset spans from 1960 to 2022 and captures dynamic career trajectories enriched with fine-grained entity-level information. In total, it comprises approximately 1.36 million work experience records across 9,582 unique companies and 244,441 unique job positions. Additionally, the dataset includes 379,332 education experience entries linked to 47,026 distinct schools and 131,180 unique academic subjects. This rich structural diversity enables comprehensive modeling of career paths across both professional and educational domains.

Each entity in the dataset is annotated with hierarchical level features to enable structured modeling of career progression across jobs, companies, academic majors, and educational institutions. For the job position hierarchy, we adopted the IPOD classification system (J. Liu et al., 2020), organizing titles into five levels based on title semantics: *Level-1 Entry* (e.g., intern, junior roles), *Level-2 Middle* (e.g., “Mid-” roles), *Level-3 Senior* (e.g., “Senior” or “Sr.” roles), *Level-4 Management* (e.g., directors, supervisors), and *Level-5 Executive* (e.g., CEO, CFO). The company hierarchy is derived from external engagement scores based on Google Trends, under the assumption that companies with higher public visibility possess greater influence (Hao & Li, 2021; Mariconda & Lurati, 2014; A. Yang & Kent, 2014). Companies are stratified into six levels: *Level-0 Non-company* (e.g., army, government), *Level-1 No score*, *Level-2 Bottom 25%*, *Level-3 25th–50th percentile*, *Level-4 50th–75th percentile*, and *Level-5 Top 25%*. To assess the validity of this visibility-based hierarchy, we further verify its correlation with company size metrics (e.g., number of employees), which serve as an independent proxy for organizational scale and influence.

For the school hierarchy follows the QS World University Rankings (Symonds, 2024), classifying institutions as *Level-0 Primary/Secondary*, *Level-1 Non-ranked* (e.g., community colleges), *Level-2 Beyond 500*, *Level-3 150–500*, *Level-4 50–150*, and *Level-5 Top 50*. Lastly, for the major hierarchy, we employed the ISCED 1997 framework (Hoffmeyer-Zlotnik & Wolf, 2003) to group academic programs into five tiers: *Level-1 Primary/Secondary/Certificates*, *Level-2 Associate/Diploma*, *Level-3 Bachelor/Post-Bachelor*, *Level-4 Postgraduate* (Master's, MPhil), and *Level-5 Doctorate* (PhD, Medical PhD).

The evaluation results presented in Table 6.1 compare the performance of several baseline models—including classical machine learning methods (Random Forest and Logistic Regression) (Breiman, 2001; Hosmer et al., 2013), sequential neural networks (LSTM, Transformer, BERT, and NEMO) (**transformer**; **bert**; Hochreiter & Schmidhuber, 1997; L. Li et al., 2017), and GNN-based models (HCPNN, UnGAE, AHEAD, UniTRep) (Meng et al., 2019; Zha, Qin, et al., 2024; Zha, Sun, et al., 2024; L. Zhang et al., 2021)—against our proposed LADDER-GNN model. The models were assessed on three critical career mobility prediction tasks: predicting the next job position, identifying the next employer, and forecasting employment duration. Metrics used for evaluation include accuracy at various ranks (ACC@1, ACC@3, ACC@5, ACC@10), mean reciprocal rank (MRR) for classification tasks (position and employer predictions), and mean absolute error (MAE) and root mean square error (RMSE) for duration prediction (L. Li et al., 2017; Meng et al., 2019; L. Zhang et al., 2021). LADDER-GNN consistently outperformed baseline models across all evaluated tasks, demonstrating significant improvements in predictive accuracy and robustness. Its superior performance is mainly due to its integration of macro-level market dynamics alongside micro-level individual career features, hierarchical job/major and company/school attributes, and parallel career experiences. The pretraining and fine-tuning approach further enhances predictive accuracy by initially learning generalizable patterns from extensive career data, which are then adapted for specific mobility tasks. We also demonstrated the effects of different training strategies and showed the results in the Appendix.

Table 6.1: Evaluation Results of Career Mobility Prediction with Average Improvements

Method	ACC@1	ACC@3	ACC@5	ACC@10	MRR	**Avg. Improv.**	ACC@1	ACC@3	ACC@5	ACC@10	MRR	**Avg. Improv.**	MAE	RMSE	**Avg. Improv.**
	Next Job Position						**Next Employer**						**Duration**		
RF	0.0718	0.1555	0.2104	0.3248	0.1497	142.85%	0.1080	0.1658	0.2543	0.3128	0.1500	167.50%	3.9272	4.9868	54.38%
LR	0.0980	0.1875	0.2555	0.4012	0.1998	90.70%	0.1225	0.2412	0.3001	0.3842	0.2002	111.57%	3.2957	4.1771	45.58%
LSTM	0.1385	0.2502	0.3314	0.5212	0.2301	46.65%	0.1771	0.2887	0.3989	0.4621	0.2701	61.82%	2.5278	3.2932	30.09%
BERT	0.1245	0.2552	0.3204	0.5115	0.2244	51.85%	0.1678	0.3192	0.4055	0.4902	0.2993	54.99%	2.4114	3.2794	28.36%
Transformer	0.1419	0.2712	0.3435	0.5103	0.2500	40.53%	0.1999	0.3201	0.4101	0.5004	0.2912	48.93%	2.4036	3.3684	29.23%
NEMO	0.1568	0.2866	0.3723	0.5100	0.2683	31.96%	0.1978	0.3002	0.4132	0.4998	0.3259	47.78%	2.3967	3.2765	28.11%
HCPNN	0.1643	0.2951	0.3689	0.5008	0.2745	29.99%	0.2102	0.3545	0.4893	0.5687	0.3877	29.03%	2.4087	3.2832	28.36%
UnGAE	0.1730	0.2984	0.3725	0.5064	0.2821	27.14%	0.3008	0.4391	0.5303	0.6301	0.4001	9.54%	2.0261	3.2520	21.53%
AHEAD	0.2040	0.3790	0.4705	0.5899	0.3241	5.57%	0.3084	0.4790	0.5593	0.6467	0.4223	4.26%	1.9874	2.9001	16.16%
UniTRep	0.2109	0.3859	**0.4894**	0.6216	0.3450	1.32%	0.2996	0.4694	0.5419	0.6299	0.4152	6.87%	1.8221	2.8999	12.39%
LADDER-GNN	**0.2201**	**0.3904**	0.4847	**0.6249**	**0.3501**	-	**0.3104**	**0.4984**	**0.5735**	**0.6810**	**0.4599**	-	**1.6504**	**2.4545**	-

