SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA

Department of Economics, Management, and Statistics
Ph.D. in Economics, Statistics, and Data Science, Cycle 37°
Curriculum Big Data & Analytics for Business

Novel approaches of Taxonomy enrichment via distributional semantics

ALESSIA DE SANTO
Num. 887637

Supervisor: Prof. Emilio Colombo
Co-Supervisor: Prof. Fabio Mercorio
Tutor: Prof. Matteo Maria Pelagatti
PhD. Program Coordinator: Prof. Matteo Manera

Academic Year 2024–2025

Abstract

In today's rapidly evolving labor market, the value of skill taxonomies such as the European Skills, Competences, Qualifications and Occupations (ESCO) and the Occupational Information Network (O*NET) is increasingly challenged by their static nature. While these resources provide a crucial backbone for Labour Market Intelligence (LMI), they struggle to keep pace with the emergence of new skills driven by digitalization and the green transition. To bridge this gap, Online Job Advertisements (OJAs) have emerged as an invaluable, real-time source of information on the demand side of the labor market. This reality frames the central question of this doctoral thesis: how can we systematically leverage the dynamic language of Online Job Advertisement (OJA)s to keep official skill taxonomies continuously updated? The research presented here, developed in close connection with institutional requests, offers a series of methodological contributions that blend advances in natural language processing with practical implementations for LMI. The thesis is structured in three main parts. The first provides the conceptual and empirical background. The second presents the core methodological contributions, showcasing a progression of enrichment pipelines:first a foundational data-driven pipeline for ESCO's green taxonomy; then the `TAXMAP` framework, a more sophisticated Artificial Intelligence (AI) approach for digital skills; and finally the `SkiLLens` pipeline, which extends these methods into a large-scale, multilingual system for detecting emerging skills across Europe. The third and last part focuses on the impact of these enriched taxonomies, using a case study on green skills to show how they enable more dynamic economic and social analyses, providing policymakers with timely intelligence on skill transitions, education requirements, and wage differentials.

Overall, the thesis argues that enriched and adaptive taxonomies are central infrastructures for the future of LMI, bridging methodological innovation with pressing policy needs. By connecting the fields of artificial intelligence and economic analysis, this work highlights the transformative potential of data-driven taxonomies and underscores that keeping them continuously updated is essential for advancing evidence-based research on the dynamics of work and skills.

Acknowledgment

Table of contents

List of figures

List of tables

List of Acronyms

AI Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

CBOW Continuous Bag-of-Words

CEDEFOP European Centre for the Development of Vocational Training

CRISP Interuniversity Research Centre for Public Services

ESCO European Skills, Competences, Qualifications and Occupations

EUROSTAT Statistical Office of the European Union

GNN Graph Neural Network

GPT Generative Pre-trained Transformer

HSS Hierarchical Semantic Similarity

IE Information Extraction

ISCO International Standard Classification of Occupations

LMI Labour Market Intelligence

LLM Large Language Model

LLMs Large Language Models

LM Language Model

ML Machine Learning

NLP Natural Language Processing

O*NET Occupational Information Network

OECD Organisation for Economic Co-operation and Development

OJA Online Job Advertisement

OJAs Online Job Advertisements

OLS Ordinary Least Squares

PILLARS Pathways to Inclusive Labour Markets

RCA Revealed Comparative Advantage

RSCA Revealed Symmetric Comparative Advantage

SG Skip-gram

SE Skill Extraction

SN Skill Normalisation

SR Semantic Retrieval

TEP Taxonomy Enrichment Problem

TF-IDF Term Frequency–Inverse Document Frequency

WIH Web Intelligence Hub

WIH-OJA Web Intelligence Hub – Online Job Advertisements

Introduction and Thesis Rationale

In an era defined by the rapid expansion of digital information and the transformative impact of AI, the need for structured and validated knowledge has never been more critical. Fundamentally, this challenge of structuring knowledge is addressed through taxonomies [108, 63, 78] —formal classification systems that organize concepts into a hierarchy of parent-child relationships. Their value is rooted in their ability to create a common language and a logical map for a given domain, serving as the semantic backbone for knowledge-intensive applications. For instance, in e-commerce, a product taxonomy (e.g., *Electronics* → *Computers* → *Laptops*) allows millions of users to navigate vast catalogues efficiently. In biology, taxonomies such as the *NCBI Taxonomy* are indispensable for classifying and integrating knowledge about living organisms. Their foundational role in enabling information retrieval, data integration, and knowledge discovery is well established across fields such as information science, Natural Language Processing (NLP), and knowledge management [77, 2].

A well-documented challenge of traditional taxonomies, however, lies in their construction and maintenance [52, 108, 121, 129]. The process has historically relied on manual curation by domain experts, a method that, while ensuring high quality, is inherently resource-intensive and slow. The resulting frameworks are consequently static, a characteristic that is manageable for stable domains but becomes a significant limitation when knowledge evolves rapidly. To overcome this bottleneck, the field of NLP has focused on automated Taxonomy Expansion, a task that aims to: "...automatically expand the taxonomy to incorporate these new concepts (without changing the existing relations in the given taxonomy)" [108]. This line of research seeks to make taxonomies more adaptive and responsive to change. This tension between static frameworks and dynamic realities is acutely felt in the labor market,

where interpreting complex information about skills and occupations is a crucial challenge for education, industry, and policy. Here, specific taxonomies like the European ESCO and the American O*NET provide an essential shared infrastructure. They offer a stable framework for classifying jobs, designing curricula, and informing policy. However, as digitalization and the green transition continuously reshape the world of work, these vital instruments struggle to keep pace. The very stability that once made them valuable now risks creating a growing gap between established classifications and the reality of the modern workforce, rendering them increasingly obsolete.

This doctoral thesis confronts this challenge by investigating how existing skill taxonomies can be evolved into dynamic, data-driven infrastructures capable of monitoring and understanding the modern European labor market. To this end, the thesis develops and validates a series of novel pipelines that leverage large-scale OJA data and advances in Natural Language Processing—from distributional semantics to Large Language Models (LLMs). This work showcases a clear methodological progression: from an initial data-driven pipeline for green skills to a more sophisticated, collaborative AI framework for digital skills, and finally to a large-scale, multilingual system for emerging skill detection. The research then moves beyond methodological innovation to demonstrate the real-world value of these enriched frameworks through applied economic analysis, providing timely evidence on skill transitions and wage differentials.

The research was developed during my doctoral studies at the University of Milano-Bicocca, in close collaboration with the Interuniversity Research Centre for Public Services (CRISP)[1], and through direct engagement with European institutions such as the European Centre for the Development of Vocational Training (CEDEFOP)[2] and Statistical Office of the European Union (EUROSTAT)[3]. This unique context was essential for grounding the research in applied policy needs and shaped its inherently multidisciplinary nature. It provided the framework to bridge advanced

[1]CRISP
[2]Cedefop Website
[3]Eurostat Website

techniques from NLP and AI with the demands of applied economic analysis and public policy—a synthesis that is a core tenet of the work presented.

The unifying thread of the different projects discussed is the conviction that dynamic, enriched taxonomies are key infrastructures for labor market intelligence, making it possible to track emerging skills, measure their diffusion, and assess their economic impact.

0.1 Contributions

This thesis makes several key contributions to the fields of LMI, NLP, and economic analysis. These contributions can be divided into two main categories: methodological and applied.

Methodological Contributions:

- **Enrichment of ESCO's Green Taxonomy:** The development and implementation of a novel data-driven pipeline to enrich the ESCO green skills taxonomy. This methodology leverages distributional semantics to analyze millions of online job advertisements, demonstrating how to identify and validate emerging green skills and link them to existing occupational structures.
- **The TAXMAP Framework for Digital Skills:** The design and formalization of TAXMAP, a collaborative framework that integrates LLMs with expert validation to augment the ESCO digital skills taxonomy. This contribution showcases a hybrid human-AI approach to taxonomy enrichment that balances automation with the need for high-quality, relevant results.
- **SkiLLens System for Emerging Skills:** The creation of SkiLLens, a multilingual system for detecting and mapping novel skills from job ads across Europe. This work provides a scalable, real-time solution for monitoring skill trends, addressing the limitations of static, manually curated taxonomies.

Applied Contributions:

- **Analysis of Green Skills and Jobs:** A comprehensive, data-driven analysis of the demand for green skills and the characteristics of green jobs in the European labor market. By applying the enriched green skills taxonomy, this

thesis provides timely evidence on skill transitions, educational requirements, and wage differentials, offering actionable insights for policymakers.

0.2 Collaboration and Dissemination

As previously mentioned, the research presented in this thesis is the result of different collaborative project in which I had the chance to bring my contribute. The work was framed by valuable interactions with European institutions: CEDEFOP and EUROSTAT, and carried out in close partnership with colleagues at CRISP. Specifically, the methodological pipelines for the enrichment of the green and digital taxonomies were developed under the supervision of Prof. Fabio Mercorio and Prof. Mario Mezzanzanica, with a key contribution from Simone D'Amico on the digital skills framework. The applied economic analysis of green jobs presented in Chapter 8 is the result of a joint effort with my supervisor, Prof. Emilio Colombo, and our colleague Francesco Trentini. Furthermore, the `SkiLLens` pipeline was a collaborative project with fellow researchers Lorenzo Malandri and Navid Nobani. I extend my sincere gratitude to all these colleagues and supervisors for their significant contributions and support.

The findings and methodologies developed have been disseminated through the following channels:

- The work on the `TAXMAP` [38] framework, presented in Chapter 5, was published in the proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing (SAC'25):
- The applied analysis of the green labor market, detailed in Chapter 8, was presented in 2025 at *"The economics of new technologies and global change" (ENTeG)* Workshop; at *"The Eight International (Astril) Conference"*; and at a CEDEFOP event, *"Using data from the web to shape next-generation labour market and skills analysis"*. It is right now under revision for the *Journal of the Association of Environmental and Resource Economists*.
- The foundational work on the green taxonomy enrichment (Chapter 4) was detailed in two internal reports for CEDEFOP.

- `SkiLLens` framework is, at the time of writing, under revision for *Knowledge-based Systems* journal.

0.3 Thesis Structure

The thesis is composed of three main parts, organized to guide the reader from foundational concepts to methodological contributions and, finally, to real-world impact.

1. **Part I – Background and Related Work:** This part establishes the theoretical and empirical foundations for the thesis. In **Chapter 1**, we review the evolution of techniques for representing meaning in text. It traces the progression from early frequency-based methods (e.g., Bag-of-Words) to static word embeddings (e.g., Word2Vec, FastText) and contextual models (e.g., Bidirectional Encoder Representations from Transformers (BERT)), culminating in modern Large Language Model (LLM). This chapter provides the technical background for the NLP methods used throughout the thesis. **Chapter 2** formally defines a taxonomy and the Taxonomy Enrichment Problem (TEP). It provides a structured overview of existing methods, categorising them as graph-based, textual, or hybrid, thereby positioning the thesis's contributions within the current state-of-the-art. **Chapter 3** contextualises the research within the domain of applied policy analysis. It discusses the role of taxonomies like ESCO in structuring labor market information and introduces OJAs as a critical data source for real-time intelligence.
2. **Part II – Methodology and Applications: Enriching Skill Taxonomies:** This part presents the core methodological contributions of the thesis, showcasing three enrichment pipelines developed in collaboration with institutional partners. In Chapter 4, we detail the first pipeline, which uses distributional semantics (FastText) to identify and validate new green skills from OJAs. It demonstrates how to construct a data-driven "skill pillar" and link it to an "occupation pillar," thereby enriching the official ESCO taxonomy with emerging, market-relevant terminology. Results of this part are discussed in two internal

reports. **Chapter 5** introduces a more advanced, collaborative LLM-based framework. It presents `TAXMAP`, a system that combines the semantic power of multiple LLMs with human-in-the-loop validation to enrich the ESCO digital skills taxonomy. Finally, **Chapter 6** expands on these approaches by presenting `SkiLLens`, a multilingual pipeline designed to detect, validate, and map emerging skills on a pan-European scale. This chapter discusses skill enrichment across 28 countries, incorporating a large-scale expert validation process to ensure the real-world relevance of the identified skills.

3. **Part III – Impact of Enriched Taxonomies:** The final part demonstrates the economic and policy relevance of the enriched taxonomies developed in the thesis. In **Chapter 8**, we apply the enriched green skills taxonomy to a large-scale dataset of nearly 29 million European job ads. This chapter provides a detailed empirical analysis of the green labor market, examining the characteristics of green jobs, their skill requirements, and their impact on wages and education, including in traditionally non-green ("brown") occupations. **Chapter 8.6.5** concludes by summarizing the key findings and discussing their implications for future policy and research.

Part I

Background and Related Work

Chapter 1

Word Representations: From Conventional Models to Embeddings

Representing words and documents in numerical form is a central task in NLP. Over the years, several methods have been developed to encode words as vectors that reflect their meaning and usage in context. These representations are interpretable, can be combined and compared through mathematical operations, and provide a solid foundation for many Machine Learning (ML) applications.

The earliest methods, known as *conventional* or *frequency-based models*, focused on converting text into a numerical format that early algorithms could process. Later, a paradigm shift led to distributional representation models, or static word embeddings, where a word's context is used to learn a single, dense vector imbued with semantic meaning. Most recently, this concept has evolved into contextual models, which create dynamic word representations computed directly from their immediate context. This chapter traces this evolution, beginning with the foundational frequency-based techniques.

1.1 Frequency-Based Representations

A foundational challenge in computational linguistics is the conversion of unstructured text into a structured, numerical format. Early approaches relied on direct

feature extraction, treating words as discrete symbols to be counted. These methods are best understood not as models of meaning, but as frequency-based vectorization techniques that create semantically impoverished, yet computationally useful, representations.

Bag of Words (BoW) The Bag of Words (BoW) model is the most intuitive of these techniques. It represents a piece of text as an unordered collection—a "bag"—of its words, disregarding grammar and word order entirely. The resulting vector for a document has a dimension for every unique word in the corpus vocabulary, and the value in each dimension is the word's frequency.

Statement 1: The meal was delicious.
Statement 2: The food was tasty.

A human understands these sentences to mean the same thing. However, their BoW representations share only the stopword "the". As shown in Table 1.1, the model would perceive them as highly dissimilar because "meal" is as different from "food" as it is from "car".

Table 1.1 BoW representation for semantically similar sentences.

	the	meal	was	delicious	food	tasty
S1	1	1	1	1	0	0
S2	1	0	1	0	1	1

The BoW representation of these statements is shown in Table 1.1. While simple, the BoW model suffers from profound limitations. It is fundamentally incapable of capturing semantic meaning; the words "food" and "meal" are treated as equally unrelated. Furthermore, its disregard for syntax means sentences with different meanings can produce identical vectors. Other limitations include vector sparsity and the disproportionate influence of frequent words, which can distort similarity measures [5].

N-grams The N-gram model was developed as a direct response to BoW's inability to capture word order. By treating contiguous sequences of n words as the atomic units to be counted, it preserves local context. For $n = 1, 2, 3$, the models are termed uni-gram, bi-gram, and tri-gram respectively. Considering again the two sentences from the BoW example, their bi-gram representation is shown in Table 1.2.

Table 1.2 Bi-gram representation for semantically similar sentences.

	the meal	meal was	was delicious	the food	food was	was tasty
S1	1	1	1	0	0	0
S2	0	0	0	1	1	1

While this refinement captures some phrasal context, it does not solve the core semantic problem and exacerbates the issue of dimensionality, as the vocabulary of n-grams is orders of magnitude larger than that of single words.

Term Frequency-Inverse Document Frequency (TF-IDF) Term Frequency–Inverse Document Frequency (TF-IDF) is a more sophisticated weighting scheme used to address a key flaw in the BoW model: that frequent words (like "the" or "was") dominate the representation. TF-IDF assesses the importance of a term by balancing its frequency in a specific document (Term Frequency) against its rarity across the entire collection of documents (Inverse Document Frequency). This allows the model to find terms that are characteristic of a given document.

The TF-IDF score is computed as:

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D) \tag{1.1}$$

where $\text{tf}(t, d)$ is the relative frequency of term t within document d:

$$\text{tf}(t, d) = \frac{f_{td}}{\sum_{t' \in d} f_{t'd}} \tag{1.2}$$

and $\text{idf}(t, D)$ is the logarithm of the ratio between the total number of documents (N) and the number of documents containing the term t:

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{1.3}$$

Although TF-IDF provides a more nuanced vector representation by down-weighting common terms, it remains a frequency-based method and does not inherently capture semantic relationships between different words.

1.2 Static Word Embeddings: Capturing Meaning in a Vector Space

The limitations of frequency-based models necessitated a paradigm shift from counting words to understanding their meaning. This shift was catalyzed by the application of neural networks to natural language, giving rise to **word embeddings**. These models are not based on explicit counting but on the distributional hypothesis: that a word's meaning can be inferred from the contexts in which it appears. By training a neural network on a massive text corpus, these models learn to represent each word as a dense, low-dimensional vector in a way that captures its semantic properties.

This learned representation is the crucial distinction. The resulting vector space encodes meaningful relationships, where words with similar meanings are mapped to nearby points. This geometric arrangement famously allows for algebraic operations that mirror semantic relationships, such as 'vector('king') - vector('man') + vector('woman')' yielding a vector proximate to 'vector('queen')'. The transition to embeddings marks a move from representing words as isolated identifiers to representing them as points in a rich, semantic space.

Formally, a word embedding e can be defined as a lookup table that maps words from a vocabulary V to continuous vectors in $\mathbb{R}^D$:

$$\begin{aligned} e: \quad & V \to \mathbb{R}^D \\ & w \mapsto e(w) = v_w^e \end{aligned} \tag{1.4}$$

This mapping is learned through an optimization process, resulting in a vector space where geometric relationships reflect semantic ones [115]. A common distinction between different types of embedding is between static and contextualised embeddings:

- **Static word embeddings** use the distributional hypothesis to learn a single, global vector for each word, condensing all its contextual uses from the training corpus.
- **Contextualised word embeddings** provide variable vectors. Word representations are multiple and are computed dynamically from their surrounding context.

1.2.1 Models in Euclidean Space

The most prominent static embedding models operate in Euclidean space, the familiar "flat" geometry that is well-suited for capturing symmetric relationships like similarity and analogy.

Word2vec

The Word2vec algorithm, introduced by Mikolov et al. [82], represented a breakthrough in applying the distributional hypothesis, as it implemented it through a two-layer neural network. The authors proposed two complementary architectures: the Continuous Bag-of-Words (CBOW) model, which predicts a target word from its surrounding context words, and the Skip-gram (SG) model, which performs the inverse task of predicting context words from a given target word.

The Skip-gram architecture, which often yields superior results for infrequent words, formalizes its objective as maximizing the average log probability across a sequence of training words $w\_1, ..., w\_T$:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{1.5}$$

where c is the size of the context window around the center word w_t. The success of Word2vec was also driven by key implementation details that made training highly efficient on massive corpora. These included methodological refinements such as negative sampling, which simplifies the optimization problem, and hierarchical softmax, which provides a computationally efficient approximation of the full softmax.

FastText

Developed by Facebook AI Research, fastText [14] is an extension of the word2vec skip-gram model. Its key innovation is to learn vectors for character n-grams in addition to full words. Each word is then represented as the sum of its constituent character n-gram vectors. This sub-word information allows fastText to generate embeddings for out-of-vocabulary (OOV) words and makes it more robust to typos and rare words. Formally, given a word w, and a dictionary of size G, G_{w} is the set of n-grams of size G appearing in w. Denoting as z_g the vector representation of the n-gram g, w will be represented as the sum of the vector representation of its n-grams and the score associated to the word w as:

$$\int(\mathrm{w},\mathrm{c}) = \sum_{g\in G_{\mathrm{w}}} z_g^{\mathsf{T}} v_{\mathrm{c}} \tag{1.6}$$

where v_{c} is the vector representing the context. This simple representation allows one to share information between words, and this makes it useful to represent rare words, typos, and words with the same root. Moreover, it allows the computation of representations for words not seen during the training phase, called Out Of Vocabulary (OOV) words.

GloVe

GloVe (Global Vectors) [98] is another foundational model that combines the strengths of two major approaches: the local context window methods of word2vec and the global matrix factorization methods (like Latent Semantic Analysis). GloVe is trained on a global word-word co-occurrence matrix, which tabulates how frequently words appear together in the corpus. Its training objective is to learn word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence.

1.2.2 A Special Case for Hierarchies: Non-Euclidean Embeddings

While powerful for capturing semantic similarity, the Euclidean geometry underlying these foundational models is ill-suited for efficiently representing asymmetric, hierarchical relationships. In a flat space, it is difficult to embed tree-like structures, which are central to taxonomies.

To address this limitation, researchers have explored **non-Euclidean embeddings**, particularly those in hyperbolic space. The geometry of hyperbolic space naturally resembles a continuous version of a tree, making it highly effective at modelling hierarchical structures. The Poincaré ball model [93] was a pioneering approach that learned hierarchical representations by embedding them in hyperbolic space, such that the distance between nodes directly reflects their semantic similarity and hierarchy. While more specialized, these models are of high relevance for tasks involving taxonomies and knowledge graphs [94, 50].

1.3 Contextual Embeddings

The primary limitation of all static models—whether Euclidean or hyperbolic—is their inability to handle polysemy. A word like "bank" has the same single vector representation regardless of whether it refers to a financial institution or a river's edge. This limitation motivated the next major leap in NLP: the development of **contextualised word embeddings**.

These models do not generate a single, static vector for each word. Instead, they compute a word's representation dynamically based on the entire sentence or context in which it appears. The embedding for "bank" in "the river bank" is therefore different from its embedding in "the world bank".

ELMo One of the first contextualised embedding model was ELMo [100], which leveraged large-scale pre-training to transfer internal representations from bidirectional language models (BiLSTM) to downstream tasks as contextual word embeddings. Unlike static embeddings, ELMo assigns a different representation to a word depending on its context within the sentence. This is achieved through a bidirectional design: the model takes into account both the preceding and the following words when generating the embedding.