Overall, LADDER-GNN consistently achieves the best or near-best performance across all evaluation metrics, indicating that incorporating hierarchical information and joint graph–sequence learning yields systematic improvements. For next job position prediction, LADDER-GNN attains the highest ACC@1 (0.2201) and MRR (0.3501), outperforming strong baselines such as UniTRep and AHEAD, particularly on top-rank accuracy, which is most relevant for practical recommendation and decision-making scenarios. Similar gains are observed for next employer prediction, where LADDER-GNN improves ACC@1 to 0.3104 and MRR to 0.4599, reflecting enhanced capability in capturing market-level transition patterns. For duration prediction, LADDER-GNN achieves the lowest MAE (1.6504) and RMSE (2.4545), indicating more accurate estimation of job tenure compared to both sequential and graph-based baselines. Notably, the performance gains over recent deep learning models are incremental rather than dramatic, suggesting that the improvements stem from more informative representations rather than increased model complexity. Taken together, these results demonstrate that explicitly modeling hierarchical structure, temporal dynamics, and their interactions leads to more robust and generalizable career mobility predictions across multiple outcome dimensions.

To complement predictive modeling, we propose a Career Projection Space that positions individuals within a two-dimensional latent space, termed the *Career Projection Space*, facilitating interpretable analyses of career dynamics and talent segmentation. This space consists of two meaningful axes: *Ability to Thrive* and *Willingness to Stay*. The *Ability to Thrive* axis A_i captures an individual's upward mobility potential, integrating the hierarchical levels of predicted next job and employer. Formally, let $\hat{l}_i^{job}$ and $\hat{l}_i^{comp}$ represent predicted level scores for the next job and company, respectively, with $\lambda \in [0,1]$ weighting job versus company level contributions. The composite thrive score is defined as $A_i = \lambda \cdot \hat{l}_i^{job} + (1-\lambda) \cdot \hat{l}_i^{comp}$. The *Willingness to Stay* axis S_i assesses the predicted duration an individual will remain in their current position, quantified simply as $S_i = \hat{d}_i$. Each individual i is projected onto this two-dimensional plane at the point (A_i, S_i), facilitating categorization into four strategic quadrants: High Thrive, High Stay (high-performing, loyal individuals suited for leadership and strategic retention), Low Thrive, High Stay (stable yet underutilized employees suitable for upskilling or lateral mobility), High Thrive, Low Stay (ambitious talents requiring engagement through promotions or challenges to prevent turnover), and Low Thrive, Low Stay (at-risk individuals potentially experiencing role misfit or disengagement).

As shown in Figure 6.2, the highest concentration of employees appears in Q4 (Low Thrive, Low Stay), indicating many valuable employees may be at risk of leaving despite having good position and compensation levels. A notable group in Q2 (Low Thrive, High Stay), representing a loyal but potentially under-leveraged workforce segment. Relatively few employees fall into Q1 (High Thrive, High Stay), which represents the ideal combination of high performance and retention. This structured representation empowers granular individual analyses alongside macro-level workforce insights, enabling HR practitioners to transform predictive data into strategic decisions for talent management and retention.

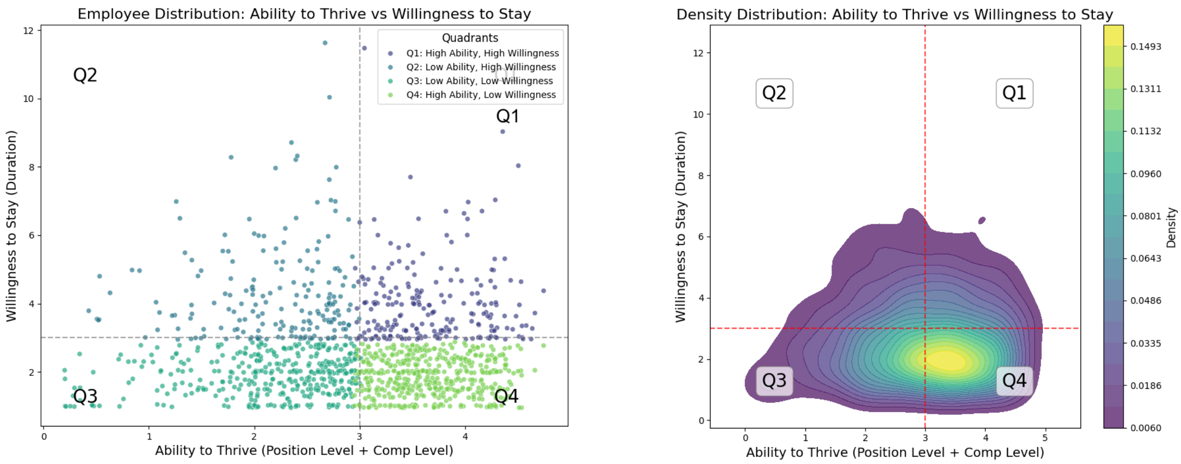


Figure 6.2: The visualization of the 2D Career Projection Space. We set Ability to Thrive $(0.25 \times \text{position\_level} + 0.75 \times \text{comp\_level})$ and Willingness to Stay (duration) axes using custom thresholds (3.0), displaying 0.5% of data points with minimal jitter for clarity.

6.5 Summary

In this chapter, we have presented LADDER-GNN, a novel level-aware graph–sequence framework that integrates hierarchical information across companies, positions, universities, and majors into career mobility modeling. By combining contrastive pretraining on large-scale trajectory data with joint fine-tuning on three downstream tasks—*next job position*, *next employer*, and *employment duration*—our model consistently outperforms both traditional classifiers and state-of-the-art sequential and graph-based baselines. Beyond predictive accuracy, the introduction of the two-dimensional Career Projection Space, defined by *Ability to Thrive* and *Willingness to Stay*, transforms opaque predictions into interpretable, actionable insights for talent management, enabling workforce segmentation that supports targeted retention and development strategies.

From a methodological perspective, LADDER-GNN contributes to the technical frontier of trajectory representation learning by incorporating hierarchical signals into both graph and sequence modules. Looking ahead, we plan to enhance interpretability by employing post-hoc attribution techniques. Specifically, SHAP will be applied to the sequence module to quantify how individual experiences (e.g., jobs, employers, or degrees) shape predictions, while GNN-SHAP will be used on the graph module to attribute embeddings and message-passing pathways to influential nodes, edges, or substructures. This line of work moves us toward a fully explainable system for fair, transparent, and accountable career decision-making.

More importantly, this chapter also lays the groundwork for integrating talent potential modeling with the AI Exposure Index developed in earlier chapters. A key future direction is to examine how external technological pressures (AI exposure) interact with internal workforce dynamics (talent potential). Several research questions arise:

- Are high-potential individuals systematically less exposed to automation risk, or do they cluster in highly exposed occupations where adaptability and continuous learning become essential?
- Does high *Ability to Thrive* (mobility potential) act as a buffer against displacement, allowing workers to pivot into less exposed roles?
- How does *Willingness to Stay* (tenure potential) interact with exposure—are individuals who remain in high-exposure occupations at greater long-term risk?

To address these questions, future work will combine the AI Exposure Index with the Career Projection Space, enabling joint analysis of exposure scores and talent dimensions. Methodologically, this will involve linking job task–level exposure measures to individual career trajectories, followed by regression and survival analysis to test associations between exposure and talent potential outcomes, as well as counterfactual simulations to explore how shifts in AI capability domains may alter career pathways. Through this integration, we aim to provide a unified perspective on how AI-driven technological change shapes not only the content of work but also the trajectories and potential of individual workers.

Chapter 7

Conclusion

This final chapter synthesizes the key insights, contributions, and broader implications of the research presented in this thesis. Building on a trilogy of studies that span network-aware occupational modeling, task-level semantic analysis, and capability-aligned AI exposure indexing, the chapter draws together a high-level reflection on how artificial intelligence is transforming the structure of work. It begins by summarizing the overarching contributions of the research, then discusses the societal and policy relevance of the findings in the context of workforce adaptation, inequality, and technological governance. The chapter also critically evaluates the limitations of the current approach, particularly in terms of data scope, modeling assumptions, and generalizability. Finally, it outlines promising directions for future research to advance the understanding of AI's impact on labor markets. This chapter thus closes the thesis by highlighting both the significance of the work completed and the opportunities that lie ahead.

7.1 Overview

Taken together, the four studies in this thesis form a coherent, multi-layered framework for understanding AI-driven transformation in the job market, progressing systematically from tasks to occupations and ultimately to individual careers. The first two chapters establish task- and occupation-level foundations by modeling explicit automation risks—such as substitution and complementarity—using supervised learning approaches grounded in expert knowledge and task semantics. These studies answer what types of work are exposed to automation and through which mechanisms. The third chapter generalizes this perspective by moving beyond predefined risk categories to an agnostic, capability-driven exposure framework that directly aligns evolving AI capabilities with job tasks and weights them by observable technological progress, capturing a broader spectrum of AI impacts including latent and future risks. Building on these exposure measures, the final chapter shifts the analytical focus from jobs to individuals, introducing a career mobility model that captures how workers move across positions, organizations, and hierarchical levels over time. This career modeling framework provides the structural foundation required to integrate AI exposure indices into longitudinal career trajectories, enabling future research on heterogeneous AI impacts at the individual level. Collectively, the thesis advances AI-driven job market research by unifying task-based risk assessment, capability-aware exposure measurement, and dynamic career modeling into a single, integrated analytical pipeline, offering both methodological innovation and a scalable foundation for policy, organizational, and workforce analysis. This thesis has presented a *granular, multi-layered framework for AI risk modeling* that bridges the gap between individual work tasks and broader occupational categories. Instead of treating occupations as monolithic entities, the research links whole occupations to their constituent tasks and further to the advancing AI capabilities that can perform those tasks. By fusing these layers, the study provides a unified perspective on how artificial intelligence may impact the labor market at different levels of granularity. Crucially, AI exposure is framed as a *dynamic, moving frontier* rather than a static snapshot, aligning risk measurement with the accelerating pace of technological progress. In doing so, this work extends the classic task-based approach by embedding each task in rich semantic representations and tying them to concrete AI benchmarks, thereby enriching the lens through which we assess technological vulnerability in jobs.

Methodologically, the thesis contributed a novel combination of techniques to assess automation risk. It leveraged network-based modeling to capture how risk can propagate through shared skills between occupations and applied state-of-the-art language models to understand the *textual content of tasks* at scale. Additionally, a capability-aligned matching approach was introduced to connect tens of thousands of human tasks with an ontology of AI tasks, tracing which specific technological capabilities map to which job activities. These complementary methods were synthesized into a *comprehensive AI Exposure Index* that weights each task by both its importance in the occupation and the recent performance trends of relevant AI benchmarks. The result is a forward-looking and fine-grained gauge of AI risk that captures not only *which* jobs might be affected, but *how and why* at the task level. At the individual level, the thesis further introduces a career mobility modeling framework that captures how workers transition across positions, organizations, and hierarchical levels over time. This model provides the structural foundation required to move beyond static job characteristics and enables future integration of AI exposure measures with longitudinal career trajectories.