Formally, given a sequence of N tokens $(t_1, t_2, \ldots, t_N)$, ELMo optimises the likelihood of the sequence in both directions. A forward language model estimates the probability of token t_k given its history $(t_1, t_2, \ldots, t_{k-1})$, while a backward language model performs the reverse, predicting a token from its future context [5].

The introduction of the Transformer architecture marked a significant turning point, proving more scalable and effective than recurrent networks like LSTMs.

The GPT (Generative Pre-Training) Later models moved from recurrent architectures such as BiLSTM to transformer decoders. Transformer-based approaches, even with larger parameter counts, have proven more effective and scalable [115]. The Generative Pre-trained Transformer (GPT) family of models exemplifies this transition. GPT employs a unidirectional (left-to-right) language model and fine-tunes pre-trained parameters on downstream tasks.

For a sequence of tokens $(t_1, t_2, \ldots, t_N)$, the language modelling objective is to maximise:

$$L_1(X) = \sum \log P(t_i | t_{i-N}, \ldots, t_{i-1}; \theta) \tag{1.7}$$

GPT relies on a multi-layer transformer decoder with self-attention, where multi-headed attention captures dependencies across tokens, and position-wise feed-forward layers refine contextual representations [5]. GPT introduced the now-

standard transformer decoder architecture for language modeling, showing that pretraining a unidirectional language model on large corpora could be successfully fine-tuned for downstream tasks.LLMs - that will be discussed in the next section - build directly on this paradigm but then have expanded it using ten to hundreds of billions of parameters.

Bidirectional Encoder Representations from Transformers (BERT) BERT [41] revolutionized the field by successfully training a deep, bidirectional Transformer model. Unlike GPT, which could only look at the left context, BERT's pre-training tasks allow it to learn from the entire sentence at once. This is achieved through two novel objectives:

- **Masked Language Modeling (MLM)**: Randomly masking some percentage of the input tokens and training the model to predict the original masked word based on the unmasked context.
- **Next Sentence Prediction (NSP)**: Training the model to predict whether two sentences are sequential in the original text.

The deep bidirectional representations produced by BERT became the new state-of-the-art for a wide range of NLP tasks.

1.4 From Contextual Embeddings to LLMs

The evolution from static word embeddings to contextual models such as ELMo and BERT represented a foundational shift in natural language processing. For the first time, lexical representations were no longer single, fixed vectors but context-sensitive embeddings capable of capturing polysemy and nuanced usage. However, the capabilities of these encoder-based models remain fundamentally descriptive; they are limited when a task demands richer world knowledge, relational inference, or concept-level reasoning, as is characteristic of taxonomy enrichment.

LLMs represent a significant architectural and functional leap beyond this paradigm. Since 2018, large-scale transformer-based models—including the GPT series [103, 104, 20], PaLM [28], LLaMA [116], and GPT-4 [96]—have extended the principle of contextual representation. Their breakthrough rests on a simple training

principle: predicting the next word in a sequence across massive corpora of text. By iteratively learning to predict tokens from vast amounts of data, LLMs internalize not only syntactic regularities but also semantic relations, factual knowledge, and even world knowledge [65]. For example, such models can infer that roses, dahlias, and peonies are all types of flowers, that "enormous" is a stronger synonym of "big," or that Virginia Woolf is the author of *A Room of One's Own*.

This predictive, autoregressive training[1] - where the model learns to forecast each next token based on all preceding ones - allows LLMs to generalize contextual representations into scalable, generative systems. The research on LLMs is evolving very rapidly and in many different ways. Contemporary surveys note their emergent abilities in few-shot prompting, reasoning, and knowledge inference [84]. In practice, this means that LLMs no longer serve only as embedding providers but also as active agents for tasks such as skill extraction, taxonomy enrichment, and dynamic concept alignment in LMI. Fig.1.1 illustrates established and emerging abilities of these models.

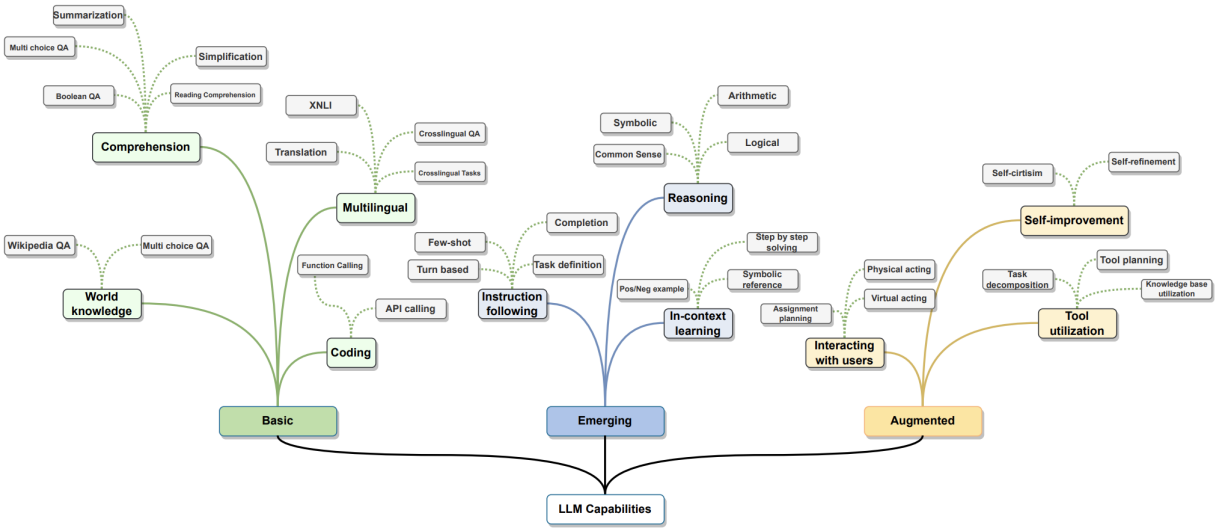


Fig. 1.1 LLM Capabilities, from Minaee et al. [83]

Critically for this thesis, the application of LLMs to knowledge-rich tasks reframes their role from passive feature extractors to *active knowledge agents*. While

[1]For technical details about their structure see, among others, [65] and [84] for a detailed survey on LLMs evolution.

LLMs can produce high-quality embeddings that outperform earlier models [114], their true potential lies in their generative and reasoning capabilities. Recent studies demonstrate this by leveraging LLMs to predict taxonomic relations [88], enrich hierarchical structures with minimal supervision [143], and expand taxonomies in dynamic environments through collaborative mapping [133]. These approaches illustrate a key transition: LLMs can be prompted not merely to represent concepts, but to actively suggest new nodes, validate hierarchical positions, and propose semantic relations—functions that static or contextual embeddings alone cannot perform.

Despite these advancements, LLMs-based methodologies introduce critical challenges, including the risk of hallucination, inconsistent reasoning over deep hierarchies, and the significant computational cost of fine-tuning. Recent surveys on integrating LLMs with knowledge bases discuss these trade-offs and outline hybrid strategies that combine symbolic and neural methods to mitigate such risks [133].

In summary, this thesis builds upon the foundation of static and contextual embedding methods but extends it by operationalizing the advanced capabilities of LLMs for taxonomy enrichment. The methodological chapters that follow will demonstrate how LLMs can be employed not just as passive embedding providers, but as active participants in the workflow to update, align, and expand taxonomies to reflect evolving labor market demands.

1.5 Implications for This Research

This thesis makes use of both static and contextual embeddings, reflecting both their complementary strengths and the institutional needs driving the research:

- **FastText** embeddings were used to enrich the ESCO green taxonomy, where scalability and robustness to out-of-vocabulary words were crucial.
- **Contextual embeddings from transformer-based LLMs** powered TAXMAP, improving precision in matching digital skills to ESCO and handling semantic ambiguities across heterogeneous sources.

The progression from static to contextual embeddings, and more recently to large language models, mirrors both the technological evolution of NLP and the trajectory of this research: from practical taxonomy enrichment tasks to more ambitious, data-driven skill intelligence infrastructures.

Chapter 2

Background on Taxonomy Enrichment

2.1 The Structure and Function of Taxonomies

Taxonomies provide a structured representation of knowledge by organizing concepts into a semantic hierarchy governed by the *is-a* relation [19]. This relationship, known in linguistics as hyponymy, connects a specific term (the *hyponym*) to its broader category (the *hypernym*) through a transitive and asymmetric link that forms the backbone of the hierarchy [59]. Such hierarchical knowledge is vital for a wide range of NLP applications, from enhancing query understanding in search engines to powering personalized recommendations in e-commerce [127, 109]. Despite the existence of large-scale knowledge bases like WordNet [47] and Wikipedia Categories, their manual construction and maintenance create a significant bottleneck. These generic hierarchies often fail to keep pace with the rapid evolution of specialized or emerging domains [121], creating a persistent need for automated methods to expand and refine them—a task known as taxonomy enrichment.

2.2 Formalizing Taxonomies and Their Enrichment

To rigorously define the problem of taxonomy enrichment, we must first establish a formal definition of a taxonomy itself.

2.2.1 A Formal Definition of Taxonomy

From a knowledge representation perspective, a taxonomy distinguishes between abstract concepts and the lexical terms that refer to them. The formalisation proposed by [73] captures this two-layered structure.

Definition 2.2.1 (Two-Layer Taxonomy)**.** A taxonomy $\mathcal{T}$ is a 4-tuple $\mathcal{T} = (C, \mathcal{W}, \mathcal{H}^c, \mathcal{F})$, where:

- C is a set of abstract concepts (nodes).
- $\mathcal{W}$ is a set of words or entities (the lexical layer).
- $\mathcal{H}^c \subseteq C \times C$ is a directed acyclic relation defining the hierarchy between concepts (e.g., $\mathcal{H}^c(\text{Canine}, \text{Mammal})$).
- $\mathcal{F} \subseteq C \times \mathcal{W}$ is a relation mapping words to the concepts they instantiate (e.g., $\mathcal{F}(\text{Canine}, \text{"dog"})$).

While this two-layered view is formally robust, many contemporary NLP approaches, particularly those focused on automated graph-based enrichment, adopt a more streamlined, practical definition. In this view, the distinction between concepts and words is collapsed, and every term is treated as a node in a single graph. This thesis will adopt the following operational definition, which aligns with recent literature in the field [108, 53].

Definition 2.2.2 (Operational Taxonomy)**.** A taxonomy $\mathcal{T}$ is a pair $\mathcal{T} = (C, H_C)$, where:

- C is a set of concepts or terms belonging to the domain of interest (nodes).
- $H_C \subseteq C \times C$ is a directed acyclic binary relation between concepts.

$H_C(c_1, c_2)$ means that concept c_1 is the *hypernym* of c_2, and c_2 is the *hyponym* of c_1.

2.2.2 The Taxonomy Enrichment Problem

Building on this operational definition, we can now formally define the TEP. While the broader problem of taxonomy enrichment can also involve identifying and merging synonyms (a form of horizontal expansion), this thesis adopts the formulation most prevalent in the literature. We define the TEP specifically as the task of **vertical**

enrichment: attaching new, more specific terms (hyponyms) to their correct parent concepts (hypernyms) within the existing hierarchy. This focus is consistent with the primary goal of increasing the taxonomy's granularity and aligns with the methods discussed in the literature review (Section 2.3).

Definition 2.2.3 (Taxonomy Enrichment Problem)**.** The Taxonomy Enrichment Problem (TEP) is a 3-tuple $(\mathcal{T}, \mathcal{D}, \mathcal{S})$, where:

- $\mathcal{T} = (C, H_C)$ is an existing seed taxonomy, as in Definition 2.2.2.
- $\mathcal{D}$ is a set of new query terms to be integrated into the taxonomy.
- $\mathcal{S} : C \times \mathcal{D} \rightarrow [0, 1]$ is a scoring function that estimates the relevance of attaching a new term $t \in \mathcal{D}$ as a hyponym to an existing concept $c \in C$.

A solution to the TEP is an enriched taxonomy $\mathcal{T}^+ = (C \cup \mathcal{D}, H_{C^+})$, where H_{C^+} contains the original relations H_C along with a new set of relations linking new terms to existing concepts, i.e., $H_{C^+} = H_C \cup \{(c, t) \mid c \in C, t \in \mathcal{D}\}$.

As established in prior studies [113, 78], we maintain the assumption that the core taxonomy $\mathcal{T}$ remains unchanged. This approach preserves the integrity of the high-level, expert-curated structure. Our focus, therefore, is on the integration of fine-grained terms under existing concepts. Furthermore, in alignment with convention [78], we do not establish hierarchical relationships among the newly introduced query terms in $\mathcal{D}$. This methodology enriches the taxonomy with greater specificity without disrupting the established hierarchical order. Figure 2.1 shows an example of taxonomy enrichment.

2.3 Taxonomy Enrichment Methods

The task of automatically enriching taxonomies has evolved significantly, shifting from rule-based systems to sophisticated neural architectures. Unless otherwise specified, the methods discussed in this review focus on the problem of vertical enrichment—attaching new terms as hyponyms under existing concepts. We provide an overview of this landscape by categorizing the literature along two primary axes:

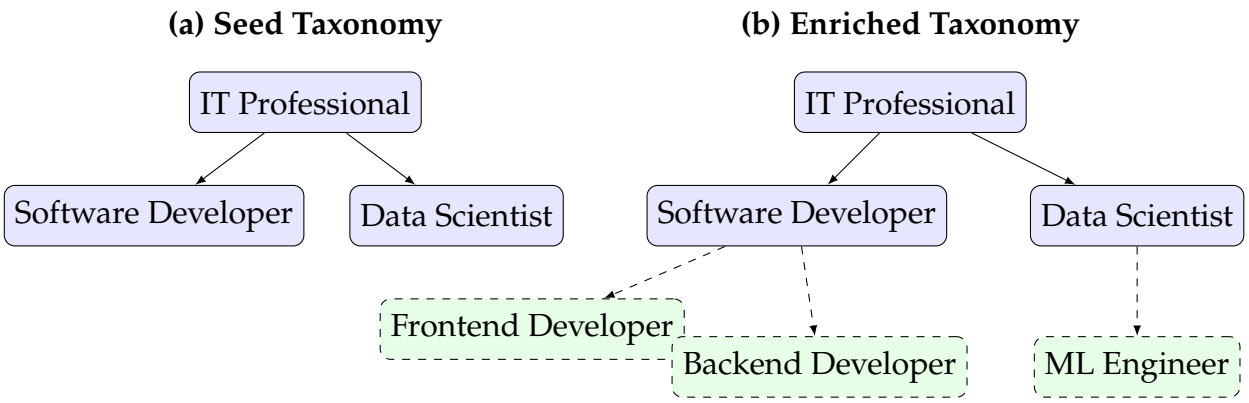


Fig. 2.1 An illustration of the taxonomy enrichment problem. Given a simple seed taxonomy (a), the task is to attach new, more specific query terms (e.g., 'Frontend Developer') to their correct parent concepts, resulting in an enriched taxonomy (b).

- **Feature Representation**: Whether a method primarily relies on the taxonomy's explicit *graph structure* or the *textual semantics* of its concepts. Many modern approaches use a *hybrid* of both.
- **Structural Depth**: Whether the model considers only the *local* context of a concept (e.g., its direct parent and siblings) or leverages the *global* hierarchical structure (e.g., paths to the root).

Table 2.1 summarizes the key works discussed according to these dimensions.

Table 2.1 Taxonomy Enrichment Methods according to feature representation and structural depth.

Contribution	**Feature Representation**			**Structural Depth**	
	Hybrid	*Graph*	*Textual*	*Global*	*Local*
TAXOEXPAN [109]		x			x
HYPEREXPAN [71]		x			x
STEAM [136]		x			x
HEF [131]		x		x	
MUSUBU [113]			x		x
TMN [139]			x		x
TEMP [69]			x	x	
TAXOPROMPT [132]			x	x	
OCTET [78]	x				x
TEAM [101]	x				x
GENTAXO [137]	x			x	
VTE [144]	x				x

2.3.1 Graph-Based Methods: Leveraging Taxonomic Structure

A significant body of work explicitly models the taxonomy as a graph, using Graph Neural Network (GNN) to learn structural representations. These methods are particularly adept at encoding the relational context of a potential attachment point.

Most graph-based methods focus on the local neighborhood of a concept. **Taxo-Expan** [109], a pioneering work in this area, introduced a position-enhanced GNN to encode the local structure of a candidate parent (an "anchor concept"). It constructs an ego-network for each anchor and uses position embeddings to distinguish relatives like parents and siblings, providing rich local context for the attachment decision. Similarly, **HYPEREXPAN** [71] uses a hyperbolic GNN, leveraging hyperbolic geometry's natural ability to model hierarchical structures within a local ego-graph. **STEAM** [136] is a self-supervised taxonomy expansion model which samples mini-paths from the existing taxonomy, and formulates a node attachment prediction task between anchor mini-paths and query terms. To solve the node attachment task, it learns feature representations for query-anchor pairs from multiple textual and lexico-syntactic views, and performs multi-view co-training for prediction.

While local methods are effective, some approaches aim to incorporate a more global view. **HEF** [131], for example, introduces a "coherence modeling" module that constructs an "ego-tree" containing all ancestors of a node, thereby capturing a deeper vertical slice of the taxonomy to ensure the new term fits logically within the broader hierarchy.

2.3.2 Textual Methods: The Power of Language Models

Another major line of research prioritizes the rich semantic information in the textual descriptions of concepts, often leveraging large pre-trained Language Model (LM).

In low-resource settings, **MUSUBU** [113] generates natural language queries from term pairs using Hearst patterns (e.g., "a {parent} such as a {child}") and feeds them to an LM-based classifier, effectively probing the relational knowledge already encoded in the LM. Other methods focus on providing LM with global structural context. **TEMP** [69] is a self-supervised taxonomy expansion method, which predicts

the position of new concepts by ranking the possible taxonomy paths they could be placed in. The authors employ pre-trained contextual encoders to capture hypernym-hyponym relations. **TAXOPROMPT** [132] introduces a framework that enriches existing taxonomies by generating hypernyms for query concepts through prompt-based methods. It enhances the expansion process by integrating the global structure of taxonomies into the language model used in the prompt via a random walk algorithm and by incorporating taxonomic and descriptive information through prompt tuning.

2.3.3 Hybrid Approaches

Recognizing the complementary strengths of structure and semantics, many state-of-the-art models adopt hybrid approaches. **OCTET** [78] exemplifies this by building a heterogeneous representation combining structural graph embeddings, semantic word embeddings, and lexical string features. **TEAM** [101] integrates a Graph Attention Network (GAT) to capture local structure with FastText embeddings for semantics, uniquely framing the problem as a multi-task challenge of deciding between an "attach" (new child) or "merge" (new synonym) operation. This makes it a notable exception in the literature; whereas most methods focus exclusively on hierarchical attachment, TEAM is explicitly designed to also handle the expansion of synsets[1] by identifying new synonyms. **GENTAXO** jointly encodes taxonomic sentences and top-down/bottom-up subgraphs to classify candidate positions and even generate concept names, bridging structure and text [137]. Finally, **VTE** brings vision into the loop, aligning textual hypernyms and visual prototypes of hyponyms in a shared space to detect hypernymy when image evidence is informative [144].

[1] A synset (short for "synonym set") is a group of words or phrases that are considered synonyms for a specific, shared concept

2.4 Foundational Tasks for Labor Market Taxonomy Enrichment

The formal definition of the Taxonomy Enrichment Problem (TEP) presupposes a set of new query terms, $\mathcal{D}$, to be integrated into an existing hierarchy. In the context of this thesis, where we aim to enrich skill taxonomies using data from Online Job Advertisements (OJAs), this set $\mathcal{D}$ is not readily available. Instead, potential new skills are embedded within vast amounts of unstructured text. Therefore, two critical upstream tasks must be addressed before the enrichment itself can occur: Skill Extraction (SE), the process of identifying skill mentions in raw text, and Skill Normalisation (SN), the task of mapping these extracted mentions to the canonical concepts within the reference taxonomy. The following sections will formally define and review the state-of-the-art for both foundational tasks. These concepts are particularly central to the methodologies developed for the green skills taxonomy enrichment (Chapter 4) and the multilingual SkiLLens framework (Chapter 6).

2.4.1 Skills Extraction

SE can be viewed as a specialised application of Information Extraction (IE) in the labor market domain. IE is the process of identifying and extracting structured information of predefined types from unstructured natural language texts [86].

Formally, let $\mathcal{T} = \{t_1, \ldots, t_n\}$ be the set of texts, $\mathcal{P} = \{p_1, \ldots, p_m\}$ the set of information types, and $\mathcal{A}$ the universe of atomic items.

Let $\tau : \mathcal{A} \to \mathcal{P}$ assign each item its type.

$$\mathrm{IE} : \mathcal{T} \times \mathcal{P} \to 2^{\mathcal{A}}, \qquad \mathrm{IE}(t, p) = \{\, a \in \mathcal{A} \mid \mathrm{occ}(a, t) \ \wedge \ \tau(a) = p \,\}. \tag{1}$$

$$\mathrm{occ}(a, t) \overset{\text{def}}{\iff} \exists 1 \le i \le j \le |t| \text{ such that } t_i \ldots t_j = a.$$

For a fixed pair (t_i, p_j) the output is a finite set

$$\mathrm{IE}(t_i, p_j) = \{a_1, \ldots, a_{k_{ij}}\}. \tag{2}$$

Drawing from this formalisation, we define *Skill Extraction (SE)* as the specific IE task of identifying and extracting skill entities from unstructured textual data commonly found in the labor market—such as job advertisements, resumes, or job descriptions. Let $\mathcal{T}$ be the set of unstructured text documents (e.g., job ads, résumés), $\mathcal{P}$ the set of information types, and $\mathcal{S}$ the universe of canonical skill entities. For a document $t_i \in \mathcal{T}$ and a target type fixed to skills, $p_s = \textsc{skill} \in \mathcal{P}$, the goal is to return all skill entities present in t_i.