Empirically, the findings revealed a significant subset of occupations that sit on the high-risk frontier of automation. At the same time, the task-level analysis exposed profound *within-occupation heterogeneity*, showing that even jobs traditionally considered "safe" from automation contain tasks highly automatable. Notably, the findings highlight the *unexpected vulnerability of education-related occupations*, challenging the prevailing assumption that knowledge-centric professions are insulated from automation. This richer understanding of job transformation dynamics provides a more actionable basis for adaptation and policy design. Beyond occupational findings, the career mobility analysis shows how hierarchical advancement, organizational context, and mobility patterns shape individual exposure and resilience to technological change. Together, these findings provide a richer, more actionable understanding of how AI reshapes work—not only in terms of which jobs are affected, but how impacts unfold across tasks, occupations, and individual careers—offering a robust empirical basis for workforce strategy, education policy, and long-term labor-market planning.

7.2 Societal and Policy Implications

The results carry important implications for workforce adaptation and policymaking. A central insight is that AI will not simply destroy jobs, but will *reconfigure the content of work*, redistributing tasks between humans and machines. This reinforces the need for *continuous workforce adaptation*, where individuals proactively cultivate skills that complement AI, such as creativity, critical thinking, and interpersonal capabilities.

From a policy perspective, the granular risk indices developed in this thesis enable *tar-*

geted interventions. Policymakers can use these insights to guide funding toward retraining programs in occupations with high automation risk, redesign curricula in education systems to prepare learners for AI-era demands, and incentivize firms to deploy AI in a *human-augmenting* rather than human-replacing manner. The identification of teachers as highly exposed, for example, suggests a need to integrate AI into educational workflows as a tool for augmenting instruction, not replacing educators.

Moreover, the research underscores the importance of addressing inequality. Automation risk is no longer concentrated solely in routine or low-wage jobs; rather, it increasingly affects medium- and high-skill roles. Without proactive measures, AI adoption could exacerbate *wage and employment disparities*, favoring those with the resources and flexibility to adapt. A just transition requires *inclusive upskilling pathways*, social safety nets for displaced workers, and career mobility programs to ensure that technological progress does not leave large segments of society behind.

In short, the granular task-level insights produced in this thesis can equip institutions with the diagnostic tools to *anticipate disruption and design inclusive, responsive, and forward-looking policies* that harness AI's potential while mitigating its dislocating effects.

7.3 Limitations

First, the analysis depends heavily on structured occupational datasets such as O*NET, which, although comprehensive, may not fully capture informal jobs, emerging roles, or cross-cultural differences. As a result, the findings may be less generalizable to developing economies or rapidly evolving industries.

Second, the risk assessments are based on the *technical potential* for automation, assuming that if AI can perform a task, it eventually will. In reality, adoption decisions are shaped by cost, regulation, social norms, and organizational strategy. As such, the exposure scores should be interpreted as indicative, not deterministic.

Third, while the study captures detailed information at a task level, it does not model dynamic changes over time or the emergence of entirely new tasks and occupations resulting from AI. Nor does it fully account for the contextual and institutional variables that mediate AI's real-world impact on labor.

Finally, although the models use advanced language representations and large-scale benchmarking, they rely in part on expert annotation and prompt-based AI judgments. These carry inherent subjectivity and are sensitive to biases in training data and prompt design.

7.4 Future Research Directions

Building on the contributions and addressing the limitations above, several directions are recommended for future exploration:

- **Real-time risk tracking.** Future models could incorporate live updates from AI benchmarks, research publications, patent activity, and job postings to build dynamic exposure indices that evolve alongside technological progress. Instead of static assessments, such systems would enable near real-time monitoring of shifting risks at the task, occupation, and industry level. This would provide policymakers and organizations with early-warning signals about which areas of the workforce are becoming newly vulnerable and where reskilling efforts should be prioritized.
- **Global comparisons.** Extending the proposed methods to international contexts could uncover how AI's labor market impact differs across institutional, cultural, and economic settings. For example, task structures in European economies, Asian manufacturing hubs, or emerging markets may display unique vulnerabilities or complementarities. Comparative analysis would allow us to test whether certain education systems, labor regulations, or industrial policies mitigate or amplify AI exposure, offering a more globally representative understanding of workforce transformation.
- **Linking AI exposure with individual-level talent potential.** A critical future direction is to integrate the AI Exposure Index developed in earlier chapters with the individual-level talent potential framework introduced in the forth work. This integration would make it possible to examine how external technological pressures interact with internal career dynamics. Future studies could link task-level exposure measures to career trajectories modeled by LADDER-GNN, apply regression and survival analysis to quantify relationships, and run counterfactual simulations to explore how accelerated AI progress in specific domains might reshape individual career pathways.

Such extensions would enrich our understanding of AI's labor implications and strengthen the practical utility of AI exposure models for both research and policy.

Bibliography

Acemoglu, D. (2021). Ai's impact on labor markets: Balancing policy and ethics. *Communications of the ACM*, *64*(10), 26–28. https://doi.org/10.1145/3488560

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In D. Card & O. Ashenfelter (Eds.). Elsevier. https://doi.org/https://doi.org/10.1016/S0169-7218(11)02410-5

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/10.1086/718327

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

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, *108*(6), 1488–1542. https://doi.org/10.1257/aer.20160696

Acemoglu, D., & Restrepo, P. (2017). *Robots and jobs: Evidence from us labor markets* (Working Paper No. 23285). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w23285

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Allen, R. C. (2009). Engels' pause: Technical change, capital accumulation, and inequality in the british industrial revolution. *Explorations in Economic History*, *46*(4), 418–435. https://doi.org/https://doi.org/10.1016/j.eeh.2009.04.004

Anelli, M., Colantone, I., & Stanig, P. (2021). Individual vulnerability to industrial robot adoption increases support for the radical right. *Proceedings of the National Academy of Sciences*, *118*(47), e2111611118. https://doi.org/10.1073/pnas.2111611118

Arntz, M., Gregory, T., & Zierahn, U. (2017). Revisiting the risk of automation. *Economics Letters*, *159*, 157–160. https://doi.org/10.1016/j.econlet.2017.07.001

Arntz, M., Gregory, T., & Zierahn, U. (2016). The risk of automation for jobs in oecd countries. https://doi.org/10.1787/5jlz9h56dvq7-en

Autor, D., & Handel, M. (2009). Putting tasks to the test: Human capital, job tasks and wages. *Journal of Labor Economics*, *31*. https://doi.org/10.1086/669332

Autor, D., Levy, F., & Murnane, R. (2003). The skill content of recent technological change: An empirical exploration. *Proceedings*, *118*. https://doi.org/10.1162/003355303322552801

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

Autor, D. H., & Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review*, *103*(5), 1553–97. https://doi.org/10.1257/aer.103.5.1553

Autor, D. H., Katz, L. F., & Krueger, A. B. (1998). Computing inequality: Have computers changed the labor market?*. *The Quarterly Journal of Economics*, *113*(4), 1169–1213. https://doi.org/10.1162/003355398555874

Bessen, J. E., Goos, M., Salomons, A., & Van den Berge, W. (2019). Automatic reaction—what happens to workers at firms that automate? *Boston Univ. School of Law, Law and Economics Research Paper No. 17-09*. https://www.bu.edu/law/files/2019/07/Automation_and_Working_Paper.pdf

Blagec, K., Barbosa-Silva, A., Ott, S., & Samwald, M. (2022). A curated, ontology-based, large-scale knowledge graph of artificial intelligence tasks and benchmarks. *Scientific Data*, *9*(1), 322. https://doi.org/10.1038/s41597-022-01435-x

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(4-5), 993–1022.

Bordes, A., Usunier, N., García-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, *26*, 2787–2795. https://papers.nips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html

Bowles, J. (2014, July). The computerisation of european jobs. https://www.bruegel.org/blog-post/computerisation-european-jobs

Braverman, H. (1998). *Labor and monopoly capital: The degradation of work in the twentieth century*. NYU Press. Retrieved June 15, 2025, from http://www.jstor.org/stable/j.ctt9qfrkf

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative ai at work. *The Quarterly Journal of Economics*, *140*(2), 889–942.

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., & McAfee, A. (2015). Will humans go the way of horses? labor in the second machine age. *Foreign affairs (Council on Foreign Relations)*, *94*, 8–14.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530–1534. https://doi.org/10.1126/science.aap8062

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.20181019

Brzeski, C., & Burk, I. (2015). *Die roboter kommen: Folgen der automatisierung für den deutschen arbeitsmarkt* (tech. rep. No. 30). ING-DiBa Economic Research. Frankfurt am Main, Germany.

Buyst, E., Goos, M., & Salomons, A. (2018). Job polarization: An historical perspective. *Oxford Review of Economic Policy*. https://doi.org/10.1093/OXREP/GRY003

Card, D., & Dinardo, J. (2002). Skill-biased technological change and rising wage inequality: Some problems and puzzles. *Journal of Labor Economics*. https://doi.org/10.1086/342055

Cessie, S. l., & Houwelingen, H. C. v. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 191–201.

Chen, M. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., & Inkpen, D. (2017). Enhanced lstm for natural language inference. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/p17-1152

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chui, M., Manyika, J., & Miremadi, M. (2016). Where machines could replace humans–and where they can't (yet). *McKinsey Quarterly*. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/where-machines-could-replace-humans-and-where-they-cant-yet

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Das, S., Steffen, S., Clarke, W., Reddy, P., Brynjolfsson, E., & Fleming, M. (2020). Learning occupational task-shares dynamics for the future of work. *arXiv preprint.* https://arxiv.org/abs/2002.05655

Dawson, N., Williams, M.-A., & Rizoiu, M.-A. (2021). Skill-driven recommendations for job transition pathways (S. Fu, Ed.). *PLOS ONE*, *16*(8), e0254722. https://doi.org/10.1371/journal.pone.0254722

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *J. Amer. Soc. for Information Science*, *41*(6), 391–407.

Deming, D., & Kahn, L. B. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, *36*(S1), S337–S369.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023, August). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. Retrieved November 28, 2023, from http://arxiv.org/abs/2303.10130

Felten, E., Raj, M., & Seamans, R. C. (2019). The Effect of Artificial Intelligence on Human Labor: An Ability-Based Approach. *Academy of Management Proceedings*, *2019*(1), 15784. https://doi.org/10.5465/AMBPP.2019.140

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Wang, D., Youn, H., & Rahwan, I. (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

Frazzetto, P., Haq, M. U. U., Fabris, F., & Sperduti, A. (2025). Graph neural networks for candidate-job matching: An inductive learning approach. *Data Science and Engineering*. https://doi.org/10.1007/s41019-025-00293-y

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)*, 148–156.

Frey, C., & Osborne, M. (2013). The future of employment: How susceptible are jobs to computerisation? *Oxford Martin*, *114*. https://doi.org/10.1016/j.techfore.2016.08.019

Gao, T., Fisch, A., & Chen, D. (2020). Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Goldfarb, A., Taska, B., & Teodoridis, F. (2023). Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings. *Research Policy*, *52*(1), 104653. https://doi.org/10.1016/j.respol.2022.104653

Goldin, C., & Katz, L. F. (2008). *The race between education and technology*. Harvard University Press. Retrieved June 15, 2025, from http://www.jstor.org/stable/j.ctvjf9x5x

Goos, M., & Manning, A. (2003). *Lousy and lovely jobs: The rising polarization of work in britain* (CEP Discussion Papers No. dp0604). Centre for Economic Performance, LSE.