We define the skill–extraction function

$$\mathrm{SE} : \mathcal{T} \times \{\textsc{skill}\} \to 2^{\mathcal{S}}, \qquad (t_i, p_s) \mapsto S(t_i, p_s), \tag{2.1}$$

which yields a finite set of recognised skills

$$\mathrm{SE}(t_i) = S(t_i, p_s) \;=\; \left\{ s \in \mathcal{S} \,\middle|\, s \preceq t_i \right\}, \tag{2.2}$$

where $s \preceq t_i$ denotes that at least one span in t_i realises the skill entity s.

Early approaches to skill extraction relied on exact matching, combining manual annotation with semantic clustering (e.g., Word2Vec) and ontology construction to classify skills from job-related texts [21, 57, 61]. To handle variability in skill terminology, fuzzy matching techniques were introduced, comparing extracted phrases with controlled vocabularies like ESCO using similarity metrics such as Levenshtein Distance and Jaccard Similarity [17]. Unsupervised topic modelling, particularly Latent Dirichlet Allocation (LDA), has been used to uncover latent skill structures by applying it directly to job descriptions or using a domain-specific vocabulary [58, 39]. Deep learning further advanced the field by treating skill extraction as sequence tagging or multi-label classification, using convolutional

networks and ranking-based methods [68, 64, 56]. Recently, transformer-based models have shown promising results by leveraging contextual embeddings, typically by fine-tuning BERT or SpanBERT with classification layers (e.g., CRFs) or adapting pretrained models for domain-specific recruitment tasks [12, 140, 10, 45].

2.4.2 Skills Normalisation

Skills Normalisation (SN) can be cast as a specialised instance of Semantic Retrieval (SR) in the labour-market setting. In SR, items are retrieved based on semantic similarity, typically by encoding text into a vector space.

Semantic Retrieval (SR). Let $Q = \{q_1, \ldots, q_n\}$ be a set of queries and $\mathcal{E} = \{e_1, \ldots, e_m\}$ a set of target elements. Assume each item is represented by an embedding, and let $\text{sim}(\cdot, \cdot)$ be a similarity function (e.g., cosine). For each $q_i \in Q$, SR returns the most similar target:

$$\text{SR}(q_i; \mathcal{E}) = \arg\max_{e \in \mathcal{E}} \text{sim}(q_i, e). \tag{2.3}$$

Skills Normalisation (SN). Let $\mathcal{S} = \{s_1, \ldots, s_k\}$ be the set of skill mentions extracted from OJAs, and let $\mathcal{E}_{\text{ESCO}}$ denote the set of canonical skills in ESCO (each with a preferred label), represented as embeddings. SN maps each extracted mention to its canonical ESCO entry:

$$\text{SN}(s_i; \mathcal{E}_{\text{ESCO}}) = \arg\max_{e \in \mathcal{E}_{\text{ESCO}}} \text{sim}(s_i, e) = \hat{s}_i \in \mathcal{E}_{\text{ESCO}}. \tag{2.4}$$

Aggregating over all mentions gives the normalised set

$$\text{SN}(\mathcal{S}; \mathcal{E}_{\text{ESCO}}) = \{\hat{s}_1, \ldots, \hat{s}_k\} \subseteq \mathcal{E}_{\text{ESCO}}, \tag{2.5}$$

where each $\hat{s}_i$ corresponds to the ESCO entry whose *preferred label* is the closest in meaning to the extracted mention s_i.

Few works address mapping skills to ESCO. We summarize them in Tab. 2.2.

SkillNER [46] extracts soft skills from texts but does not normalize them to ESCO entries. Kompetencer [141] categorizes skills into 23 ESCO-aligned groups

Table 2.2 Comparison of Skill Normalization Approaches

Framework	Mapping Approach	Lang.	Maps to ESCO Skills?	Used?	Reproducibility	Notes
Kompetencer [141]	Rule-based classification to 23 ESCO categories	en, da	No	No	Data: ✓ Code: ✓	Only high-level classes.
ESCOXLM-R [142]	Pretrained LM, data from [141]	en, fr, da, de	No	No	Data: ✓ Code: ✓	Only high-level classes.
SkillGPT [67]	LLM: summarization + vector search	en, fr, nl	Yes	No	Data: ✗ Code: ✓	No validation.
Decorte et al. [40]	Skill extraction + ESCO mapping	en	Yes	Yes	Data: ✓ Code: ✓	Granular ESCO mapping.
Clavie et al. [30]	Zero-shot LLM to ESCO entries	en	Yes	Yes	Data: ✓ Code: ✓	Granular mapping (en only).

without using the ESCO skills for the mapping. The same dataset is used to evaluate ESCOXLM-R [142], a pretrained language model for skills extraction and classification. SkillGPT [67] employs a language model for skill extraction and standardization but lacks empirical validation. Closest to our work are [40] and [30], which map extracted skills to ESCO; the latter adopts a two-step zero-shot pipeline. In Section 6.2 we benchmark `TAXMAP` against these methods and show that it outperforms both.

2.5 Implications for This Research

The field of taxonomy enrichment has made remarkable progress, moving from local, single-modality methods to more holistic models that integrate global structure with rich semantics. The dominant paradigm involves self-supervised training on an existing seed taxonomy, where models learn to re-insert known concepts and then apply this knowledge to attach new query terms.

Despite these advancements, several critical challenges and research gaps remain:

- **The Prototypical Hypernym Problem:** Many models trained on semantic similarity struggle to distinguish between true hyponyms and other strongly related concepts. For example, a model may incorrectly attach "Apple Juice" to "Fruit" due to its strong semantic association with "Apple", failing to recognize the correct hypernym is "Juice".

- **Static and Generic Domains:** Most benchmark datasets are derived from well-structured, relatively static domains like computer science or general knowledge (WordNet). Methods developed on this data may not generalize well to domains that are more abstract, dynamic, or specialized.
- **Reliance on Rich Textual Definitions:** Several state-of-the-art methods depend on the availability of detailed textual definitions for concepts. These are often unavailable in real-world, domain-specific taxonomies, such as product catalogs or skills classifications, where concepts may be represented by short, ambiguous phrases.

The domain of **green skills**, for example, presents a unique combination of these challenges. Concepts are often abstract (e.g., "Sustainability Mindset"), rapidly evolving with policy and technology, and lack the clean, hierarchical structure found in traditional domains. Few, if any, existing methods have been designed to handle the specific ambiguities and complexities inherent in such a socio-technical skills taxonomy.

This thesis aims to address this gap by developing and applying a data-driven methodology for the enrichment of institutional skill taxonomies, with a particular focus on the domain of green and digital skills. The proposed approach builds upon recent advances in distributional semantics and labour market intelligence to extract and validate new skill concepts emerging from OJAs. In doing so, it aims to demonstrate how large-scale, unstructured labour market data can be systematically leveraged to complement and extend existing frameworks such as ESCO.

To put this into practice, our enrichment pipelines frame the Taxonomy Enrichment Problem as a similarity-ranking task within a vector space. The core of our methodology involves representing both the existing concepts within a seed taxonomy (e.g., ESCO) and the new candidate terms as dense numerical vectors, or embeddings. The attachment of a new term to its most appropriate parent concept is then determined using **cosine similarity** as our primary scoring function. Cosine similarity measures the cosine of the angle between two vectors; in our context, a score closer to 1 signifies a strong semantic relationship, as the terms' vector representations point in a similar direction. By calculating this score between a new

skill and all potential parent concepts, we can identify the most plausible hypernym and systematically enrich the taxonomy. This foundational approach is applied and extended in the following chapters to enrich both the green and digital skill taxonomies.

Chapter 3

Labor Market Intelligence

In recent years, European labour demand channelled through specialised web portals has surged, catalysing the emergence of *Labour Market Intelligence* (LMI): the design and application of AI/ML algorithms and end-to-end frameworks to analyse labour-market data for decision support (e.g., [138, 123, 51]). Today, monitoring and interpreting labour-market change both *timely* and at *fine-grained* geographical resolution is practically critical. Examples include estimating the effects of robotisation on occupations in the US [48] and quantifying skill relevance using the O*NET standard taxonomy [1].

In Europe, Cedefop launched a large-scale initiative in 2016 to collect and classify online job vacancies OJAs across the then 28 EU Member States, using ESCO as the reference hierarchy and handling the Union's multilinguality. A typical OJA contains a mix of structured and unstructured information, presenting a rich but complex data source, as illustrated by the following stylized example:

OJA Example

Digital Marketing Manager –Future Growth Solutions.
We are seeking an experienced Digital Marketing Manager to develop and execute our online strategy. The ideal candidate will be responsible for managing all digital campaigns, from SEO/SEM to social media and email marketing, to drive brand awareness and lead generation. This role requires a high degree of *creativity* and strong *analytical skills* to monitor campaign performance, optimize for ROI, and report on key metrics using tools like Google Analytics. You will work closely with our content and sales teams, so excellent *teamwork and communication skills* are essential. The position demands proven experience in *digital advertising*, proficiency in *project management*, and the ability to handle multiple priorities in a fast-paced environment. Familiarity with CRM software is a definite plus.

Such advertisements are rich sources of information, but their unstructured nature—blending formal skills, specific tools, and soft competencies—poses a significant analytical challenge. Early results from the Cedefop initiative focused on lexicon analysis from OJAs and the identification of novel occupations and skills [52, 17, 16, 74]. The availability of classified OJAs has, in turn, enabled third-party studies: for instance, [31] used OJAs to estimate the impact of AI on job automation and to measure the salience of digital/soft skills; during the COVID-19 pandemic (May 2020), CEDEFOP used OJAs to create the Cov19R index to identify workers at higher exposure risk[1].

While ESCO supports cross-country comparability, it does not—by itself—encode country-specific market peculiarities: both the *mix of skills* advertised and the *semantics* of terms vary by national context and maturity of local labour markets. This motivates LMI pipelines that (i) link unstructured vacancy text to shared taxonomies and (ii) *enrich* those taxonomies with emerging terminology.

The role of taxonomies in LMI. Taxonomies are pivotal to LMI because they provide the semantic scaffold for mapping raw signals (skills, tasks, occupations) into interpretable structures. Several contributions illustrate this role. Alabdulkareem et

[1] https://www.cedefop.europa.eu/en/news-and-press/news/cedefop-creates-cov19r-social-distancing-risk-index-which-eu-jobs-are-more-risk

al. [1] analyse the relevance and polarisation of skills within the US O*NET taxonomy, showing how taxonomy-grounded skill signals reveal structural shifts. Giabelli et al. propose *WETA*, a domain-independent method that uses distributional semantics and classification for automatic taxonomy alignment, and demonstrate bridging between the Italian occupation taxonomy and ESCO [53]. Malandri et al. [75] employ word embeddings to refine/expand taxonomies and test on ESCO with European OJAs, highlighting how data-driven similarity can surface missing nodes and edges. In the business domain, Arslan and Cruz [4] use BERTopic to automatically augment a given taxonomy with concepts extracted from online news, showing how topic models can support taxonomy maintenance when domains evolve quickly. Given their central role, it is crucial to understand the design and characteristics of the specific taxonomies that underpin modern LMI.

3.1 Key Taxonomies: A Comparative Analysis of ESCO and O*NET

In the fields of labor economics, education, and workforce development, occupational and skills taxonomies provide the critical infrastructure for analyzing and managing human capital. Among the most influential are Europe's ESCO classification and the United States' O*NET. While both aim to structure labor market information, they are predicated on different designs and objectives.

ESCO: A Multilingual Framework for European Labor Market Integration ESCO is a multilingual classification system managed by the European Commission. Its primary goal is to support job mobility across Europe and foster a more integrated and efficient labor market by offering a common language for occupations and skills. This "common language" facilitates better communication between the education and training sector and the EU labor market. ESCO is available for free and can be accessed through the official ESCO portal. For a comprehensive overview, the European Commission also provides a detailed ESCO handbook.

The classification is structured upon three interconnected pillars:

- **Occupations**: This pillar contains descriptions of almost 3,000 occupations relevant to the European labor market. Each occupation is mapped to the International Standard Classification of Occupations (ISCO).
- **Skills and Competences**: This pillar acts as a thesaurus of over 13,500 concepts, detailing the knowledge, skills, and competences needed for the occupations. It systematically shows the relationships between these different concepts.
- **Qualifications**: This pillar provides a framework for organizing information on qualifications, such as degrees and certificates, awarded by national authorities.

A defining feature of ESCO is its availability in 27 languages, which is crucial for ensuring interoperability between national systems and supporting both employers and jobseekers across the EU.

O*NET: A Data-Rich Content Model of the U.S. Economy The Occupational Information Network, O*NET , sponsored by the U.S. Department of Labor, is the primary source for occupational information in the United States. It provides a comprehensive, free online database designed to help students, job seekers, and professionals understand the U.S. world of work. The architecture of O*NET is based on a **Content Model**, a robust framework that organizes information into six major domains: **Worker Characteristics**: Enduring attributes of a worker, such as abilities, interests, and work styles; **Worker Requirements**: Attributes developed through education and experience, including knowledge and skills; **Experience Requirements**: Backgrounds related to previous work, including training, certifications, and licensing; **Occupational Requirements**: Characteristics of the work itself, such as work activities, context, and organizational context; **Workforce Characteristics**: Variables that describe the broader labor market, such as wages and employment outlook;**Occupation-Specific Information**: Detailed information unique to a specific job, including tasks, tools, and technology.

To ensure the database remains current, data is continuously collected from a large, statistically selected sample of incumbent workers across hundreds of occupations each year. This data, gathered through standardized web-based and

paper questionnaires, provides a real-world foundation for the detailed descriptions in O*NET.

Key Conceptual and Structural Distinctions A primary distinction lies in the explicit separation of skills and tasks within the O*NET framework.

- In **O*NET**, **Skills** are defined under "Worker Requirements" (e.g., "Critical Thinking") and are rated for importance and level. **Tasks**, under "Occupational Requirements," are specific work activities (e.g., "Analyze data to resolve operational problems.").
- **ESCO** does not maintain this formal structural separation. While task-like descriptions are present within occupational profiles, skills are linked directly to occupations without the intermediary layer of specific work activities.

In summary, ESCO provides a broad, multilingual framework designed for interoperability, while O*NET offers a highly detailed, data-driven model that systematically differentiates between the requirements of the work and the attributes of the worker.

3.2 Implication for this research

In this thesis, LMI is both a *data source* (OJAs and related corpora) and a *validation ground* for taxonomy methodologies. We (i) link labour-market text to ESCO using distributional semantics and later LLM-based mapping, and (ii) enrich ESCO with emergent skills and terms, enabling policy-relevant indicators (e.g., greenness, pervasiveness) and country-specific interpretations while retaining cross-country comparability.

Part II

Taxonomy Enrichment via Semantic Representations: From Word Embeddings to LLMs

A Comparative Framework for Taxonomy Enrichment

The following three chapters constitute the methodological core of this thesis, presenting a series of frameworks designed to bridge the gap between static labor market taxonomies and the dynamic reality of OJAs. These frameworks represent both distinct technical approaches and they reflect the rapid evolution of NLP between 2021 and 2025 and were developed in close collaboration with European institutions to address specific, real-world analytical needs.

Chapter 4 focuses on the enrichment of green skills. This framework was conceived during the 2021-2022 period, a time when LLMs had not yet reached the maturity or accessibility required for stable, production-grade labor market intelligence. Developed in the context of collaborations with Cedefop and Eurostat, the pipeline relies on static word-embedding models (e.g., FastText). At that time, these models represented the gold standard for institutional use due to their computational efficiency and their ability to capture broad distributional shifts in emerging terminology. This approach remains highly appropriate for researchers working in stable, high-resource language settings where the primary goal is identifying new terms rather than resolving deep semantic ambiguities. As the focus shifts to the digital domain in Chapter 5, the methodology evolves to reflect the shift toward contextual semantic representations. The `TAXMAP` system was designed to address the high degree of polysemy found in technological terms, where technical nuances require deeper disambiguation than static models can provide. Moving into the era of Transformer architectures, this pipeline prioritizes analytical fidelity and precision. It is the preferred choice for fine-grained, within-

occupation analyses where the cost of a "false positive" mapping is high and where the computational resources to deploy contextual models are available. Finally, the challenge of pan-European scalability is addressed in Chapter 6 through the `SkiLLens` pipeline. Developed within the framework of the PILLARS project, this system synthesizes previous lessons to confront a diverse linguistic landscape of 22 languages and millions of observations. This framework argues for a hybrid approach: leveraging the speed of embeddings for large-scale candidate discovery while employing LLMs for the complex task of cross-lingual refinement and expert-led validation. This represents the most appropriate solution for international policy-monitoring applications that require a harmonized, cross-border view of skill trends. Viewed collectively, these chapters demonstrate that taxonomy enrichment is not a one-size-fits-all process. The transition from the static methods of 2021 to the hybrid LLM pipelines of 2025 illustrates a movement toward higher semantic resolution and broader geographic reach. By navigating the trade-offs between computational overhead, contextual precision, and multilingual scalability, this Part provides a modular toolkit for maintaining the relevance of labor market intelligence in an era of constant technological change.

Chapter 4

Enriching the ESCO Green Skills Taxonomy

This chapter presents the first strand of methodologies developed in the thesis for *taxonomy enrichment*, focusing on data-driven techniques based on distributional semantics (Word2Vec, FastText) applied to OJAs. Building on the conceptual framework and literature reviewed in the previous chapters, we implement the idea that occupations can be characterised—and ultimately enriched—through their observable skills and tasks in the labour market. Concretely, we develop an enrichment pipeline for the ESCO green skills taxonomy that (i) prepares and filters OJAs, (ii) learns word embeddings and selects models via a hierarchical semantic similarity metric aligned to ESCO, and (iii) constructs two linked pillars: a *skill pillar* (green terms and their validated linguistic variants) and an *occupation pillar* (ESCO occupations quantified by green pervasiveness and greenness).

Methodologically, this chapter showcases how classic word-embedding approaches can surface both generic and emerging green terminology and align it back to ESCO through semantic similarity and conservative matching rules. Substantively, the enriched taxonomy supports monitoring and analysis of green skill demand across occupations.

Chapter 5 extends the methodological arc from distributional word embeddings to LLMs, showing how contextual and generative representations can further generalize the enrichment framework and support additional mapping and validation tasks.

Together, the two chapters provide a coherent progression of semantic methods for enriching occupational and skills taxonomies.

The empirical work documented in the following sections has been collected in two internal reports produced within the Web Intelligence Hub – Online Job Advertisements (WIH-OJA) collaboration with CEDEFOP,[1] which this chapter synthesises and reframes within the thesis narrative.

4.1 Introduction and Motivation

The urgency of the climate crisis and the increasing demand for sustainability have led most countries to recognize the importance of transitioning to greener models of production and consumption. This transformation has created new employment opportunities in the so-called green economy, while simultaneously underscoring the need for appropriate knowledge, abilities, and competences—*green skills*—to design, implement, and use environmentally sustainable technologies across a variety of sectors [18].

Understanding how the green transition reshapes the demand for skills is crucial for policy and labour market actors. For instance, identifying the degree of similarity between green and non-green jobs can help determine the extent of retraining required for workers to adapt to greener occupations [18]. Yet, this process is slowed by the lack of consensus on what constitutes a green occupation, task, or skill [6]. Multiple initiatives have therefore sought to define and classify green work, often contributing to the enrichment of existing occupational taxonomies (citarne alcune) .

In this context, being part of the Crisp research center I had the chance to collaborate with Cedefop on the *Web Intelligence Hub* project (WIH-OJA, 2021–2023)[2], that, as one of its many tasks, had the goal to augment the ESCO green skills taxonomy using large-scale evidence from OJAs.

[1] Internal reports: *D1.1 Cedefop Green Skills* (2022) and *OF7 Green Skills* (2023).

[2] The project was called "Towards the European Web Intelligence Hub – European system for collection and analysis of online job advertisement data (WIH-OJA)".

The main contributions of this work were: Firstly, as we just mentioned, online jobs ads data were used in constructing the taxonomy, so that the latter results as fully data-driven since it catches terms as they appear in OJAs. Then, and this was a crucial point in shaping our work, the methodology and the taxonomy we constructed assume that green occupations are defined by their skills and tasks content and not vice versa as it is in Top-down builded taxonomies: in our work, the skills are the ones that shape an occupation and define whether (and how much) it is green or not. Therefore a green occupation is one that must include green skills or tasks and for which its greenness is defined by the skills/task found in the market. Then, our contribution is directly linked to ESCO taxonomy equipped with a semantic similarity value obtained through AI techniques developed by CRISP[54]. Exploiting ESCO content of skills and occupations we were able to directly connect the skills and tasks obtained from the job market data with the ESCO occupations. In this way, we are contributing to the enrichment of the taxonomy with new skills, and providing a clear view of how the ESCO occupations are affected by green skills.

4.2 Related work

In recent years, interest in green skills has intensified, with significant contributions from actors such as LinkedIn, the Organisation for Economic Co-operation and Development (OECD), and UNIDO [6], yet major gaps remain in terms of data coverage and data-driven approaches. This raises key research and policy questions: do taxonomies like O*NET or ESCO adequately reflect the skill requirements of emerging green jobs? Can we capture the upcoming trends in the greening of industries? To what extent will these transformations require reskilling or adaptation of education programmes?

One of the first systematic efforts was undertaken by O*NET in the United States. Its *Green Economy Program*, launched in 2009 [42], relied on academic and technical sources to classify around 100 occupations more closely involved in the green economy. Using the so-called "output approach" [15], occupations were categorized into three groups: (i) existing occupations in high demand due to the

greening of the economy (*Green Increased Demand*); (ii) occupations undergoing significant changes in task content (*Green-Enhanced Skills*); and (iii) new occupations emerging in the green economy (*New & Emerging Green*) [124]. This classification later inspired the "task approach" [125, 126, 124], which used granular ONET task data[3] to derive continuous measures of "greenness" and to identify the general skills most intensively used in greener occupations. The result was a set of 16 skills, primarily engineering and technical, defined as green [124].

Table 4.1 Green Skills from O*NET Taxonomy Vona [124]

	Engineering & Technical
2C3b	Engineering and Technology
2C3c	Design
2C3d	Building and Construction
2C3e	Mechanical
4A3b2	Drafting, Laying Out, and Specifying Technical Devices, Parts, and Equipment
4A1b3	Estimating the Quantifiable Characteristics of Products, Events, or Information
	Operation Management
2B4g	Systems Analysis
2B4h	Systems Evaluation
4A2b3	Updating and Using Relevant Knowledge
4A4b6	Provide Consultation and Advice to Others
	Monitoring
2C8b	Law and Government
4A2a3	Evaluating Information to Determine Compliance with Standards
	Science
2C4b	Physics
vero2C4d	Biology

In Europe, the ESCO framework sought to systematize available knowledge and incorporate green skills and occupations into its taxonomy. Adopting CEDEFOP's definition of green skills as the "knowledge, abilities, values, and attitudes needed

[3]It is important to highlight how the specific structure of O*NET allows for this distinction. O*NET contains, in fact, detailed information – coming from workplace survey, occupational experts and experts analysis - on both tasks (e.g., what workers are expected to do at the workplace – the 'demand side') and skills (e.g., the abilities and competences that workers should possess to perform work tasks - the 'supply side'). See Sec.3.1.

to live in, develop and support a society which reduces the impact of human activity on the environment" [24], ESCO implemented a three-step expert- and machine-learning-based procedure that labelled 571 skills and knowledge concepts as green (381 skills, 185 knowledge concepts, and 5 transversal skills). However, these remained relatively generic (e.g. *sustainability*, *waste management*) and did not capture more specialized or emerging labour market terminology (e.g. *flood risk assessment*, *ecologistics*).

4.3 Methodology

The methodological process for constructing the green skill taxonomy involved several distinct stages, transforming raw OJAs into a structured and semantically coherent representation of skills. This approach builds on a distributional semantics framework, combining data preprocessing, intrinsic model evaluation, and large-scale embedding training. The methodology is organised into two main phases: the first phase (§4.3.1) concerns data preparation and the selection of the most suitable embedding model. The second phase (§4.3.2) leverages this model to extract, enrich, and validate green-related terms, forming the basis of the new green skill pillar.

4.3.1 Preliminary steps and embedding selection

This phase focuses on the creation of a linguistically robust corpus and the intrinsic evaluation of embedding quality with respect to ESCO's semantic structure. It includes three main steps, summarised below.

Step 1: Setting the pipeline. This step prepares the OJAs dataset to be processed by the word-embedding algorithm to capture similarities and co-occurrences. It is data-independent but language-dependent, so it can be applied to any language dataset. The preparation includes:

(i) **Pre-processing**: apply standard text cleaning, including (a) lowercasing; (b) converting numbers to words or removing numbers; (c) removing punctuation; (d) removing accent marks and diacritics; (e) trimming extra whitespace; (f)

expanding frequent abbreviations (e.g., *aka*, *asap*); (g) removing stop words[4], sparse terms, and nuisance tokens; (h) lemmatisation.

(ii) **Identify green OJAs**: retain only OJAs containing a list of *sentinel* green words (partly provided by Cedefop and partly discovered later via data-driven enrichment).

(iii) **n-gram generation**: we scan sentences to form multiword expressions (up to 4-grams). For example, "problem" and "solving" become the single token "problem_solving". `problem_solving`.

(iv) **Skills anchor**: To reduce the noise and improve the overall embedding quality, each OJA is converted into a set of texts that are more likely to encode skills. The idea is to use sentinel words[5] as anchors of a window that should only contain the skill-related part of the OJAs to restrict word-embedding processing only to skill sentences to remove the bias. Notably, this step helps in removing green terms related to the description of the company, rather than the job that is expected to be green.

At the end of Step 1, the corpus is transformed into a set of normalised *OJA skill sentences* ready for embedding.

Step 2: Build the HSS to select the best embedding. The quality of identified green skills critically depends on the quality of the underlying embeddings. To evaluate model quality intrinsically and ensure taxonomic faithfulness, we use the Hierarchical Semantic Similarity (HSS) metric. HSS estimates how well an embedding preserves semantic neighbourhoods implied by a taxonomy (here, ESCO). Higher HSS indicates better preservation of ESCO's hierarchical relations in the embedding space. In some recent research activities, we implemented and refined the HSS to

[4]This was done using NLTK package from python and a list of additional stopwords related to the job market identified by the authors.

[5]A term we use to indicate particular words that are considered appropriate to identify relevant portions of the OJV descriptions. Example of the words used are: 'Qualification', 'qualifications' 'skill', 'skills', 'task', 'tasks' 'duties', 'role', 'abilities', 'background', 'knowledge', 'competence', 'ability', 'duty','candidate','candidates', 'priorities', 'priority' ,'responsibilities', 'responsibility', 'requirements', 'required

work on both (i) ESCO and (ii) general taxonomies as well. A detailed description and formalization of how HSS is built can be found in [51].

Step 3: Build embeddings (with reproducible settings). In this step, a range of embedding models (Word2Vec, GloVe, and FastText) were trained on a corpus of approximately six million UK OJAs (2019). In total, 260 models were produced through systematic variation of hyperparameters. The performance of each model was evaluated using the HSS metric. The FastText model with Continuous Bag-of-Words (CBOW) architecture, 300 dimensions, 100 epochs, and a learning rate of 0.1 achieved the highest intrinsic correlation ($\rho = 0.29, p < 0.001$) and was therefore selected for subsequent analyses. More complex transformer-based models (e.g., BERT) did not show a significant improvement for this specific task and were excluded due to computational inefficiency.

4.3.2 Extraction and validation of green skills

As discussed in Section 2.4.1, skill extraction refers to the process of identifying skill-related expressions within unstructured textual data, while skill normalisation aligns these expressions with standard taxonomies such as ESCO. In this section, we apply these principles to the specific domain of green skills, adapting the general extraction and mapping framework to enrich the ESCO taxonomy with new, data-driven content.

Building upon the embedding model previously selected, this phase focuses on extracting, expanding, and validating a lexicon of green-related terms from OJAs to construct a new *green skill pillar* for the ESCO taxonomy. The entire workflow, illustrated in Figure 4.1, is structured into three main stages, each defined by specific inputs and outputs.

Stage 1: Enhancement of the Green Terminology. The process begins with two primary inputs: an initial list of 141 sentinel green terms provided by CEDEFOP and a large corpus of UK job advertisements from 2019. First, the OJAs corpus is filtered to retain only postings relevant to the green economy. Then, using a FastText

embedding model trained on this filtered corpus, we perform a semantic expansion of the initial list. For each of the 141 sentinel terms, the Top-5 most similar words are retrieved via cosine similarity. After curating these candidates to remove duplicates and trivial variations, the lexicon is expanded into the *"Cedefop List Enhanced"*, which contains a more comprehensive set of 227 unique green terms.

Stage 2: Green Skills Extraction and Expert Validation. In the second stage, three key inputs are combined: the enhanced list of 227 green terms, the original OJAs dataset, and the official ESCO Green Taxonomy. The objective is to identify how the 227 abstract terms are described in practice within real-world job postings. We search the OJAs dataset for these terms and extract their surrounding contextual phrases to identify linguistic variants and collocations (e.g., "sustainability" becomes "sustainability agenda"). These extracted *mentions* are then systematically mapped against the existing concepts in the ESCO Green Taxonomy to establish a formal link. This process yields a list of potential new green skills, which undergoes a rigorous validation by Independent Country Experts (ICEs). The experts verify the relevance and quality of each mention, resulting in the final, validated *"Cedefop Green Skills"* database containing 182 distinct items.

Stage 3: Green Occupations Identification. The final stage leverages the validated set of 182 green skills to enrich the overall taxonomy. This curated list serves as the primary input for identifying which occupations are intrinsically linked to green activities. By analyzing the co-occurrence of these validated skills within specific job advertisements, we can identify and flag "Green Occupations." This final step results in an "Enriched Taxonomy" where occupations are augmented with a new, data-driven green dimension, fulfilling the primary objective of the workflow.

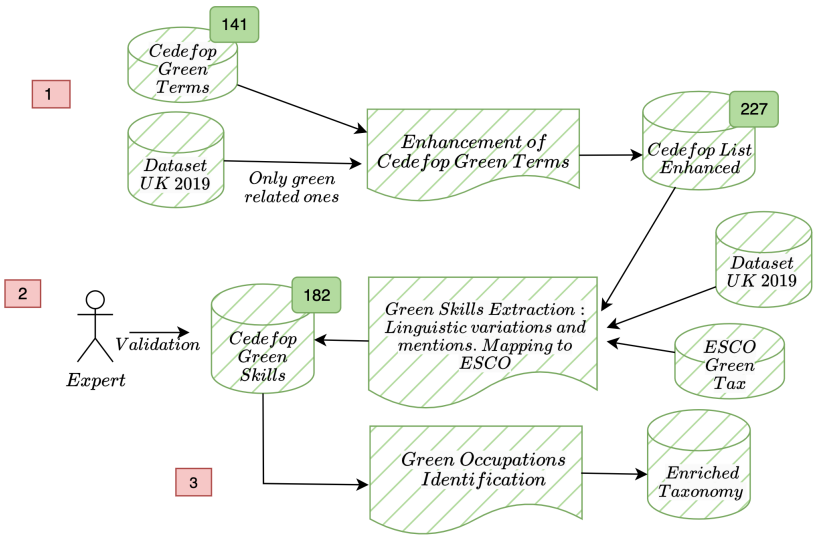


Fig. 4.1 Workflow of the proposed approach to identify and validate green terms.

4.4 Results: Skills and Occupations Pillar

Skill pillar

The skill pillar has the structure shown in Figure 4.2. The first layer contains the green terms identified in OJAs. The second layer contains validated mentions (linguistic variations of green terms). Finally, each green term is associated with the most similar (the cosine similarity measure is employed) ESCO green skills according to the embedding model.

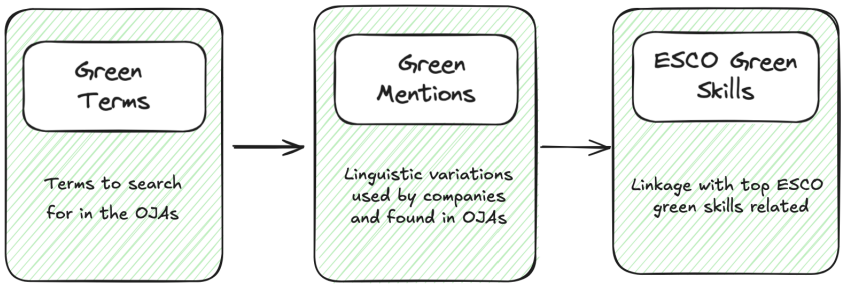


Fig. 4.2 The skill pillar structure, resulted in 182 Green Mentions.

In the following table, there are some examples of green terms, green mentions and ESCO skills associated that are in the skill pillar. For each mention, we also have the information about it being a task or a skill, validated by the ICE.

Table 4.2 Examples of green terms, mentions, and ESCO skill associations.

Green term	Green mention	Top 1 ESCO skill
Air hygiene	Provide air hygiene services	Perform water treatment
Flood risk assessment	Flood risk assessment methods	Environmental engineering
Eco-friendly materials	Use eco-friendly materials	Waste management

Occupation pillar

The **occupation pillar** establishes the link between green skills and ESCO occupations at the IV-digit level through a data-driven approach that draws on both green and non-green skills identified in OJAs. Each ESCO occupation, as classified in the Web Intelligence Hub (WIH), is described using three quantitative indicators calculated on the UK-2019 benchmark dataset:

- **Green pervasiveness**: the share of OJAs for a given occupation that include at least one green skill, relative to all OJAs for that occupation.
- **Greenness**: the proportion of green skills among the total set of skills requested in OJAs for that occupation.
- **Link to the skill pillar**: the specific green mentions observed in OJAsfor the occupation, together with their relative frequency, indicating which skills are most relevant.

It is worth noting that the *green pervasiveness* and *greenness* indicators have since gained wider relevance: they are now employed in Cedefop's online dashboard to monitor the evolution of the green labour market across Europe.

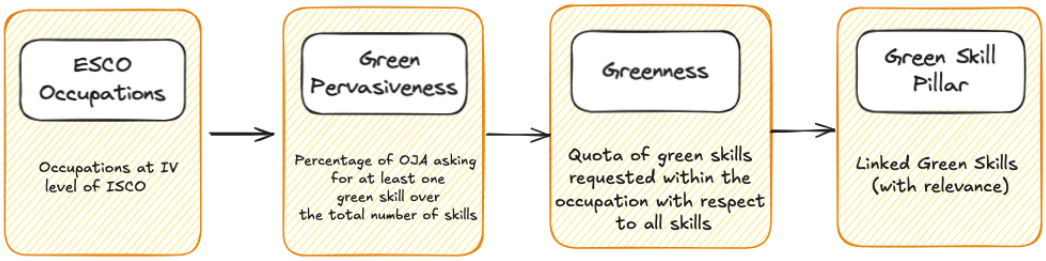


Fig. 4.3 The occupation pillar structure.

For illustration, Table 4.3 reports the case of the *Environmental Engineer* (ESCO 2143). When considering only the CEDEFOP green mentions, around 16% of the vacancies included at least one green skill, and 5.5% of the requested skills were classified as green. Using the ESCO green skills, the same occupation displays higher pervasiveness (36.9%) and slightly higher greenness (6.9%), largely driven by the inclusion of broader, generic terms such as *sustainability* or *waste management* that occur frequently in job ads. Finally, when combining both CEDEFOP and ESCO taxonomies, the indicators increase further (40.0% and 8.4%), and the skill profile integrates both generic and more specific mentions.

This comparison shows how the enrichment process expands ESCO's coverage: the CEDEFOP additions complement the existing taxonomy by capturing emerging and more fine-grained skill mentions, while the ESCO set provides the broader terms. The resulting combined taxonomy offers a richer and more nuanced representation of how green skills are embedded in this occupation.

Table 4.3 Comparison of Environmental Engineer (ESCO 2143) under three taxonomies: Cedefop only, ESCO only, and combined.

ESCO Occupation	Taxonomy	Pervasiveness	Greenness	Top green skills (relevance)
2143 - Environmental Engineer	Cedefop only	16%	5.5%	Geo-environmental engineering (31.3), Safety environmental management (9.7), Geology environmental science (8.8), Flood risk assessment (7.9), Environmental science engineering (6.4)
2143 - Environmental Engineer	ESCO only	36.9%	6.9%	Sustainability (27.0), Geology (20.7), Environmental engineering (20.3), Environmental legislation (20.3), Waste management (19.9)
2143 - Environmental Engineer	Cedefop + ESCO	40.0%	8.4%	Sustainability (24.9), Geology (19.1), Environmental engineering (18.7), Environmental legislation (18.7), Waste management (18.4)

The green occupation pillar relies on the WIH classification of OJAs at the ESCO IV-digit level.[6] Inevitably, some misclassifications can occur in the pipeline – for example, vacancies titled *Senior Ecologist* were on occasion assigned to the category of software developers, leading to misleading matches. However, a substantial effort was devoted during the project to refining and correcting the classification pipeline in order to minimise such errors. As a result, the overall impact of misclassifications is limited, and the consistency of the occupation-level indicators is preserved. Looking forward, additional refinements at the V-digit level could further improve precision.

[6]See the Cedefop Web Intelligence Hub documentation: https://www.cedefop.europa.eu/en/tools/skills-online-vacancies

Chapter 5

Enriching ESCO Digital Skills via LLMs: The TAXMAP Approach

Building on the methodological framework introduced in the previous chapter, which focused on the enrichment of the ESCO *green skills* taxonomy through distributional semantics and expert validation, this chapter presents the second major case study of the thesis: the **TAXMAP** system for the enrichment and monitoring of the *digital skills* taxonomy. While the previous chapter illustrated the use of static word-embedding models to capture and align emerging green terminology, here the focus shifts to the use of LLMs and contextual semantic representations to improve concept discovery, disambiguation, and mapping.

Methodologically, the chapter extends the taxonomy enrichment pipeline from static embeddings (e.g., FastText) to contextual models that leverage transformer architectures, allowing for richer semantic interpretation and better handling of polysemy in digital skill expressions. Substantively, it applies this enhanced framework to the domain of digitalisation, where the fast pace of technological change requires constant updating of the taxonomy to reflect new tools, frameworks, and competencies.

Together, the two chapters represent complementary applications of the same overarching research aim: developing scalable, data-driven methods for the semantic enrichment of skill taxonomies in evolving domains of the European labour market.

Some results of this Chapter were published in [38].

5.1 Introduction and Motivation

Lexical taxonomies provide a natural framework for representing semantic relationships between words and concepts through IS-A hierarchies. They are foundational tools used in various industries and by policymakers to support knowledge dissemination and information retrieval. However, as domains evolve, taxonomies need to be regularly updated to accommodate new categories and relationships, ensuring their continued relevance [113]. This maintenance process, however, is labor-intensive, domain-specific, and expensive, making it a time-consuming task [52]. Consequently, there is a growing need for automated methods to enrich taxonomies efficiently. In this work, we introduce a production system that integrates generative models with expert knowledge to automate the taxonomy enrichment process while ensuring accuracy through human validation. Our method leverages contextual embeddings, which are pre-trained on large-scale unlabeled data and have demonstrated state-of-the-art performance across various natural language processing tasks such as text classification, question answering, and text summarization [41, 135]. This motivated our decision to utilize LLMs as encoders in the system.

This research is conducted within the scope of an EU-funded project titled SkillOvate. The project aims to integrate OJAs into official labor market statistics by creating an AI-driven system that collects and analyzes millions of job ads across Europe[1]. In alignment with this project, our study focuses on developing an automated pipeline to enrich the ESCO[2] digital skills taxonomy with updated information from the Web, as illustrated in Fig. 5.1. Taxonomies like ESCO are crucial in identifying necessary skills for specific jobs and in monitoring changes in the labor market. Policymakers rely on such structured data to make informed decisions about future workforce interventions. However, the traditional top-down approach for developing and maintaining taxonomies, like ESCO, is not only time-consuming but also prone to errors and high costs. This emphasizes the need for a system that can automatically update taxonomies with real-world data, particularly for digital

[1] As of 2019, the project has collected over 300 million unique online job ads from 32 EU countries.
[2] ESCO is the official European multilingual classification of Skills, Competences, and Occupations.

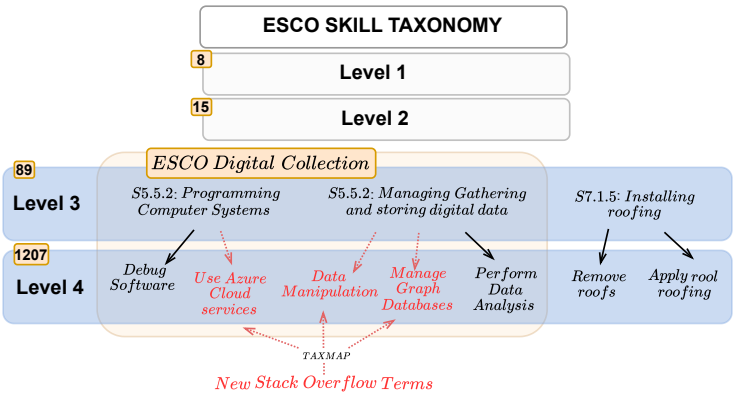


Fig. 5.1 A simplified example of TAXMAP deployment, illustrating the enrichment of the ESCO taxonomy. The numbers represent the count of digital skills within each ESCO level. Our focus is on the last level of the taxonomy.

skills, which are rapidly evolving with the continuous growth of knowledge and the emergence of new concepts.

This work makes three primary contributions:

1. **TAXMAP - TAxonomy eXpansion through Collaborative LLM MAPping**. We propose and formalize a taxonomy enrichment method, TAXMAP, which uses contextual word embeddings as encoders to enrich hierarchical taxonomies with related terms. This method is flexible and can be applied to any taxonomy with a hierarchical structure of concepts.
2. **Real-world deployment**. We apply TAXMAP in the labor market domain to enrich the official ESCO taxonomy by evaluating over 40,000 digital-related terms. This framework is already deployed within an EU research project, with plans for periodic updates to ESCO.
3. **Baseline comparison and human validation**. We construct a baseline using the existing ESCO structure to evaluate the effectiveness of TAXMAP. Additionally, the results are validated by human experts from the European Network of Regional Labor Market Monitoring[3].

[3] http://regionallabourmarketmonitoring.net/

5.1.1 Deployment challenges

A crucial aspect considered during the application of TAXMAP to the ESCO taxonomy is the latter's role as a support tool for European organizations. ESCO serves as a valuable decision-making instrument, both at the national and European levels. In this regard, the quality of enrichment has been prioritized over the quantity of added terms. The primary objective is to expand the taxonomy automatically to ensure constant updates, while simultaneously aiming to produce high-quality enrichment, even at the expense of the quantity of terms added. Therefore, significant attention has been paid to term selection, both through the use of LLMs and subsequent human validation. The adopted approach involved enriching ESCO with a limited number of terms while ensuring their coherence with ESCO's existing structure. Therefore, the main challenges encountered in the development of the project have been: (i) the presence of terms and descriptions that are highly heterogeneous, often ambiguous, not relevant as skills, or incomplete (see Tab. 5.1); (ii) the need to maintain consistency between existing taxonomy structure – constructed following a top-down approach by experts - and new data-driven terms collected from the Web; (iii) the need to validate the finals results provided by algorithms with experts, empowering a Human-AI teaming approach. Our solution exploited LLMs performances and distributional semantics to preprocess, clean, and select the data. This led to a restricted number of quality matches that allowed for human experts' validation, reducing costs and efforts that would have come from a manual enrichment of ESCO Digital Taxonomy and achieving the validation of almost a hundred possible new digital skills.