Graetz, G., & Michaels, G. (2015). *Robots at work* (Discussion Paper No. 8938). IZA. https://ssrn.com/abstract=2589780

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, *18*(5-6), 602–610.

Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, *30*, 1024–1034. https://papers.nips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html

Hao, Y., & Li, S. (2021). Does firm visibility matter to debtholders? evidence from credit ratings. *Advances in Accounting, 52*, 100515. https://doi.org/https://doi.org/10.1016/j.adiac.2021.100515

Hernandez, D., & Brown, T. B. (2020). Measuring the algorithmic efficiency of neural networks. *arXiv preprint arXiv:2005.04305.*

Hinton, G. E., & Roweis, S. T. (2003). Stochastic neighbor embedding. *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, 833–840.

Ho, T. K. (1995). Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition (ICDAR 1995)*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hoffmeyer-Zlotnik, J. H. P., & Wolf, C. (2003). International standard classification of education, isced 1997. In *Advances in cross-national comparison: A european working book for demographic and socio-economic variables* (pp. 195–220). Springer US. https://doi.org/10.1007/978-1-4419-9186-7_10

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd). John Wiley & Sons.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 417–441, 498–520. https://doi.org/10.1037/h0071325

Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM 2013)*, 2333–2338. https://doi.org/10.1145/2505515.2505665

International Federation of Robotics. (2020, September). *World robotics 2020: Industrial robots.*

International Monetary Fund. (2025). *Ai adoption and inequality* (IMF Working Paper No. 2025/068). International Monetary Fund.

https://www.imf.org/en/Publications/WP/Issues/2025/04/04/AI-Adoption-and-Inequality-565729

Jaccoud, F. (2025). Robots & ai exposure and wage inequality: A within occupation approach. *Eurasian Business Review*. https://doi.org/10.1007/s40821-025-00306-w

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. *Business Horizons*, *61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Jiang, P., & Cai, X. (2024). A survey of text-matching techniques. *Information*, *15*(6), 332.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Karabarbounis, L., & Neiman, B. (2013). The global decline of the labor share*. *The Quarterly Journal of Economics*, *129*(1), 61–103. https://doi.org/10.1093/qje/qjt032

Katz, L. F., & Murphy, K. M. (1992). Changes in relative wages, 1963-1987: Supply and demand factors. *The Quarterly Journal of Economics*, *107*(1), 35–78. Retrieved June 15, 2025, from http://www.jstor.org/stable/2118323

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 3146–3154.

Khan, I. S., Ahmad, M. O., & Majava, J. (2023). Industry 4.0 innovations and their implications: An evaluation from sustainable development perspective. *Journal of cleaner production*, *405*, 137006.

Khandelwal, U., He, H., Qi, P., & Jurafsky, D. (2020). Generalization through memorization: Nearest neighbor language models. *International Conference on Learning Representations (ICLR)*.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1609.02907

Kogan, L., Papanikolaou, D., Schmidt, L. D. W., & Seegmiller, B. (2021, December). *Technology, vintage-specific human capital, and labor displacement: Evidence from* *linking patents with occupations* (Working Paper No. 29552). National Bureau of Economic Research. https://doi.org/10.3386/w29552

Kojima, T., Gu, S., Reid, M., Matsuo, Y., & Iwasaki, S. (2022). Large language models are zero-shot reasoners.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*, 221–232.

Kristal, T. (2020). Why has computerization increased wage inequality? information, occupational structural power, and wage inequality. *Work and Occupations*, *47*(4), 466–503. https://doi.org/10.1177/0730888420941031

Krugman, P. R. (1994). Past and prospective causes of high unemployment. *Economic Review*, *79*(Q IV), 23–43. https://ideas.repec.org/a/fip/fedker/y1994iqivp23-43nv.79no.4.html

Lacity, M., & Willcocks, L. P. (2017). *Robotic process automation and risk mitigation: The definitive guide*. SB Publishing.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. *8th International Conference on Learning Representations (ICLR)*.

Lane, M., & Saint-Martin, A. (2021). The impact of artificial intelligence on the labour market: What do we know so far? *OECD Social, Employment and Migration Working Papers*, *256*. https://doi.org/10.1787/7c895724-en

Langlotz, C. P. (2025). The effect of AI on the radiologist workforce: A task-based analysis. *medRxiv*. https://doi.org/10.64898/2025.12.20.25342714

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, 1188–1196. https://doi.org/10.48550/arXiv.1405.4053

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Li, L., Jing, H., Tong, H., Yang, J., He, Q., & Chen, B.-C. (2017). Nemo: Next career move prediction with contextual embedding. *Proceedings of the 26th International*

Conference on World Wide Web Companion, 505–513. https://doi.org/10.1145/3041021.3054200

Li, Y., Zhao, H., Jiang, H., Pan, Y., Liu, Z., Wu, Z., Shu, P., Tian, J., Yang, T., Xu, S., et al. (2024). Large language models for manufacturing. *arXiv preprint arXiv:2410.21418*.

Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2181–2187. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9742

Liu, J., Ng, Y. C., Wood, K. L., & Lim, K. H. (2020). Ipod: A large-scale industrial and professional occupation dataset. *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, 323–328. https://doi.org/10.1145/3406865.3418329

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.

Luo, M., Xu, X., Liu, Y., Pasupat, P., & Kazemi, M. (2024). In-context learning with retrieved demonstrations for language models: A survey. *ArXiv*, *abs/2401.11624*. https://doi.org/10.48550/arXiv.2401.11624

Mäkelä, E., & Stephany, F. (2024). Complement or substitute? how ai increases the demand for human skills. *arXiv preprint*. https://arxiv.org/abs/2412.19754

Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R., & Sanghvi, S. (2017). Jobs lost, jobs gained: Workforce transitions in a time of automation. https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages

Mariconda, S., & Lurati, F. (2014). Being known: A literature review on media visibility, public prominence and familiarity with implications for reputation research and management. *Corporate Reputation Review*, *17*(3), 219–236. https://doi.org/10.1057/crr.2014.11

Martínez-Plumed, F., Tolan, S., Pesole, A., Hernández-Orallo, J., Fernández-Macías, E., & Gómez, E. (2020). Does AI Qualify for the Job?: A Bidirectional Model Mapping Labour and AI Intensities. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 94–100. https://doi.org/10.1145/3375627.3375831

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, *27*(4), 12. https://doi.org/10.1609/aimag.v27i4.1904

McKinsey & Company. (2020). Global survey: The state of ai in 2020. *McKinsey Analytics.* https://www.mckinsey.com/.../Global-survey-The-state-of-AI-in-2020.pdf

Meng, Q., Zhu, H., Xiao, K., Zhang, L., & Xiong, H. (2019). A hierarchical career-path-aware neural network for job mobility prediction. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 14–24.

Micheli, A. (2009). Neural network for graphs: A contextual constructive approach. *IEEE Transactions on Neural Networks*, *20*(3), 498–511. https://doi.org/10.1109/TNN.2008.2010350

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 3111–3119.

Milano, S., McGrane, J. A., & Leonelli, S. (2023). Large language models challenge the future of higher education. *Nature Machine Intelligence*, *5*(4), 333–334.

Mokyr, J., Vickers, C., & Ziebarth, N. L. (2015). The history of technological anxiety and the future of economic growth: Is this time different? *Journal of Economic Perspectives*, *29*(3), 31–50. https://doi.org/10.1257/jep.29.3.31

Moro, E., Frank, M. R., Pentland, A., Rutherford, A., Cebrian, M., & Rahwan, I. (2021). Universal resilience patterns in labor markets. *Nature Communications*, *12*(1), 1972. https://doi.org/10.1038/s41467-021-22086-3

Mueller, J., & Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016)*, 2786–2792. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12094

National Center for O*NET Development. (2023). *O*NET database, release 28.1*. U.S. Department of Labor, Employment and Training Administration. https://www.onetcenter.org/database.html

National Research Council. (1983). *The long-term impact of technology on employment and unemployment*. The National Academies Press. https://doi.org/10.17226/19470

Nedelkoska, L., & Quintini, G. (2018). Automation, skills use and training. *OECD Social, Employment and Migration Working Papers*. https://doi.org/10.1787/2E2F4EEA-EN

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, *104*(1), 11–33. https://doi.org/10.1109/JPROC.2015.2483592

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192.

OECD. (2024a). *Artificial intelligence and wage inequality* (OECD Artificial Intelligence Papers No. 13). OECD. https://www.oecd.org/en/publications/artificial-intelligence-and-wage-inequality_bf98a45c-en.html

OECD. (2024b). *Emerging trends in ai skill demand across 14 oecd countries* (tech. rep.). OECD. https://www.oecd.org/en/publications/emerging-trends-in-ai-skill-demand-across-14-oecd-countries_7c691b9a-en.html

Osman, H. E., Yousef, M., Hakami, A. A., Jolha, L. M., Albasha, N., Daghmash, W. K., Takroni, Z. M., Patel, A. S., Mohammad, Y. J., Alsalamah, A., Albadri, S., & Khalid, A. (2025). Impact of artificial intelligence on the future of radiology: A national cross-sectional study among medical students and radiology professionals. *Journal of Radiation Research and Applied Sciences*, *18*(4), 102025. https://doi.org/https://doi.org/10.1016/j.jrras.2025.102025

Pajarinen, M., Rouvinen, P., & Ekeland, A. (2015). *Computerization threatens one-third of finnish and norwegian employment* (ETLA Brief No. 34). ETLA, The Research Institute of the Finnish Economy.

Peláez-Sánchez, I. C., Velarde-Camaqui, D., & Glasserman-Morales, L. D. (2024). The impact of large language models on higher education: Exploring the connection between ai and education 4.0. *Frontiers in Education*, *9*, 1392091.

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of ai on developer productivity: Evidence from github copilot. https://arxiv.org/abs/2302.06590

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP 2014)*, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proc. of NAACL-HLT 2018*, 2227–2237.

Phan, T. T., Pham, V. Q., Nguyen, H. D., Huynh, A. T., Tran, D. A., & Pham, V. T. (2021). Ontology-based resume searching system for job applicants in information technology. *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices (IEA/AIE 2021)*, 261–273. https://doi.org/10.1007/978-3-030-79457-6_23

PwC. (2025). Ai drives productivity and wage growth, pwc global study reveals. https://press.pwc.be/ai-drives-productivity-and-wage-growth-pwc-global-study-reveals

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training* (tech. rep.). OpenAI. https://cdn.openai.com/research%E2%80%90covers/language%E2%80%90unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners* (tech. rep.). OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proc. of the 2019 Conference on Empirical Methods in Natural Language*

Processing (EMNLP) and 9th International Joint Conference on Natural Language Processing (IJCNLP), 3973–3983.

Ricardo, D. (2015). *On the principles of political economy, and taxation*. Cambridge University Press.