5.2 Related Work

The methodology proposed in this chapter is situated at the intersection of two established research fields: Taxonomy Enrichment and Labor Market Intelligence (LMI). The former provides the foundational techniques for automatically expanding and maintaining hierarchical knowledge structures, which is the core technical challenge we address. A comprehensive survey of these methods, from early

Table 5.1 Descriptions of Digital Terms from Stackoverflow

Digital Term	Description
Computer vision	Computer vision is an interdisciplinary field that deals with how computers can be made to gain high-level understanding of digital images and videos.
Scala	Scala is a general-purpose programming language providing support for functional programming and a strong static type system.
plot	Wikipedia uses the following definition of plot:
ipad	iPad is a tablet computer designed by Apple running the iOS operating system. iPad applications are usually written in Objective-C and Swift in the Xcode IDE, although it's possible to use other tools to build iPad applications as well.

distributional approaches to modern deep learning models, is provided in the literature review in Chapter 2. Our application is deeply rooted in the domain of Labor Market Intelligence, where taxonomies are pivotal for structuring and analyzing dynamic data from sources like online job advertisements. The broader context of LMI and its analytical challenges are discussed in detail in Chapter 3. This chapter, therefore, synthesizes concepts from both fields to present a novel framework that applies advanced enrichment techniques to solve a critical, real-world LMI problem.

5.3 Methodology

This section details the `TAXMAP` methodology. To ground our approach in established theory, we first revisit the formal definitions of a taxonomy and the Taxonomy Enrichment Problem (TEP). These concepts, introduced in detail in the literature review (Chapter 2), are briefly presented here to ensure the chapter is self-contained and to set the stage for our specific solution. The subsequent sections will then

explore how TAXMAP incorporates new concepts into an existing taxonomy, enhancing its depth and scope.

5.3.1 Setting the stage

Starting from the formalization of [108] and [53], we define taxonomy and the taxonomy enrichment problem.

Definition 5.3.1. A taxonomy $\mathcal{T} = (C, H_C)$ where:

- C is a set of concepts $c \in C$ belonging to the domain of interest (aka, nodes).
- H_C is a directed taxonomic binary relation existing between concepts, that is $H_C \subseteq (c_i, c_j) \in C^2 | i \neq j$.

$H_C(c_1, c_2)$ means that concept c_1 is a generalization of c_2, and c_2 is a specification of c_1. In other words, we refer to c_1 as the *hypernym* of c_2 and c_2 as the *hyponym* of c_1.

Definition 5.3.2. Taxonomy Enrichment Problem (TEP). A Taxonomy Enrichment Problem is a 3-tuple $(\mathcal{T}, \mathcal{D}, \mathcal{S})$, where:

- $\mathcal{T}$ is a taxonomy, as in Definition 5.3.1.
- $\mathcal{D}$ is a set of new terms to be included in the taxonomy with $t \in \mathcal{D}$.
- $\mathcal{S} : C \times \mathcal{D} \rightarrow [0, 1]$ is a scoring function that estimates the relevance of t with respect to an existing concept c in the taxonomy.

A solution to TEP is a new taxonomy $\mathcal{T}^+ = (C^+ = (C \cup \mathcal{D}), H_{C^+})$, where H_{C^+} is the enriched set of hierarchical relations defined as follows: $H_{C^+} \subset H_C \cup \{(c, t) : c \in C, t \in \mathcal{D}\}$. This means that H_{C^+} contains the original relations in H_C along with new directed relations from existing concepts $c \in C$ to new terms $t \in \mathcal{D}$.

As established in prior studies such as [113] and [78], we maintain the assumption that the core taxonomy $\mathcal{T}$ remains unchanged. This approach ensures the preservation of high-level, expert-curated hypernym pairs, which are crucial for maintaining the integrity of the taxonomy's foundational structure. Our focus, therefore, shifts towards the integration of fine-grained terms within the framework of the existing terms in the seed taxonomy. Furthermore, in alignment with the conventions [78], we avoid establishing hierarchical relationships among the newly introduced terms $\mathcal{D}$. This methodology allows us to enrich the taxonomy with additional specificity

and detail without disrupting the established hierarchical order. By concentrating on the attachment of fine-grained terms, we enhance the taxonomy's granularity and utility, ensuring that it continues to reflect expert knowledge while accommodating new insights and terms.

5.3.2 TAXMAP

Employing the term vector encoding, TAXMAP introduces an innovative approach to assign a new concept as the child of a concept already present in the seed taxonomy. Algorithm 1 shows the steps of the methodology, it receives as input the taxonomy $\mathcal{T}$ to be updated, the set of terms $\mathcal{D}$ to be added to $\mathcal{T}$, the scoring function $\mathcal{S}$ and encoder model $\mathcal{E}$:

1. In the encoding step, the embedding model $\mathcal{E}$ is used to obtain a vector for each concept $c_i \in C$ and term $t_j \in \mathcal{D}$.
2. In the scoring step, using the scoring function $\mathcal{S}$, the scoring matrix M is computed. Each cell represents the scoring between the concepts $c_i \in C$ and the term $t_j \in \mathcal{D}$ to be added. Any measure of vector similarity can be used.
3. In step 3 the best match is selected: for each new term, its best parent in the taxonomy is identified as the one whose vector maximizes the score with the vector of the new term. In this way the new taxonomy $\mathcal{T}^+$ is constructed, the set C^+ is the union of the set of concepts already presented in $\mathcal{T}$ with the set of new terms $\mathcal{D}$ and H_C is extended with the best match for each new term.

TAXMAP approach is model-independent, not tied to a specific embedding encoding model or scoring function. Moreover, it facilitates the utilization of multiple vector encoding models to enhance the quality of matches. Given that each model acquires distinct patterns during training and may provide different answers for the same task, we hypothesized that combining multiple models simultaneously could enhance performance. This approach aims to mitigate the likelihood of encountering errors or biases inherent in any individual model. Consequently, we opted for three LLMs and applied algorithm 1 to each, employing cosine similarity as the scoring function. We then extended the TAXMAP algorithm by combining their results. Specifically, we selected the best match from each model and considered only those matches where

Algorithm 1: TAXMAP

Data: $\mathcal{T} = (C, H_c), \mathcal{D}, \mathcal{S}, \mathcal{E}$
Result: $\mathcal{T}^+ = (C^+, H_{C^+})$

/* Step1: Encoding */
$V_C \leftarrow \emptyset$
for $c_i \in C$ **do**
  $V_C \leftarrow V_C \cup \overrightarrow{\mathcal{E}[v_{c_i}]}$
$V_D \leftarrow \emptyset$
for $t_j \in \mathcal{D}$ **do**
  $V_D \leftarrow V_D \cup \overrightarrow{\mathcal{E}[v_{t_j}]}$
/* Step2: Scoring */
$M \leftarrow matrix(m, n)$
for $v_{c_i} \in V_C$ **do**
  for $v_{t_j} \in V_D$ **do**
    $M[i, j] \leftarrow \mathcal{S}(v_c i, v_t j)$
/* Step3: Select best match */
$C^+ = C \cup D$
$H_{C^+} = H_C$
for $t_j \in D$ **do**
  $H_{C^+} \leftarrow H_{C^+} \cup \{(t_j, c_i) | \mathcal{S}(c_i, t_j) = max(M[:, j])\}$
return (C^+, H_{C^+})

the three models *"agreed"* in identifying the same concept c as the best parent for the term t. More details of this choice will be given in the next section.

5.4 Evaluation against a Baseline

5.4.1 Choice of the Models

As mentioned, we employed three pre-trained LLMs for the implementation of TAXMAP. To choose among the different models, we relied on previous work from [89] who constructed a benchmark for Text Embedding models. Their framework (MTEB) aims to clarify how models perform on various embedding tasks and thus provides a comprehensive view of the state of text embedding models. They compare the running time, embedding dimension, and performance (evaluated on different tasks)

of more than 50 models; considering the STS (Semantic Textual Similarity) task, LLMs score higher than word embedding models. The leaderboard with the ranking of the best models is available on the HuggingFace[4] platform and is continuously updated as new models are released. We then selected the top three open source pre-trained models available on June 2024: *mixedbread-ai/mxbai-embed-large-v1*[5] with 335M params, *w601sxs/b1ade-embed*[6] with 335M params and *Labib11/MUG-B-1.6*[7] with 335M params.

5.4.2 Baseline Evaluation

Lacking the required test data to confirm the efficacy of our method in this specific task, we created a baseline to evaluate `TAXMAP` performance in the enrichment of ESCO taxonomy. The ESCO digital collection is hierarchically organized, with each subsequent level providing further specification of the one above it. Thus, to create a baseline, we proposed mimicking the functionality of our framework on the higher level of the taxonomy; specifically, we aimed to enrich the second level of the taxonomy using the third level, one already present in ESCO. Then, we utilized the outcomes of this process for performance evaluation. If `TAXMAP` demonstrates good performance, it indicates its ability to capture the specialization involved in transitioning from a higher level to a lower one and could be applied for our real goal.

Establish correctness criteria. To start with, we needed criteria for correct and incorrect matches in `TAXMAP`. Referring to Def. 2.2.2: given ESCO as the taxonomy $\mathcal{T}$, the match (S6.2, S6.2.1) is considered correct since it belongs to H_C. Therefore, if this match is identified as a possible match by `TAXMAP`, it counts as a correct prediction for the evaluation of `TAXMAP` itself. Similarly, given (S6.2, S5.2.1) as another possible match proposed from `TAXMAP`, it counts as an incorrect match, given that it is not in

[4] https://huggingface.co/spaces/mteb/leaderboard
[5] https://www.mixedbread.ai/
[6] https://huggingface.co/w601sxs/b1ade-embed
[7] https://huggingface.co/Labib11/MUG-B-1.6

H_C. Thus, the ratio between the correct matches and the total number of matches, known as the Positive Predictive Value (PPV), is calculated.

Single Model evaluation First, we perform an individual evaluation of the three models used. For each LLM, we calculate the Positive Predictive Value (PPV). We obtained a PPV rate of 78% for both *MUG-B-1.6* and *b1ade-embed* while *mxbai-embed-large-v1* achieved 79%.

Can models supervise each other? Research in collaborative machine learning has shown that using multiple models concurrently can enhance algorithm performance, particularly in improving confidence in correct output [70, 130]. This approach leverages the idea that models trained differently will have varying strengths and weaknesses, thus allowing them to complement and correct each other. We investigated this idea by examining instances where one model suggested an incorrect match for a particular skill (i.e., identified the wrong parent). If at least one of the other two models provided the correct match, it indicated that the models do not make errors on the same skills. This lack of overlapping errors suggests that the models can correct each other's mistakes. Our study revealed that the three models produced an incorrect match (i.e., none of the models got the right match) only 19% of the time. To verify our intuition regarding the behaviour of LLMs, we analyzed their performances across different top-k matches. Each top-k, for k in all matches in the baseline, is determined by the similarity scores computed by each model. The chart below illustrates the number of accurate matches for each model (on the y-axis) relative to the number (k) of matches assessed (on the x-axis). For instance, the initial segment, up to 50, displays the percentage of correct matches for the top 50 matches per model. The objective is to evaluate the models' performance in both high-confidence matches and those with greater uncertainty, in order to identify any inconsistencies or weaknesses in their individual outputs. Fig. 5.2 shows that the models' performances fluctuate throughout the matches. There is no best model across all partitions of k-top matches; models have been trained in different ways and hence they perform differently on the same task. Initially, they all performed

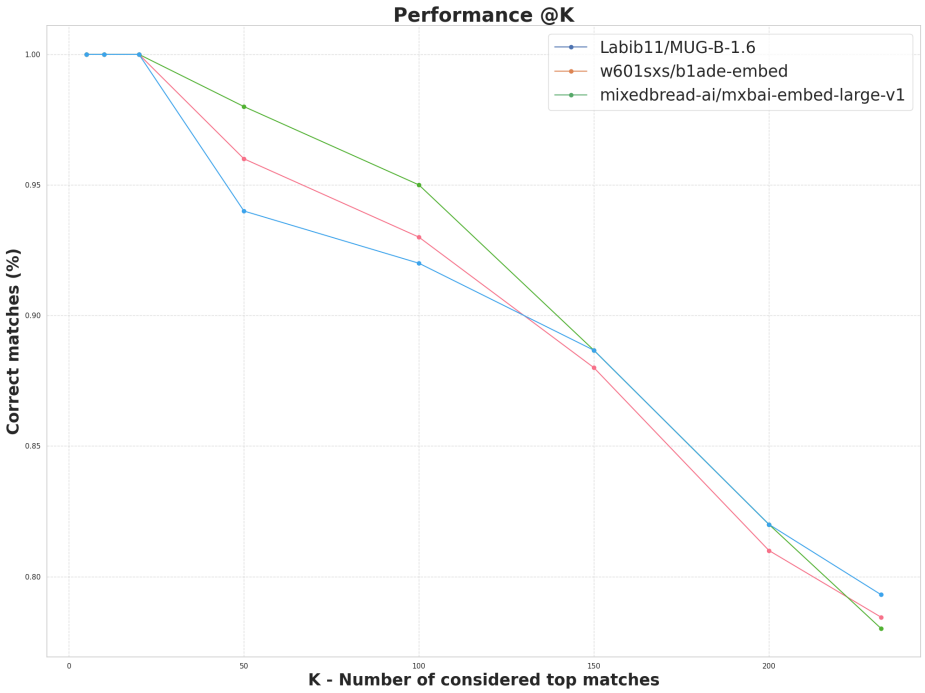


Fig. 5.2 Models performances

well and in the same way, but as we considered a greater number of top matches, their performances deteriorated and became unstable.For instance, LLM *b1ade-embed*, which initially gave the highest number of correct matches in the early k, eventually ended up with the lowest percentages in the last k-partitions. Conversely, LLM *mxbai-embed-large-v1*, which had the poorest performance in the initial partitions, ultimately became the best performer, with the highest number of correct matches. This graph highlights the inconsistency and instability in the performance of each LLM, underscoring the potential risks of relying on any single model in isolation. To solve this issue, we resorted to an intuitive solution: combining the model results. Our intuition was that we could reduce instability and inconsistency by obtaining a consensus among the models. To do so, we retained only the matches where all three models agreed, ensuring the highest consensus among them and possibly the highest confidence. While this approach resulted in fewer overall matches, these matches had a higher confidence of being correct. To further verify the effectiveness of combining the models, we analyzed the performance in cases where all three models agreed on a match. Out of 233 skills, the three models had a common match

in 220 cases. Of these 220 matches, 178 were correct, resulting in a 81% of corrected matches by all three models.

5.5 Experiment

5.5.1 Data Description

This study is situated within the framework of an EU-funded project titled SkillOvate, which seeks to integrate Online Job Ads into official labor market statistics by developing an AI-based system capable of collecting and analyzing millions of Online Job Ads across Europe. Within this project scope, our paper aims to establish an automated pipeline for enhancing the ESCO digital skill taxonomy with current information gathered from the web. Our primary focus lies on the ESCO digital collection, which encompasses skills explicitly identified as digital in the latest version of the ESCO taxonomy (ESCO 1.1.1). This collection is structured hierarchically, as depicted in Figure 1. Our objective is to augment the taxonomy's lower-level nodes, which include 88 specific skills. These nodes, targeted for enrichment, are accompanied by an additional, unstructured level containing more specialized skills (such as defining firewall rules, creating 3D environments, and using CAD for soles). To implement this task, we received a list of 40,561 digital terms mentioned as tags in Stack Overflow Questions and extracted through web scraping in July 2022. These platforms serve as major repositories for programming resources and expert discussions within the computer science and ICT communities. The list encompassed a wide array of highly specific digital keywords and tools, such as *javascript, tensorflow, ubuntu, data encryption, ssh keys*. While this extensive group of terms could offer insights into emerging trends and evolving skill demands in the digital realm, its diversity and specificity presented challenges in curating and refining the dataset to include only pertinent terms for a skill taxonomy. Therefore, to enhance our list and remove irrelevant terms, we focused exclusively on those with a higher-than-average number of mentions on Stack Overflow. This approach yielded 4,215 digital terms suitable for integration into the ESCO skills taxonomy. For the

taxonomy expansion, we specifically targeted the 88 leaves within the ESCO digital collection, as they constitute the last structured level of the taxonomy, and their *child* nodes already encompassed keywords and programming languages. Higher levels of the taxonomy wouldn't be suitable for enrichment through our terms since they contain more general competencies while the nodes we want to attach are more fine-grained (see Fig. 5.1).

5.5.2 Deployment

After achieving optimal results in the baseline evaluation, we utilized the `TAXMAP` to process the terms gathered from the web. We followed the steps mentioned in section 5.3.2 and matched the 4,215 terms with the 88 leaves of ESCO. Fig. 5.3 shows the workflow for the enrichment of ESCO Taxonomy:

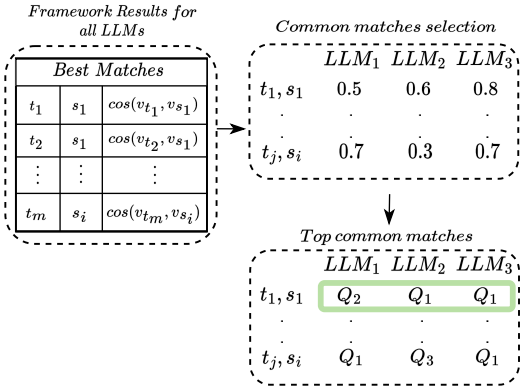


Fig. 5.3 Workflow of `TAXMAP`

1. The first block synthesize the results of `TAXMAP`, where the best matches are selected for each LLM: t_j is the generic digital term, s_i is the ESCO skill and $cos(v_{t_j}, v_{c_i})$ is the cosine similarity used as the scoring function. To increase the quality of the vector representation we also embedded descriptions and examples of use for ESCO skills and digital terms' description.
2. The second block of the figure shows the next step of the matching process. After finding the best match for each model separately, we consider only those matches where all three models identify the same ESCO skill as the best parent (e.g. t_1, s_1). This selection criterion is based on the evaluation results.

3. Finally, we implement a careful approach to refine our parent-child selection on the common matches. To ensure agreement across models in terms of similarity scores, we separately ranked each model scores using quartiles. Only matches with scores in Q1 or Q2 for all models (see the green box in Fig. 5.3) were retained, resulting in a final selection of 924 terms. Here we are assuming that the most similar candidate parent - contained in the top quartiles of the similarity distribution of each model - will be the true hypernym of the word, as [95].

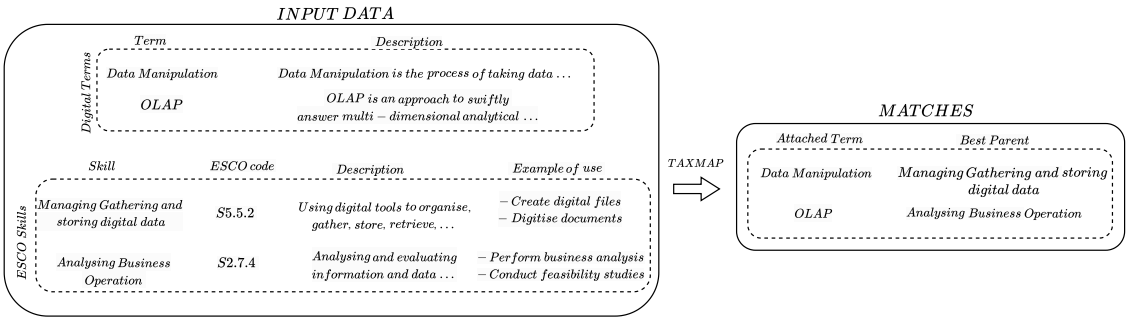


Fig. 5.4 Examples Matches of TAXMAP

The use of complex models such as LLMs usually requires the use of high-performance machines. TAXMAP uses these models as encoders and this allows this methodology to be used even on mid-performance machines, specifically, the framework was implemented in Python[8] and deployed using a machine provided by a cloud service with third-generation AMD EPYC processors, 32 cores and 64 GB of memory. To further optimize the encoding and matching process, we decided to use chroma[9], an open-source embedding database. During the initial encoding phase, three collections were created, one for each LLM, in which the vector representations of both digital terms and the ESCO taxonomy were stored using the specific LLM as the encoder. In the matching phase, the database's query functionalities were utilized to efficiently retrieve the highest similarity matches for the new terms to be added. The overall running time depends on the machine used. In our case, the encoding phase took approximately 15 minutes, while the matching phase took around 100

[8]The source code and the results obtained will be published upon publication.
[9]https://www.trychroma.com/

minutes. The limited execution time allows the use of TAXMAP to continuously update the taxonomies.