Rodrik, D. (2015, February). *Premature deindustrialization* (Working Paper No. 20935). National Bureau of Economic Research. https://doi.org/10.3386/w20935

Rubin, O., Herzig, J., & Berant, J. (2022, July). Learning to retrieve prompts for in-context learning. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 2655–2671). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.191

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Santarelli, E., Carbonara, E., & Tripathi, I. (2025). Assessing the impact of ai on labor market outcomes: A meta-analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5126345

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80. https://doi.org/10.1109/TNN.2008.2005605

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. *The Semantic Web – ISWC 2018*, *11136*, 593–607. https://doi.org/10.1007/978-3-030-00668-6_38

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

Sperduti, A. (1993). Encoding labeled graphs by labeling RAAM. In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in neural information processing systems 6 (nips 1993)* (pp. 1125–1132). Morgan Kaufmann. http://papers.nips.cc/paper/860-encoding-labeled-graphs-by-labeling-raam

Sperduti, A., & Starita, A. (1997). Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, *8*(3), 714–735. https://doi.org/10.1109/72.572108

Sperduti, A., Starita, A., & Goller, C. (1995). Learning distributed representations for the classification of terms. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI 95), Montréal, Canada*, 509–517. http://ijcai.org/Proceedings/95-1/Papers/067.pdf

Stephany, F., & Teutloff, O. (2022). What is the price of a skill? the value of complementarity. *arXiv preprint*. https://arxiv.org/abs/2210.01535

Sun, L., Yuan, Y., Yao, Y., Li, Y., Zhang, H., Xie, X., Wang, X., Luo, F., & Stillwell, D. (2025). Large language models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity*, *57*, 101870. https://doi.org/https://doi.org/10.1016/j.tsc.2025.101870

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., & Wang, G. (2023). Text Classification via Large Language Models. https://doi.org/10.48550/arXiv.2305.08377

Symonds, Q. (2024). *Qs world university rankings*. https://support.qs.com/hc/en-gb/articles/4405955370898-QS-World-University-Rankings

Tolan, S., Pesole, A., Martínez-Plumed, F., Fernández-Macías, E., Hernández-Orallo, J., & Gómez, E. (2021). Measuring the Occupational Impact of AI: Tasks, Cognitive Abilities and AI Benchmarks. *Journal of Artificial Intelligence Research*, *71*, 191–236. https://doi.org/10.1613/jair.1.12647

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature medicine*, *25*(1), 44–56.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex embeddings for simple link prediction. *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, *48*, 2071–2080. https://proceedings.mlr.press/v48/trouillon16.html

UNESCO-UNEVOC. (2021). *Understanding the impact of artificial intelligence on skills development* (tech. rep.). UNESCO-UNEVOC International Centre for Technical, Vocational Education, and Training. https://unevoc.unesco.org/pub/understanding_the_impact_of_ai_on_skills_development.pdf

University of Technology Sydney. (2024). Graduate research candidature management: Thesis preparation and submission procedures. https://www.uts.edu.au/globalassets/sites/default/files/2024-02/grs-graduate-research-candidature-management-thesis-preparation-and-submission-procedures.pdf

U.S. Bureau of Labor Statistics. (2024). *Occupational employment and wage statistics (oews), 2012–2022*. U.S. Department of Labor. https://www.bls.gov/oes/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 5998–6008.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1710.10903

Vermeulen, B., Kesselhut, J., Pyka, A., & Saviotti, P. P. (2018). The impact of automation on employment: Just the usual structural change? *Sustainability*, *10*(5), 1661. https://doi.org/10.3390/su10051661

Wang, S., & Jiang, J. (2016). Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Wang, X., Jiang, Z., Peng, L., & Lu, Z. (2021). A deep-learning-inspired person-job matching model based on sentence vectors and subject-term graphs. *Complexity*, *2021*, 1–11. https://doi.org/10.1155/2021/6206288

Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014). Knowledge graph embedding by translating on hyperplanes. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, 1112–1119. https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8402

Webb, M. (2019). The impact of artificial intelligence on the labor market. *Available at SSRN 3482150*.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ..., & Le, Q. V. (2022). Chain of thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*.

World Economic Forum. (2025, April). *AI and jobs: International Workers' Day*. https://www.weforum.org/stories/2025/04/ai-jobs-international-workers-day/

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386

Xu, D., Yang, H., Rizoiu, M.-A., & Xu, G. (2022). Being automated or not? risk identification of occupations with graph neural networks. *International Conference on Advanced Data Mining and Applications*, 520–534.

Xu, D., Yang, H., Rizoiu, M.-A., & Xu, G. (2023). From occupations to tasks: A new perspective on automatability prediction using bert. *2023 10th International Conference on Behavioural and Social Computing (BESC)*, 1–7. https://doi.org/10.1109/BESC59560.2023.10386407

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1810.00826

Yang, A., & Kent, M. (2014). Social media and organizational visibility: A sample of fortune 500 corporations. *Public Relations Review*, *40*(3), 562–564. https://doi.org/https://doi.org/10.1016/j.pubrev.2014.04.006

Yang, B., Yih, W.-t., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6575

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 5753–5763.

Yoon, K., & McDonnell, M. D. (2020). Convolutional neural networks for sentence classification using 1d convolutions. *IEEE Access*, *8*, 37043–37056.

Zha, R., Qin, C., Zhang, L., Shen, D., Xu, T., Zhu, H., & Chen, E. (2024). Career mobility analysis with uncertainty-aware graph autoencoders: A job title transition perspective. *IEEE Transactions on Computational Social Systems, 11*(1), 1205–1215. https://doi.org/10.1109/TCSS.2023.3239038

Zha, R., Sun, Y., Qin, C., Zhang, L., et al. (2024). Towards unified representation learning for career mobility analysis with trajectory hypergraph. *ACM Transactions on Information Systems, 42*(4). https://doi.org/10.1145/3651158

Zhang, L., Zhou, D., Zhu, H., Xu, T., Zha, R., Chen, E., & Xiong, H. (2021). Attentive heterogeneous graph embedding for job mobility prediction. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2192–2201. https://doi.org/10.1145/3447548.3467388

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 649–657.