5.5.3 Human Experts Validation

To further validate the methodology and its application within the project context, two domain experts were engaged to review TAXMAP.The experts' evaluation addressed two key questions aimed at assessing the accuracy of the matching algorithm and the suitability of the terms for inclusion in a digital taxonomy. The first question (*Q1: Is the match correct in terms of consistency of meaning between the father and the son?*) evaluates whether the relationship between the hypernym term and its hyponym was accurate in terms of semantic consistency (Y) or not (N). The second question (*Q2: The term deserves to be included in a digital taxonomy?*) delved into the relevance of the terms within the ICT field and the ESCO Skill Taxonomy, evaluating whether they represented essential or usable skills or knowledge. This question was crucial as the initial list of terms encompassed a wide range of concepts, some of which may have been relevant to the digital domain but not necessarily suitable for inclusion in a skill taxonomy. For instance, terms like *"gzip"*, *"variables"*, and *"cache"* were excluded as they pertained to the digital realm but did not directly correlate with specific skills or knowledge areas. Both experts independently evaluated the entire set of matched terms: for *Q1* the experts agreed on the evaluation of 86% of matches, for *Q2* the agreement was 81%. A subsequent discussion was conducted for the terms where their evaluations initially diverged, leading to a consensus on the final evaluation

5.6 Results

Fig. 5.4 provides illustrative examples of the matches generated by our system. On the left side of the graph, the input data is displayed, which includes newly identified digital terms along with their descriptions. Additionally, several ESCO skills are shown, complete with their corresponding descriptions and practical usage examples. This input data represents the digital terms identified for potential inclusion in the

Table 5.2 Performance comparison across Q1 and Q2 tasks

Q1: Correct and incorrect matches

Code	ESCO Skill	Tot	Correct	Correct (%)	Incorrect
S1	Communication, collaboration and creativity	147	109	74%	38
S2	Information skills	192	160	83%	32
S3	Assisting and caring	23	22	96%	1
S4	Management skills	74	55	74%	19
S5	Working with computers	414	365	88%	49
S6	Handling and moving	63	39	62%	24
S7	Constructing	1	0	0%	1
S8	Working with machinery and specialised equipment	10	7	70%	3
	Total	**924**	**757**	**81.92%**	**167**

Q2: Detailed breakdown of match correctness*

Code	ESCO Skill	Correct matches			Tot
		Yes	Yes, but misleading	No	
S1	Communication, collaboration and creativity	29 (27%)	9 (8%)	71 (65%)	109
S2	Information skills	40 (25%)	24 (15%)	96 (60%)	160
S3	Assisting and caring	5 (23%)	2 (9%)	15 (68%)	22
S4	Management skills	8 (15%)	9 (16%)	38 (69%)	55
S5	Working with computers	134 (36.7%)	39 (10.7%)	192 (52.6%)	365
S6	Handling and moving	1 (2.6%)	5 (12.8%)	33 (84.6%)	39
S8	Working with machinery and specialised equipment	3 (42%)	1 (14%)	3 (42%)	7
	Total	**220**	**89**	**448**	**757**

* The table presents detailed results for Question 2 of the evaluation, categorizing "Correct Matches" into three groups: "Yes, the term is suitable for inclusion," "Yes, the term is suitable but might be misleading," and "No, the term is not suitable for inclusion in a Digital Skill Taxonomy."

ESCO taxonomy. On the right side, the output from TAXMAP is shown, revealing the top-ranked parent matches identified by the algorithm, which connects the digital terms to the appropriate hierarchical positions within the ESCO taxonomy. To achieve these results, two domain experts were closely involved in the validation process. Their role was to ensure that the matches generated by the system were not only accurate but also aligned with the conceptual structure of the ESCO taxonomy. The validation process was designed to meet two key objectives: (i) to identify and confirm accurate matches between new digital terms and ESCO skills, and (ii) to filter out terms that were unsuitable for inclusion due to misalignment with the taxonomy's purpose or content. Regarding the precision of the matching process (*Q1*), the system evaluated 924 matches, of which 757 were deemed correct by the experts, representing a precision rate of 81.92%. This outcome underscores the effectiveness of the TAXMAP system in accurately aligning digital terms with the ESCO taxonomy (see *Q1* section of Tab. 5.2). By leveraging generative models and expert validation, TAXMAP was able to achieve a high level of precision while significantly reducing the manual effort traditionally required for taxonomy enrichment tasks. The second question (Q2) sought to evaluate if the identified digital terms for inclusion corresponded with the skills represented in the ESCO taxonomy. The intent was to check whether these terms accurately captured digital skills or competencies. Among the 757 correct matches, 220 were digital terms suitable for inclusion, making up nearly 30% of the total correct matches. This finding illustrates that the terms obtained from the scraping process were quite varied, with many not ideally suited for a specialized setting like a skill taxonomy. Some excluded terms were deemed valid for inclusion but could be confusing without context. For example, the term "mask" was linked to ESCO Skill S2.4.2: "entering and transforming information," but listing just the term "mask" in a digital skill taxonomy would not be appropriate. Certain terms, while correctly matched, do not qualify as skills. For instance, terms like "storage," "editor," and "back-up" correspond to S6.2.3: "storing goods and material," S5.6.2: "using word processing, publishing, and presentation software," and S5.5.2: "managing, gathering and storing digital data," respectively. The evaluation results are detailed in the *Q2* section of Tab. 5.2. The terms vary widely, encompassing both specific frameworks

and general language related to skills, abilities, and knowledge. The current ESCO level already features approximately 1,200 terms associated with the digital domain. Nonetheless, TAXMAP has allowed us to add around 200 contemporary terms to this list.

Distribution of new terms on ESCO The new 220 terms enriched 35 taxonomy branches (out of 88). In general, the most enriched branch was the one explicitly related to ICT skill, *working with computers* (S5), to which 134 terms were attached. The majority of terms were linked to data-managing-related concepts: 41 were connected to leaf S5.5.2 *"Managing, gathering and storing digital data"*; 23 to *"using word processing, publishing and presentation software"* (S5.6.2), to 20 to *"using digital tools for processing sound and images"* (S5.6.4), and 15 to S5.5.1 *"browsing, searching and filtering digital data"*. Another consistent part, 29 terms, was linked to programming skills: 13 to S5.2.1, *"Setting up computer systems"*), 8 to *"designing ICT systems or applications"* (S1.11.1), and 8 to *"designing electrical or electronic systems or equipment"* (S1.11.2).

These findings indicate that including these terms has enriched diverse branches of the taxonomy, reflecting homogeneity in the enrichment process, and ensuring that the focus of enrichment wasn't solely directed towards specific third-level entities. Moreover, the fact that certain branches showed more enrichment than others could already yield valuable insights for labor market analysis. Considering that these data were sourced from the web, specifically chosen for their frequent mentions in online questions and queries, it suggests that the degree of enrichment serves as an indicator of how relevant a branch might be becoming in the digital job market context. However, this outcome would require validation using additional data sources on occupations and employer demands.

5.7 Conclusion

The developed system, which synergizes LLMs with expert knowledge, successfully automated and validated taxonomy enrichment for an EU project focused on the ESCO digital skills taxonomy. By employing contextual word embeddings,

a similarity-maximizing function, and a diverse pool of LLMs as encoders, the system effectively linked new web-scraped terms to the taxonomy's leaf nodes. This approach demonstrated robustness, achieving an 81% Positive Predictive Value (PPV) when combining all models, a significant improvement over the baseline. The successful deployment of this framework has substantially reduced the manual effort required for taxonomy enrichment. These results are directly relevant to labor market intelligence; enriching specific branches of the ESCO taxonomy can signal the growing importance of certain digital skills. When used in conjunction with other labor market analysis tools, this methodology can help identify emerging skills and trends across the EU, ensuring the taxonomy remains aligned with current labor market demands.

Chapter 6

SkiLLens: A Multilingual Pipeline for Emerging Skill Detection and Mapping

Introduction

The preceding chapters have established the critical need for dynamic labor market intelligence and have explored methodologies for enriching static skill taxonomies like ESCO. While frameworks such as `TAXMAP` demonstrate the power of combining LLMs with expert validation for a single language context, a significant challenge remains: scaling these enrichment efforts across the diverse linguistic landscape of the European labor market. The ability to monitor skill trends in near real-time, across borders and in multiple languages, is essential for effective policymaking.

This chapter directly confronts this challenge by introducing **`SkiLLens`**, a multilingual pipeline designed to detect, validate, and normalise emerging skills from millions of OJAs. Developed within the Pathways to Inclusive Labour Markets (PILLARS) project, `SkiLLens` implements the core concepts of this thesis on a pan-European scale. It has been applied to a large-scale dataset of over 18 million job ads from 28 countries, covering 22 distinct languages. The methodology presented here advances our work in two critical dimensions. First, it implements a robust, three-step process that uses word embeddings to identify novel skill candidates before leveraging LLMs for refinement and mapping to ESCO. Second, it incorporates a large-scale,

qualitative validation exercise involving national labor market experts from across Europe, ensuring the real-world relevance and accuracy of the identified skills. This chapter details the architecture of the `SkiLLens` pipeline, presents a comprehensive quantitative and qualitative evaluation of its performance, and demonstrates its capacity to serve as a powerful tool for cross-border labor market analysis. The work presented in this chapter has been accepted to the Industry Track of the 2026 EACL Conference.

6.1 Building SkiLLens

Building on the concepts of *Skill Extraction* and *Skill Normalization* introduced in Chapter 2, this section presents `SkiLLens`, a multilingual, multi-stage framework designed to identify, map, and evaluate novel skill expressions emerging from OJAs. As shown in Figure 6.1, the process follows a three-stage pipeline. Each stage is described in detail below, outlining its input, output, and contribution to the overall workflow.

To the best of our knowledge, no prior study has conducted skill extraction over a comparably diverse set of languages or systematically evaluated the novelty of the extracted skills. To address these gaps, we employ a rigorous expert validation process involving 28 labour market specialists, each operating within their national context. This ensures that the extracted skills are both contextually relevant and genuinely novel.

Step 1: Extraction of Candidate Skills

This initial phase of the pipeline focuses on identifying candidate new skills from a large, multilingual corpus of OJAs. The workflow begins by establishing the data foundation and then applies word embedding techniques to extract semantically relevant terms.

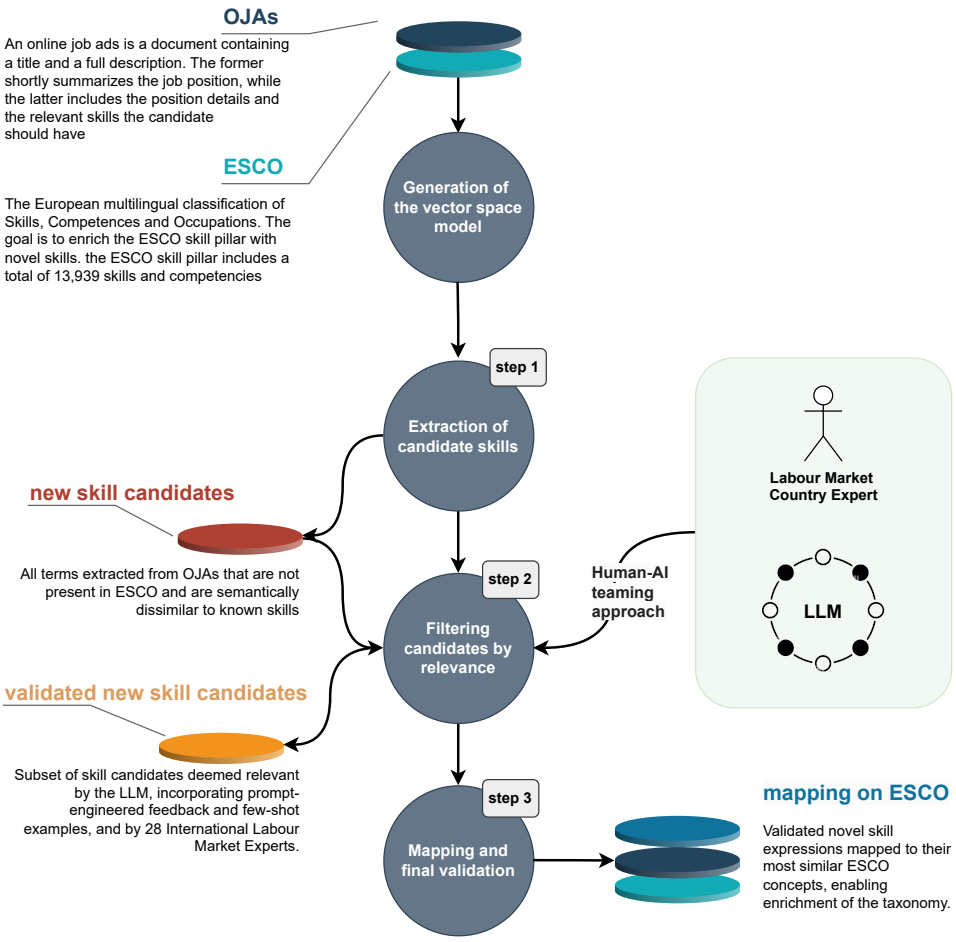


Fig. 6.1 Overview of the SkiLLens framework for extracting and mapping novel skill expressions from OJAs across 28 countries.

Data Corpus The OJAs used in this study were collected through the WIH[1], a core part of Eurostat's Trusted Smart Statistics (TSS) initiative. From the full WIH-OJA dataset of over 450 million ads, we use a representative sample (*NLP sample v3*, release r20240226) spanning 2018–2023. For this study, we selected the most recent one million ads per country where available, resulting in a final dataset of approximately

[1]https://cros.ec.europa.eu/wih

18 million ads focused on recent labor market dynamics[2]. Table 6.3 provides a breakdown of the OJA counts for the 27+1 countries included.

Table 6.1 Total OJA counts by country (capped at 1M)

Country														
Total OJAs	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	792,032	888,423
Country														
Total OJAs	651,722	371,652	496,823	631,463	665,033	317,259	285,274	289,078	174,351	198,966	139,813	120,007	113,436	59,473

Emerging Skills Extraction via Word Embeddings With this dataset as our foundation, we proceed with the extraction process. The input for this step is the raw text from the OJA descriptions, and the output is a list of candidate novel skills.
Input: Raw OJA texts.

an extracted part from an OJA

We're seeking a detail-oriented Project Manager/Data Scientist with strong quantitative skills and a knack for uncovering insights from complex datasets. You'll play a pivotal role in analysing large volumes of data using Python, R, or Java, alongside a deep understanding of machine learning algorithms. Your expertise in programming, statistical analysis, and data visualisation tools will be crucial in communicating findings effectively to diverse stakeholders. The role also requires interaction with security personnel to ensure data compliance and organizational safety. If you're passionate about transforming data into actionable insights, contributing to strategic decision-making, and exploring prompt engineering for AI-driven workflows, we'd love to have you join our team

Output: List of candidates novel skills.

[2]We access the WIH-OJA *NLP* dataset through our participation in the "*PILLARS*" project consortium. The data can also be requested from EUROSTAT, subject to approval and a non-disclosure agreement (NDA)

Candidate Novel Skills = {attention to detail, mathematical aptitude, quantitative reasoning and problem-solving skills, expertise in statistical analysis, Python, R, Java, machine learning, SQL, data visualisation, project manager, interaction with security personnel, programming, prompt engineering}

The core idea is to train country-specific word embeddings and use the existing ESCO skills as "seeds" to discover new, semantically similar terms. For each known ESCO skill s, we identify the top n most similar terms in the embedding space. These terms, which are close in meaning but lexically distinct, form our initial pool of candidate new skills.

Implementation Details To achieve this, we first apply standard preprocessing to the OJA texts, including lowercasing, stopword removal, and n-gram creation. We then train country-specific word embeddings using FastText [14], as it showed consistent performance across countries in previous evaluations [52, 38]. To select the optimal model for each country, we perform an extensive grid search over 160 configurations, varying parameters such as the algorithm (Skip-gram (SG) vs. CBOW), embedding size, and learning rate. Each configuration is evaluated using the Hierarchical Semantic Similarity (HSS) metric [54], which measures how well the ESCO semantic structure is preserved. The best-performing setup was consistently found to be: model=SG, dim=100, epoch=5, lr=0.01. Using this optimal model, we treat the existing ESCO skills as "seeds" to discover new terms. For each of the 14,000 ESCO skills, we retrieve the top-5 most similar terms based on cosine similarity. This threshold was selected empirically to balance relevance and duplication. To ensure the extracted candidates are genuinely novel, a final filtering step is applied. We discard any skill candidate that is syntactically too similar to an ESCO skill. We apply a filtering step using the `fuzz.ratio` function from the `rapidfuzz` library[3], keeping only expressions with a similarity score ≤ 70. The ratio metric is based on

[3]https://rapidfuzz.github.io/RapidFuzz/

normalized Indel similarity[4] This constraint ensures that novel skills remain lexically distinct from existing taxonomy entries.

Step 2: Filtering candidates by relevance

Input: List of candidate novel skills.

Candidate Novel Skills = {attention to detail, mathematical aptitude, quantitative reasoning and problem-solving skills, expertise in statistical analysis, Python, R, Java, machine learning, SQL, data visualisation, project manager, interaction with security personnel, programming, prompt engineering}

Output: List of relevant candidate skills.

Relevant Candidate Skills = {project manager, interaction with security personnel, statistical analysis, programming, prompt engineering}

In this step, we utilise ESCO skills as anchors to discover additional, potentially relevant skills. We employ word embeddings to retrieve terms that are semantically close to ESCO skills and are also sufficiently common in OJAs. To combine these two criteria we score each candidate c for a given ESCO skill s as

$$S(c,s) \;=\; \alpha\, c_{\text{sim}}(c,s) \;+\; (1-\alpha)\, f_{\text{zipf}}(c), \tag{6.1}$$

where $\alpha \in [0,1]$ controls the trade–off between semantic similarity and corpus usage (selected on a held–out set; following [52] formulation α is set at 0.85). The similarity term is

$$c_{\text{sim}}(c,s) \;=\; \frac{\text{cos\_sim}(c,s)+1}{2} \in [0,1], \tag{6.2}$$

Following Zipf's law of word frequencies [145], widely documented in IR/NLP [77, 102], we use a rank-normalised frequency signal:

[4]Given two words w_1 and w_2, the Indel similarity is computed as $Indel = 1 - (LevenshteinDistance_{w_1,w_2}/(len(w_1)+len(w_1))$

and the Zipf-based frequency signal is

$$f_{\text{zipf}}(c) \;=\; 1 - \frac{\log r(c)}{\log V} \in [0,1], \tag{6.3}$$

with $r(c)$ the rank of c by corpus frequency (1 = most frequent) and V the vocabulary size.

In practice, we (i) compute (6.1) only for the top-k nearest neighbours of s in the embedding space, and (ii) filter out rarely used terms by retaining only those candidates whose $f_{\text{zipf}}(c)$ contribute to the top 95% of the cumulative f_{zipf} mass within this top-k set. The remaining terms form the pool of candidate new skills.

Candidates Validation

Input: List of relevant candidate novel skills.

Relevant Candidate Skills = {project manager, interaction with security personnel, statistical analysis, programming, prompt engineering}

Output: Validated list of candidate novel skills.

We perform a two-part validation process to ensure the quality and relevance of the extracted novel skill candidates before proceeding to their mapping onto ESCO.

Automated Validation via LLM To assess the plausibility of new skill candidates, we first employ an LLM validation step based on LLaMA 3.1 8B[5]. When using LLMs for classification-like tasks, carefully designed prompts can significantly enhance the consistency and reliability of the responses. In this stage, we apply the following strategies:

- **Contextual framing:** We provide rich contextual information within the prompt to improve the model's comprehension. This includes:

[5]https://ollama.com/library/llama3.1:8b

  - A definition of the concept of *skill*, aligned with the one used in ESCO and the broader literature.
  - Instructions on the expected format of the response, explicitly asking the model to begin with "yes" or "no" to ensure machine-readable outputs.
  - A representative example of an OJA that contains the candidate skill term in a natural context, helping the model assess whether the term functions as a skill in real usage.
- **Few-shot learning:** We prepend the actual query with a curated set of example prompts and their corresponding responses. Each example includes five candidate terms along with their usage in OJAs. The model is asked to determine whether each term constitutes a skill, beginning its answer with "yes" or "no" and justifying the choice with a short explanation. This calibration helps guide the model toward more accurate and coherent responses.

This LLM-based validation serves as a crucial filtering step before expert evaluation. The use of LLMs was necessary due to the large volume of candidate terms—approximately 5,000 per country—which made manual expert validation of all entries impractical. To address this, we designed the prompt to be deliberately conservative: in cases of uncertainty, the model was instructed to flag terms as potentially relevant rather than discard them. This conservative behavior is reflected in the overall expert-validated precision of around 70%, indicating that we prioritized high recall over precision to ensure that relevant terms were retained. Although a full quantitative evaluation of LLM filtering is out of scope, we conducted random checks on filtered-out terms for two markets (Italy and the UK), and these showed that the recall of valid skills was above 98%, confirming that very few relevant candidates were erroneously excluded.

Manual Validation by Experts In the second part, the filtered list of candidates is reviewed by Labor Market Experts, who are specialised in the labor market of their respective countries. Their role is to validate the plausibility and contextual relevance of each candidate skill, ensuring that the expressions reflect meaningful

and novel concepts within the national context. Table 6.2 shows an example of the output from this two-step validation process.

Table 6.2 An example of candidate novel skill validation.

Extracted Skill	LLM Validation	LLM Motivation	Experts Validation	Experts Motivation
Project manager	✗	No, it is an occupation	*null*	*null*
Interaction with security personnel	✓	Yes, because it requires effective communication	✓	this is 'experience' not skills
Statistical analysis	✓	Yes, it is a core methodological competence	✓	
Programming	✓	Yes, it is a specialized skill.	✗	too broad... programming activities, computer programming, etc.
Prompt engineering	✓	Yes, it is a learned competence in AI.	✓	

Step 3: Mapping and Final Validation

This phase aims to map the extracted new skill expressions to their most appropriate position within the ESCO taxonomy, ensuring their standardisation and enabling the potential enrichment of the taxonomy with relevant, emerging skills.

Sentence Embedding Mapping to ESCO via LMs

Input: Validated list of candidate novel skills.

Output: List of n recommendations of mapping for each candidate novel skill.

We apply semantic similarity analysis between new skill candidates and ESCO entries using sentence embeddings. Multiple models, including Sentence-BERT and multilingual transformer encoders, were explored. Each novel skill is paired with its top 3 most similar ESCO skills based on cosine similarity. To construct a multilingual benchmark, we extract all ESCO skill labels (13,939 preferred terms) and their associated alternative labels for all 22 languages represented in the new skills

corpus[6]. The task is to correctly match each alternative label to its corresponding preferred label. Since the new skills do not yet have an established position in ESCO, this benchmark serves as a proxy to evaluate the ability of our method to correctly place out-of-taxonomy terms in their most semantically appropriate location within the existing skill hierarchy.

Implementation Given the cross-lingual nature of our corpus, we tested both multilingual sentence embeddings and an alternative approach where all texts were translated into English and then encoded. The latter proved more effective. For translation, we used the DeepSeek v3. Embedding models to map candidate novel skills to ESCO were selected based on the MTEB (Massive Text Embedding Benchmark) leaderboard [89], which compares over 50 models across tasks such as Semantic Textual Similarity and also considers encoding time and dimensionality. We excluded commercial or restricted-weight models and added three widely adopted open alternatives commonly integrated in NLP libraries and APIs[7]. Overall, we evaluated a diverse set of open-weight models covering a range of dimensions and architectures. To assess robustness, we computed similarity using thirteen different distance metrics, including cosine, dot product, Euclidean, and correlation-based variants. This allowed us to identify the most reliable embedding–metric combination empirically. Specifically, we used an English dataset derived from ESCO to select both the best-performing embedding model and the most effective similarity metric. Once these were fixed, we applied them to two English benchmark datasets to compare `SkiLLens` against the state of the art (Sec. 6.2), and subsequently analyzed the multilingual performance on the ESCO-based dataset (Sec. 6.2). We conducted evaluations considering Precision@1, Precision@3, Precision@10 and Mean Reciprocal Ranking (MRR), following previous literature [40, 30]. For the precision@1 setting, we refined the top-3 suggestions provided by sentence embeddings through an LLM-based re-ranking step described in the following section.

[6]In ESCO, alternative labels can include synonyms (words with similar or identical meanings), spelling variants, declensions, and abbreviations. These labels help connect ESCO concepts to the real labor market by providing alternative ways of referring to them.

[7]all-MiniLM-L6-v2, all-mpnet-base-v2, and gte-large-1.5b

LLM-Based Selection of Best Match

Input: List of n recommendations of mapping for each candidate novel skill.

Output: Final mapping of new skills into ESCO

For each alternative label, we retrieve the top-3 most similar ESCO preferred terms based on cosine similarity in the sentence embedding space. These candidates are then passed on to an LLM which selects the best match using a prompt that includes instructions, few-shot examples, and a brief explanation of the task. This hybrid approach leverages the semantic power of embeddings with the reasoning capabilities of LLMs, and results in improved disambiguation and classification accuracy. The listing 6.1 shows the prompt template used for each candidate.

```
messages = [
    {
        "role": "user",
        "content": (
            "I need to select the correct match for the following skill: " + candidate +
            " The correct match should be the most similar term. Now I will give you three possible terms and their descriptions: " +
            "1. Term: " + alternatives[0] + " Description: " + descriptions[0] +
            " 2. Term: " + alternatives[1] + " Description: " + descriptions[1] +
            " 3. Term: " + alternatives[2] + " Description: " + descriptions[2] +
            " Tell me which is the most similar term to the candidate. Give me only the term without comments. " +
            "Always give me an answer among the three. If you think none match, try to guess anyway. " +
            "Never provide an answer outside the given options."
        )
    }
]
```

Code 6.1 Prompt Template

Implementation For each alternative label, we retrieve the top-3 most similar ESCO preferred terms based on cosine similarity in the sentence embedding space. To chose the best alternative, we evaluated three language models on the english dataset, both open weight and not: GPT4, Gemini 2.0 flash, and Mixtral-8x22B. The best performances are reached by GPT4, which we chose for the final step. Tab. 6.5 presents the empirical results for the best match selection.

Step 3 is performed both on two benchmark dataset in english to compare `SkiLLens` against the state of the art (Sec.6.2) and on a multilingual baseline (Sec.6.2) based on ESCO to assess its performances in a multilingual setting.

6.2 Experimental Results

In this section, we present the results of our experiments, structured into three main evaluations: (i) expert validation of extracted skills, (ii) evaluation against the state of the art, and (iii) benchmarking against ESCO's preferred labels. Each evaluation provides insights into the effectiveness of our approach in identifying and categorizing skills within the ESCO taxonomy.

Expert Evaluation of Extracted Skills To validate the quality of the skills extracted from job advertisements, we conducted a human evaluation involving labor market experts. These experts were recruited through the European Network of Regional Labour Market Monitoring[8], an international community of labour market experts committed to developing, disseminating, and applying innovative concepts, methods, and tools tailored to labour market analysis. The primary objective of this evaluation was to assess whether the skills identified by our model were relevant and correctly formulated in the context of real-world job descriptions. We involved experts from the 27+1 EU countries in the evaluation of the proposed candidates for each country. The results reveal that: (i) among the 4,941 candidates proposed as *novel skills*, 3,552 (71.9%) were recognized as skills by the experts; (ii) this value varies considerably among countries, as shown in Fig. 6.2. It is worth noticing that the latter reflects expert evaluations conducted on the original job ads in their respective languages, not on translated versions. Therefore, the performance differences across countries stem from a combination of factors: the subjective choices of local experts, the varying quality of language-specific embeddings, and differences in the number and nature of OJAs available in each language; (iii) it is also important to note that

[8]https://www.regionallabourmarketmonitoring.net

DE, CY, and NL did not consider the ESCO knowledge concepts[9] in the validation, affecting their overall scores.

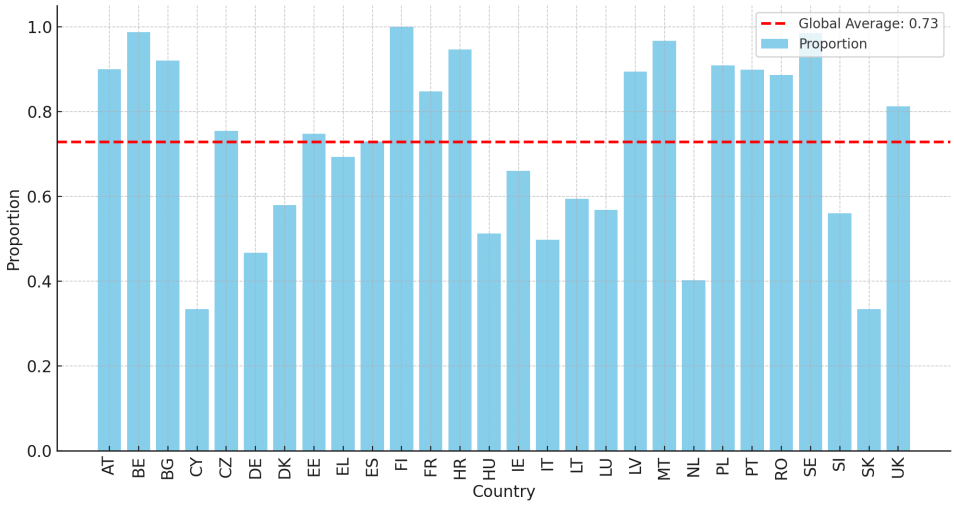


Fig. 6.2 Proportion of Candidates skills evaluated as novel and valid as skills

Tab. 6.3 presents the numerosity of OJAs validated by the country experts for each country.

Table 6.3 Number of new skill candidates evaluated by each expert per country.

Country														
Num. candidates	400	395	350	316	292	287	279	242	241	229	224	215	209	207
Country														
Num. candidates	191	163	145	142	130	107	90	89	76	62	45	33	30	19

Mapping of New Skills into ESCO The second evaluation step assessed the ability of our model to map extracted skills into the ESCO taxonomy. This evaluation was conducted in two ways.

Comparison Against State-of-the-Art

We compared our approach with the methods in [40] and [30], both of which focus on the English language, the only language analyzed in their studies. Specifically,

[9]https://esco.ec.europa.eu/en/classification/skill_main

we replicated their task of mapping soft skills to the ESCO taxonomy, rather than classifying skills in job posts, which falls outside the scope of this paper, as our focus is solely on novel skills. Table 6.4 presents the comparison results for English. The authors evaluated their approaches on two datasets: 'house' and 'tech.' Our method outperforms theirs in terms of RP@1, RP@5, RP@10, and MRR—the key metrics used in their studies.

Table 6.4 Comparison of Skill Mapping Performance on Tech and House datasets

Method	RP@1	RP@5	RP@10	MRR
Tech				
Decorte et al. [40]	n/a	0.3171	0.3919	0.339
Clavié and Soulié [30]	0.465	0.615	0.689	0.537
SkiLLens (ours)	**0.8450**	**0.898**	**0.928**	**0.823**
House				
Decorte et al.[40]	n/a	0.308	0.387	0.299
Clavié and Soulié [30]	0.630	0.567	0.610	0.507
SkiLLens (ours)	**0.814**	**0.924**	**0.937**	**0.845**

Baseline Evaluation

We constructed a baseline to evaluate SkiLLens' performance in enriching the ESCO taxonomy. Our focus is on the fourth level, the most granular one. Since our objective is to identify sibling relationships within the taxonomy—allowing us to map newly extracted skills—we concentrate on matching *alternative labels* to their corresponding *preferred labels*[10]. Moreover, we focused on alternative labels to maintain the integrity of the ESCO hierarchy, ensuring consistency with its ontology while improving coverage with real-world terms. If our framework performs well in this task—where we have a ground truth for evaluation—it demonstrates its ability to capture semantic similarity among skills. This ensures its suitability for enhancing ESCO by refining alternative label mappings. To evaluate this, we sampled 200 alternative labels from the fourth level of the ESCO skills taxonomy, translated them into English using Deepseek v3, and generated their vector representations. We then applied cosine similarity to retrieve the top three preferred label candidates

[10]In ESCO, alternative labels can include synonyms (words with similar or identical meanings), spelling variants, declensions, and abbreviations. These labels help connect ESCO concepts to the real labor market by providing alternative ways of referring to them.

for each alternative label. For encoding and matching, we used ChromaDB[11], an open-source vector database optimized for efficient storage and retrieval of vector embeddings. During the initial encoding phase, we created separate collections for each language taxonomy, storing the vector representations of both alternative and preferred labels using the selected language model (LM). In the matching phase, we leveraged the database's query functionalities to efficiently retrieve the top three most similar preferred labels for each alternative label based on their embeddings. This approach significantly improved the efficiency and scalability of our matching pipeline. We tested the top ten models on STS from the MTEB leaderboard[12] and found that *thenlper/gte-large* with 335M params achieved the best performance. Finally, to refine the selection and determine the best unique match, we designed a structured prompt. We submitted it to ChatGPT-4, which selected the most appropriate match from the top three candidates. After obtaining the final match from the model, we compared it against the ESCO ground truth to compute the accuracy score. This pipeline was implemented across all 22 languages in our framework. The evaluation results—including Accuracy@1, Accuracy@3, NDCG@3, and refinement accuracy—are presented in Tab. 6.5. The table highlights the strong performance of `SkiLLens` across all languages considered.

[11]https://www.trychroma.com/

[12]Available at https://huggingface.co/spaces/mteb/leaderboard and is continuously updated as new models are released.

Table 6.5 Performance Metrics (%) for each language

Lang	Countries	Retrieval			Refinement
		Acc@3	NDCG (@3)	Acc@1	LLM Accuracy
bg		78.50	77.46	68.00	**68.84**
cs		87.62	86.70	77.23	**77.39**
da		80.00	79.29	64.50	**68.84**
de		82.50	81.58	**73.50**	**73.50**
el		77.23	75.22	**63.37**	60.10
en		89.00	87.15	83.50	**84.50**
es		84.58	83.39	72.14	**73.00**
et		59.90	57.92	46.53	**50.00**
fi		78.50	77.09	64.00	**65.83**
fr		84.08	82.52	71.14	**73.00**
hr		90.50	88.67	78.00	**80.71**
hu		77.00	76.51	62.50	**71.21**
it		80.00	78.07	65.50	**70.71**
lt		75.74	75.76	61.39	**63.00**
lv		81.00	79.89	70.00	**72.50**
nl		84.00	82.25	70.00	**70.35**
pl		77.61	75.65	65.67	**67.00**
pt		85.50	84.74	72.00	**73.74**
ro		62.69	60.79	46.77	**54.50**
sk		74.00	70.90	**57.50**	56.78
sl		84.50	82.94	72.50	**73.37**
sv		82.00	81.82	73.00	**74.00**

Note: The flags indicate countries in which an OJA in that language has been found.

6.3 Conclusion

In this paper, we introduced `SkiLLens`, a multilingual pipeline for extracting and mapping novel skills from over 18 million job ads across 27+1 European countries. Our two-step methodology combines embedding-based extraction with LLM-assisted validation and mapping into the ESCO taxonomy. The results demonstrate strong alignment with expert assessments, with over 70% of extracted skills recognized as valid by national labor market experts. Regarding mapping, our method outperforms existing baselines in all the metrics. However, performance varies across languages, highlighting areas for improvement in low-resource or morphologically rich contexts.

Future work will focus on integrating feedback loops for taxonomy enrichment and improving cross-lingual alignment through domain-adaptive fine-tuning and prompt optimization.

The work of `SkiLLens` is the last step of the methodological contributions of this thesis. Having established effective pipelines for enriching both green (Chapter 4) and digital skills (Chapter 5), and now a scalable multilingual system, the final part of this dissertation will shift from methodology to broader impact and future directions.

Chapter 7

Ethical Considerations in the Use of Online Job Advertisements and LLMs

The increasing availability of OJAs and recent advances in NLP, including LLMs, have enabled novel data-driven approaches to labour market analysis and skill extraction. These methods offer significant analytical advantages in terms of scale, timeliness, and granularity. At the same time, their use raises a number of ethical considerations that are particularly relevant when analytical outputs are intended to inform labour market intelligence, education planning, and public policy.

A first set of ethical concerns relates to the use of OJAs as a data source. Although job advertisements are publicly accessible, they do not constitute a neutral or exhaustive representation of labour market demand. OJAs primarily reflect employer-side preferences and tend to over-represent large firms, high-skilled occupations, and digitally mediated recruitment channels, while under-representing informal work, small enterprises, and segments of the labour market where hiring occurs through non-digital or institutional pathways. As a result, analyses based on OJAs may systematically exclude certain occupations or worker groups and risk reinforcing existing inequalities in labour market visibility when used for policy-relevant purposes. The implications of these data-related limitations, as well as the strategies adopted in this thesis to mitigate them—such as focusing on within-occupation analyses—are discussed in detail in Section 8.6.1.

The second and complementary ethical issue concerns the use of LLMs and embedding-based models for skill extraction, classification, and semantic mapping, which is more specifically related to the topic of the work presented in this thesis. While such models are effective at capturing linguistic regularities at scale, a substantial body of literature has shown that they tend to inherit and amplify biases present in their training data, potentially reinforcing stereotypes related to gender, occupation, and other sensitive attributes [110, 60, 49, 81]. Empirical studies comparing LLM outputs with authoritative labour market statistics, such as data from the U.S. National Bureau of Labor Statistics [55], reveal that these biases can manifest even in out-of-the-box model deployments, with important implications for recruitment-related applications. In the context of skill analysis, biased or over-generalised model outputs may lead to inaccurate mappings between skill mentions and standardised taxonomies, or to an over-emphasis of generic and socially desirable skills at the expense of more task-specific or less salient ones [11, 55].

Recent work on LLM-based skill classification further emphasises that fine-grained taxonomies such as ESCO, while essential to preserve analytical fidelity and avoid premature aggregation, introduce substantial complexity for downstream tasks. Highly granular skill spaces can increase interpretative uncertainty and place a greater burden on automated systems, making careful aggregation strategies, error analyses, and human validation necessary to avoid distorted conclusions [128]. These challenges are compounded by the limited interpretability of LLMs, which are often characterised as black-box models due to their scale and architectural complexity. Limited explainability hampers error diagnosis, bias detection, and trust, particularly in high-stakes or policy-relevant applications where justifiable decision-making is required [134].

These issues are further amplified in multilingual settings, where models rely on shared semantic representations trained on corpora with uneven linguistic and cultural coverage. As a consequence, biases and dominant associations from high-resource languages often influence behavior in lower-resource ones, leading to cross-lingual bias propagation. This phenomenon is formalized by Měchura [80], who introduces a taxonomy of bias-causing ambiguities in machine translation. This

taxonomy categorizes how source-language underspecification—such as gender-neutral terms (e.g., "doctor"), number ambiguities (singular vs. plural "you"), or formality levels—forces models to make biased "guesses" when the target language requires those properties to be explicit. Such structural mismatches provide a concrete mechanism for how stereotype leakage occurs at the interface of different linguistic systems. While his work focuses on translation, these structural mismatches are equally problematic for skill extraction, where a model must 'decide' on a gendered or formal attribute for a skill term that was neutral in the source text. Moreover, performance may degrade in languages or domains that are poorly covered by the underlying taxonomy or encoder. This raises ethical concerns for cross-lingual skill extraction and classification, as semantic equivalence across languages cannot be assumed to be culturally neutral, and errors or biases may disproportionately affect under-represented linguistic communities [128, 92].

A further concern arises from the downstream use of AI-derived labour market indicators. International organisations have cautioned that AI-based analyses of skill demand, if not properly contextualised, may encourage over-confidence in automated signals and obscure the uncertainty inherent in model-based inferences[1]. This risk is exacerbated by the phenomenon of model hallucination, where LLMs may confidently identify skills or qualifications that do not exist within the underlying OJAs. If left uncorrected, these fabricated insights can distort the perceived reality of the labour market, potentially leading to misinformed education planning or reskilling initiatives [62]. This is particularly relevant in policy environments, where analytical outputs may influence education systems, reskilling initiatives, or funding allocations. Ethical responsibility therefore extends beyond model development to the communication and interpretation of results, including the transparent reporting of assumptions, limitations, and potential sources of error.

In light of these considerations, this thesis adopts several mitigation strategies aligned with emerging best practices for responsible AI use. The methodological chapters explicitly acknowledge the partial and selective nature of OJA data and

[1] https://www.oecd.org/content/dam/oecd/en/publications/reports/2021/03/demand-for-ai-skills-in-jobs_679000d8/3ed32d94-en.pdf

avoid claims of population-level representativeness. In addition, human-in-the-loop validation is employed, particularly in distinguishing between skill mentions, skill entities, and standardised skill concepts, and in multilingual contexts where automated mappings are more error-prone. Finally, transparency and reproducibility are prioritised through detailed documentation of pipeline design choices and by framing analytical outputs as decision-support signals rather than definitive measurements.

Overall, while the combined use of OJAs and LLM-based pipelines enables valuable insights into skill demand dynamics, their ethical implications must be carefully considered. Responsible use requires critical awareness of data limitations, model biases, and potential downstream impacts, particularly when analyses intersect with labour market governance and public policy.

Part III

Impact of Enriched Taxonomies

Chapter 8

Understanding green skills and jobs through online job advertisement

A central challenge in studying the green transition lies in how we define and measure green jobs and green skills. Taxonomies play a crucial role in this process: by systematically classifying skills, tasks, and occupations, they provide the conceptual infrastructure needed to capture the evolving nature of work. Enriching these taxonomies allows researchers not only to identify "green" activities within existing occupational structures but also to reveal how new competencies emerge and spread across sectors.

In this chapter we apply a taxonomy-based approache to nearly 29 million job postings from 26 European countries and the UK (2019–2023). Using classifications from ESCO and CEDEFOP to define green skills, the study identifies "green jobs" as those advertisements requiring at least one such skill. This taxonomy-driven identification enables a detailed comparison between green and non-green occupations, highlighting differences in educational requirements, experience, and wages. Importantly, the analysis also uncovers how the presence of green skills reshapes traditionally non-green ("brown") occupations, underscoring the fluidity of boundaries between occupational categories once taxonomies are enriched.

By grounding its methodology in the systematic enrichment of occupational and skills taxonomies, this study demonstrates how online job advertisements can be harnessed to provide timely, granular evidence on the labor market implications of

the green transition. Its findings reinforce the importance of dynamic taxonomies as analytical tools for both academic research and policy design, particularly in guiding reskilling and training strategies for a sustainable workforce.

This chapter presents work partially discussed in [32], which is currently under revision for publication in *Labour Economics*.

8.1 Introduction and Motivation

Green jobs and green skills are at the forefront of the global and European political agenda for the ongoing effort to transition to an environmentally sustainable economy. In 2019, the European Green Deal set out a comprehensive development strategy that pivoted on a structural transformation of the EU into a climate-neutral competitive economy by 2050. What distinguished the European Green Deal from previous strategies was its core focus on environmental aspects, which have quickly translated into a set of legislative actions [37, 97] and backed by substantial investments. The European Commission has dedicated €1 trillion towards sustainable investments by the year 2030, utilizing Next Generation EU and the EU's seven-year budget. The execution has been strengthened by major EU initiatives, including the European Year of Skills set for 2024.

The green transition is expected to bring significant changes to labor markets. On the one hand, it offers new opportunities for job creation in areas such as the circular economy, sustainable transport, and renewable energy production. On the other hand, it may negatively impact certain sectors, leading to job losses in highly polluting industries [76, 26]. Job mobility from polluting activities to non-polluting ones is already a driver of greening [72], but targeted policies should be introduced to mitigate these challenges while leveraging opportunities offered by the investments in this area [23, 120]. Evidence from EU countries shows that climate policies might be skill-biased, with stronger positive effects on high-skilled workers—especially technicians and professionals—while manual workers face employment losses, highlighting the need for re-skilling strategies [79]. Comprehensive industrial and labor policies are also necessary to leverage the characteristics of local economies

[105, 87, 9] considering the interplay with other factors, especially digitalization [29, 106]. To address this critical need, our analysis focuses specifically on identifying and analyzing green skills, an aspect that has been somewhat under-explored by the rest of the literature. We do so by analyzing a large and unique set of OJAs across 26 European countries for the period 2019-2023. Our paper contributes to the literature in several dimensions. First, we shift the level of observation from tasks to skills. The latter provides the opportunity to examine the competencies required to perform a job, rather than the activities to be performed, as in the former. This is a significant methodological innovation that offers greater flexibility in the analysis of green jobs. In fact, the categorization of tasks in green and non-green ones leads to a definition of green occupations that is invariant across dimensions that are of high interest for the economic analysis, namely different local labor markets, economic sectors, and time. On the contrary, we are able to account for those differences at a very granular level. Second, we provide a different perspective in analyzing green jobs. According to the task approach, green jobs correspond to green occupations; our approach enables an analysis *within* an occupation, distinguishing between OJAs that contain green skills and those that don't.

This framework does not conflict with the task approach but rather complements it, enabling the study of the role of green skills outside the usual boundaries, for example in the demand of brown occupations. Third, the detail at skill level also allows us to derive evidence on the specificity of skill bundles for green occupations, differences in skill demand at the extensive and intensive margin, and complementarities between green skills and other skill types. Finally, our work presents an analysis of green jobs in the European labor market. While most of the existing literature on green jobs relies on U.S. data, only a few studies have explored the European context. Connolly et al. [33] and Sulich and Soloducho-Pelc [112] use top-down employment statistics, such as sector-level data from Eurostat or national sources, to estimate green employment, focusing on broad sectoral trends rather than occupational detail. Sulich et al. [111] similarly examines the role of green jobs in addressing youth unemployment using aggregated employment data. More recently, Elliott et al. [43] apply a task-based approach to Dutch administrative data by linking ONET's green

tasks to ISCO-coded occupations through a crosswalk, enabling firm-level analysis of green employment composition. However, their focus remains at the occupational level rather than at the skill level. The remainder of the paper is structured as follows: section 8.2 illustrates the data set and the methodology, section 8.3 presents the empirical analysis on green OJAs and green occupations, section 8.4 illustrates the findings on skill bundles, and finally section 8.5 concludes.

8.2 Data and methodology

This study uses data from the Web Intelligence Hub (WIH), a platform created by Eurostat to standardize methods and tools for web data collection for the production of official statistics. The most advanced application of this platform is the WIH-OJA, developed in cooperation with CEDEFOP. The system automatically collects from around 1,000 sources throughout Europe, encompassing major public and private entities within the online labor market, such as specialized job boards, public employment services, websites of private employment agencies, and job sections of national newspapers. These OJAs are downloaded and analyzed to extract key information, including job title, job description, geographic location, the economic sector of the employer, and required skills [25]. The information extracted is utilized to classify the data based on the European Union's official taxonomy, namely the ESCO, reaching the 4th level of the International Standard Classification of Occupations (ISCO). In addition, various other elements of OJAs are extracted and classified in accordance with major international standards: location (NUTS), education (ISCED Levels),[1] economic activity (NACE[2] Rev.2 Divisions), and skills (ESCO skill concepts). More specifically, skill categorization relies on the ESCO Skills pillar taxonomy. This taxonomy serves as a reference dictionary, comprising nearly 14,000 distinct skills, with skill names and synonyms acting as inputs for a classifier.[3] Using the ESCO taxonomy we are able to extract approximately 3000

[1] For a detailed classification, refer to table 8.10

[2] https://ec.europa.eu/eurostat/web/nace

[3] In this work we exclude skills that belong to the ESCO *languages and knowledge* concept groups, to focus on ESCO concepts that represent skills: professional, transversal, and digital skills.

skills in the text contained in OJAs; given the large number of skills identified in the data, we present a broad classification that differentiates between Cognitive, Digital, Manual, Management, Social, and Communication skills (refer to Table 8.9 for more information).

Information on the level of experience is recorded on an 8-point scale as detailed in Table 8.12, while wage data is reported a 13-point scale corresponding to the wage bands defined in Table 8.11.

To perform the analysis, we kept all OJAs for which we have information on occupation, wages, education, experience, and skills for the period 2019-2023. This leaves us with around 30 million observations, spread across 353 ISCO-08 IV digit occupations, 21 sectors, and 26 European countries.

Due to the general absence of wage details in most OJAs, relying on wage data significantly narrows the sample size. However, this does not pose a problem for our analysis for three reasons. Firstly, even with this limitation, we still have nearly 30 million observations, representing 10% of the entire dataset, which is sufficient for a comprehensive analysis. Secondly, since our analysis is performed *within* specific detailed occupations, the risk of sample selection bias is minimized. Lastly, the findings pertaining to other variables such as education and experience in a larger dataset are consistent with those from the subset that includes wage data.[4]

8.2.1 Green skills, green jobs and green OJAs

There is a substantial amount of research investigating the notion of green jobs, yet a universally acknowledged definition continues to be elusive. There are two primary approaches to empirically identifying green jobs [3]. The first approach classifies industries or companies as green, considering all workers within these entities as holding green occupations, and is therefore referred to as the entity-level approach. To identify green industries and companies, researchers often analyse the type of product or the technology employed in production.

[4]Results available upon request.

The second approach begins with the examination and classification of specific occupations into green or non-green categories, termed as the occupation-specific approach. This method can be enhanced by evaluating the tasks within each occupation, transitioning from a simple green or non-green job dichotomy to a continuous metric that quantifies the extent of greenness in occupational activities. Under this approach, green jobs are not confined to specific sectors but can be present across a wide range of industries, with their degree of "greenness" determined by how relevant green tasks are in each occupation [18, 126]. The studies using task-based green measures have generally aimed to identify the characteristics of green jobs as opposed to non-green ones [99, 34–36][5].

Green jobs vs green OJAs

Our approach differs considerably from the studies described above. Instead of identifying green occupations according to a description of tasks and activities such as the ones by O*NET(see Vona et al. [126]), we look *into* occupations differentiating jobs according to their skill content. We use OJAs data and build an analysis of green jobs, identified through green skills requirements and explore their characteristics and skill content.

To better illustrate the nuances that our framework is able to capture, the box presents an example of two OJAs from our 2023 UK sample. Even though both job postings are for the identical specific role of a Civil Engineer (code 2142 of ESCO IV-digit classification), there is a significant variation in the skills required (emphasized in italics). Consequently, the first OJA is categorized as non-green because it lacks green skills, whereas the second one includes them.

Therefore the key feature of our approach is that we can distinguish, *within* occupations, green OJAs from other advertisements. In this regard, we complement the task-based approach by looking at the competencies required for workers to complete such tasks, i.e., skills requested by employers in job postings for such

[5]Another subset of the literature examines the relationship between environmental policy and green employment, but this research mainly focuses on regional levels and explores how green initiatives can foster green skills and create green jobs [91, 125, 126].

Occupation: Civil Engineer

- **OJA A: Civil Engineering Apprentice (Not Green)**
  We are looking for a talented and enthusiastic apprentice to join the Civil Engineering team. We work together to deliver highway and drainage designs, cut fill analysis and much more. We use the leading industry modelling software to build cutting edge designs delivering maximum value for our clients. Within this role you will be entrusted to undertake various tasks to support the design team. Civil Engineering requires excellent *problem solving and analytical skills* and you will be encouraged to think outside the box from the very beginning. Within your role you will be given full training and support to develop design and drafting skills you will continue to expand throughout your career. You will be encouraged to join engineers on site, to observe the construction of designs and meet with adopting authorities and clients and to gain a wider understanding of the construction process. Requirements include: a good standard of *maths and written communication, good ICT skills, an interest in AutoCAD or computer-aided design, and the ability to manage your workload efficiently*.
- **OJA B: Flood Risk Engineer (Green)**
  Due to the increased risk of flooding and the Council's recent declaration of a climate emergency, we look for a Civil Engineer, with a focus on flood risk, to *support development planning and flood risk management in line with relevant legislation*, including the duties of the Lead Local Flood Authority under the Flood and Water Management Act.
  The role involves supporting *strategic flood risk activities, attending countywide meetings, and operating the Council's flood reporting database*. The engineer must assess and manage features that significantly affect flood risk and carry out responsibilities such as investigating flood incidents and overseeing consenting and enforcement work related to ordinary watercourses. The position also includes contributing to the preparation of new legislation and arrangements for the *adoption and maintenance of sustainable drainage systems (SuDS)*.

The text is from the original job advertisement. Words in italic identify the extracted skills.

positions.[6] Moreover, by focusing on the skill content of jobs, we address one of the main criticisms of the task-based method, which is that even within similar occupations, there may be substantial differences in task content that are not captured by occupational-level data [3]. By using the skill content of the OJA, we are able to address this issue and observe within-occupation differences in skill requirements.

Green skills

Since our identification of green OJAs is based on the analysis of their skill requirements, the definition of green skills plays a central role in our approach. We adopt the

[6]In the literature, Autor's definition of a task is quite prominent: a task is "a unit of work activity that produces output"[7]. In contrast, a skill refers to the capability of executing tasks. Skills are determined by various factors such as education, training, and experience [3].

green skills taxonomy developed by EUROSTAT, which integrates methodologies from both ESCO and CEDEFOP. Specifically, a skill is classified as green if it meets two criteria. First, it must be included in the ESCO Green Skills group [44].[7] Second, the classification is augmented using a set of green terms identified by CEDEFOP [27], which capture additional relevant competencies not explicitly listed in ESCO[8]

The outcome of this procedure is a data-driven taxonomy that enhances the standard ESCO classification by incorporating a broader and more dynamic set of skills aligned with environmental objectives. To provide some examples, the set of terms used includes: *advise on offshore renewable energies subjects*, *develop flood remediation strategies*, *use sustainable materials and components*, *coordinate waste management procedures*, *conduct environmental surveys*, *identify energy needs*, *geothermal power generation methods*

Green jobs

As stressed above, the main focus of our analysis is conducted within occupations distinguishing green from non-green OJAs; however, to deepen the analysis and to compare it with the literature, we will also distinguish between green and non-green occupations by employing the OECD Greeness measure [107] which assigns a "green' value to each ISCO-08 IV digit occupation.[9]

To summarize, we will use two distinct concepts to categorize green job ads:

- **Green OJA**. Refers to an OJA that includes at least one Green skill. Each OJA is assigned an occupation code at the ISCO-08 IV digit level. The distinction between green and non-green OJAs is therefore done *within* occupation.
- **Green/Brown Occupation**. An occupation classified as "Green" if it is characterized by a positive value of OECD Greenness score as opposed to "Brown"

[7]The report explains how green skills are identified through a three-step process combining manual labelling, machine learning classification, and final validation. The complete list of ESCO Green skills can be downloaded from ESCO together with the whole Skill Taxonomy.

[8]The detailed description of the construction of the green skill taxonomy is provided in Appendix 8.6.3.

[9]See 8.3.4 for a detailed discussion of this Index.

occupations. The value is assigned at the ISCO-08 IV digit level occupation code and is therefore invariant for OJA belonging to the same occupation code.

8.2.2 Descriptive statistics

The final dataset comprises 26 Countries,[10] 353 ISCO Occupations and 29,236,393 unique OJAs, of which 1,419,207 (5%) are classified as green.

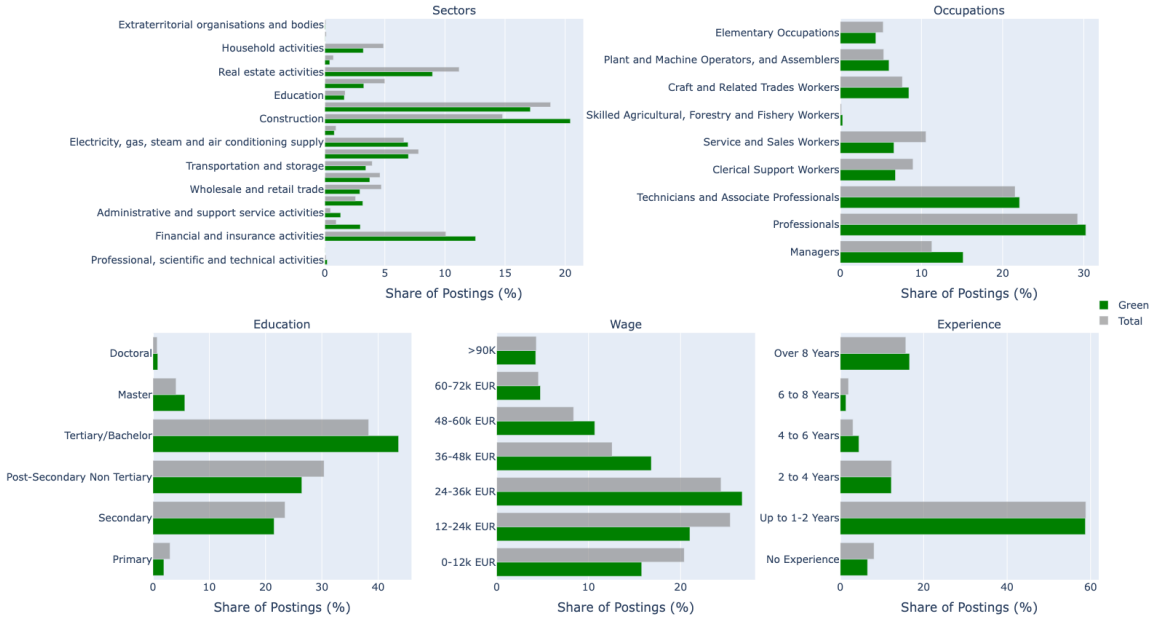
As shown in Figure 8.1, the distributions for green and non-green OJAs highlight some general features of green jobs. Green OJAs are more concentrated in professional, scientific, and technical activities (M), manufacturing (C), energy supply (D), water supply and sewage management (E), construction (F) and financial and insurance activities (K). Green OJA are also more concentrated in high education levels and high wages, while no differences emerge concerning experience.[11] Regarding occupation groups, OJA are more concentrated in high-skill occupations (managers, professionals, and technicians, ISCO-08 Major group 1, 2, and 3, respectively) and specialized workers (craft and related trades workers and plant and machine operators, ISCO-08 Major group 7 and 8, respectively), while they have lower shares of support personnel, retail sales workers, and elementary occupations (clerical support workers, service and sales workers, and elementary occupations' workers, respectively ISCO-08 Major groups 4, 5, and 9).[12]

[10]This includes all EU countries plus the UK and minus Estonia and Lithuania excluded due to the low number of observations.

[11]For sake of readability, we have introduced a coarse group classification for wage, experience, and education with respect to that used in the regression analysis, which are presented in tables 8.12, 8.11, and 8.10.

[12]Compared to the actual employment distribution in Europe, the distribution of OJAs is skewed toward high-skill occupations. This overrepresentation arises from the higher likelihood that high-skill positions are advertised online relative to low-skill ones. Importantly, this sampling characteristic does not bias our analysis: there is neither theoretical justification nor empirical evidence to suggest that green job advertisements follow systematically different online posting patterns compared to non-green positions.

Fig. 8.1 Distribution of Green vs Total Online Job Ads Across Job Characteristics



Note: authors' calculations on WIH-OJA data. The horizontal bars represent the share of green and Total OJA for each category.

8.3 Findings

8.3.1 Profiling of green OJA

We start by presenting the main overall result for European countries. We explore green OJA profiles by relying on a model specification similar to Consoli et al. [34] and summarised by the following equation:

$$y_{i,c,s,t} = \alpha + \beta \, \text{Green}_{i,c,s,t} + \gamma_i + \delta_c + \theta_s + \lambda_t + \epsilon_{i,c,s,t} \tag{8.1}$$

Where the dependent variable is in turn education, experience, and wage; $Green$ is the variable of interest and is a dummy variable indicating if the OJA contains at least one green skill (1) or not (0) and therefore if it is classified as green. The regression is saturated with fixed effects for IV digit occupation (γ_i), country (δ_c), sector (θ_s), and year (λ_t).

Table 8.1 shows that the coefficients for *green* are significant across all models, with positive sign for education and wage, and a negative for experience. Given the number of controlling fixed effects, the interpretation of our result is that within occupations, sector, country, and years, OJAs that contain green skills have higher education and wage and less required experience. Thus, green skills carry a wage and occupation premium, and an experience discount.

Table 8.1 Profiling of green occupations: education, experience and wage

	Education	Experience	Wage
Green OJA	0.0466***	-0.0158***	0.171***
	(0.000987)	(0.00191)	(0.00254)
Constant	4.282***	3.401***	5.381***
	(0.000212)	(0.000408)	(0.000555)
Isco FE	✓	✓	✓
Time FE	✓	✓	✓
Sector FE	✓	✓	✓
Country FE	✓	✓	✓
Observations	29236392	29236392	29236392
R^2	0.281	0.068	0.152

Source: Authors' calculation on WIH-OJA data.
Note: Each observation consists of an OJA. OLS regression using education, wage and experience as the dependent variable. Robust standard errors in parentheses *** p < 0.001, ** p < 0.01, * p < 0.05.

Since the seminal paper by Mincer [85] there is a well-documented correlation between wage experience and education. Therefore, we augment the wage regression following a Mincer-style earning function by adding education and experience as controls. We also augment the experience regression by controlling for the level of education. Table 8.2 shows that the sign and magnitude of the $green$ dummy barely change, even adding as controls interaction terms between occupation and economic activity sector.

It is important to note that the results presented in the main text are based on Ordinary Least Squares (OLS) regressions, which may not be fully appropriate given that the dependent variable is categorical rather than continuous. A more appropriate candidate model would be an ordered logit. Our choice is motivated by practical considerations. Specifically, estimating nonlinear models such as ordered logit in

Table 8.2 Regression Results with additional regressors

	Education	Experience	Wage
Green OJA	0.0466***	-0.0166***	0.169***
	(0.000987)	(0.00191)	(0.00253)
Education		0.0174***	0.0701***
		(0.000356)	(0.000499)
Experience			0.0857***
			(0.000257)
Constant	4.282***	3.326***	4.789***
	(0.000212)	(0.00158)	(0.00236)
Sector FE	✓	✓	✓
Isco FE	✓	✓	✓
Isco*Sector FE	✓	✓	✓
Time FE	✓	✓	✓
Country FE	✓	✓	✓
Observations	29236392	29236392	29236392
R^2	0.281	0.068	0.152

Source: Authors' calculation on WIH-OJA data.
Notes: Each observation consists of an OJA. OLS regression using education, wage and experience as the dependent variable. Robust standard errors in parentheses *** p < 0.001, ** p < 0.01, * p < 0.05.

high-dimensional panel settings, like the one used in this study, involves maximizing complex log-likelihood functions, which is computationally very intensive. However, to assess the robustness of our results—and, more importantly, to interpret the implied odds ratios—we estimate an ordered logit model on a representative sample of 1.5 million observations. The results of this specification are fully consistent with those of the main analysis, supporting the validity of our approach. Further details are provided in the Appendix (Sec. 8.6.5).

The magnitude of the estimated coefficients may appear modest compared to findings in the rest of the literature. However, it is important to emphasize that our estimates are identified *within* detailed occupational groups, whereas most existing studies estimate effects *across* occupations. To revisit the earlier example, our analysis is akin to comparing two civil engineers working in the same country, sector, and year, where the primary difference lies in their skill requirements. Given the narrow scope of analysis, large effect sizes are not expected.

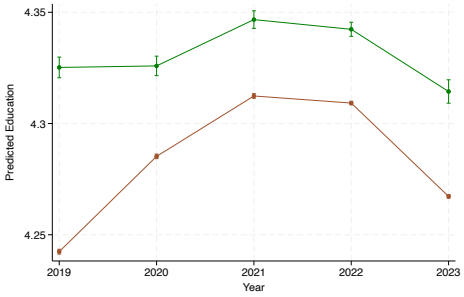
However, estimated coefficients are not as small as they appear. The analysis presented in Appendix (Sec. 8.6.5), based on a representative sample and estimated using an Ordered Logit model, provides odds ratios (Table 8.18) that quantify the association between green skills and job requirements. The results indicate that OJAs mentioning green skills are associated with 10.9% and 17.2% higher odds of requiring a higher education and wage category, respectively, and with 2.4% lower odds of falling into a lower experience category, compared to OJAs without green skills, holding other factors constant.

8.3.2 Profiling of green OJAs over time

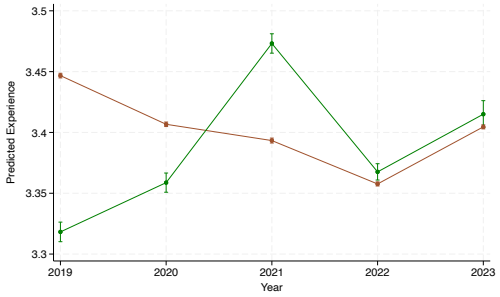
Albeit our data does not provide long time series, it is worth analyzing whether the evidence collected above is stable over time. This is done by interacting the green dummy with a time index. Figure 8.2 plots the estimated coefficients. Overall, the general message is that the education and wage premium carried by green skills holds irrespective of the time period. Note that the wage premium is considerably reduced in 2020 and 2021. This can be explained by the fact that during the Covid period the slowdown in labor demand heavily compressed wage differentials. This pattern is consistent with the sharp increase in experience requirements during the same period. The labor market slack induced by the Covid-19 enabled employers to impose higher experience requirements without corresponding wage premiums, reflecting their enhanced bargaining power relative to job seekers. Note also that as the green premium is clearly present already in 2019, it is likely to be driven by structural elements rather than by policy factors such as the Next Generation EU 2020's plan, which poses a significant emphasis on the green transition. With respect to experience, the negative premium displayed in tables 8.1 and 8.2 appears to be mainly driven by the first years of the sample. Yet, the time series reveals that in general experience requirements have been having opposite developments through time: non-green OJA have witnessed a reduction in requirements, while green OJA's requirements have been growing across the period, with an alignment in 2022 and 2023. This evidence suggests that green technologies have reached a level of maturity and, therefore, standardisation, which makes them similar to

other firms' technologies, making the recruitment policies for green and non-green workers align.

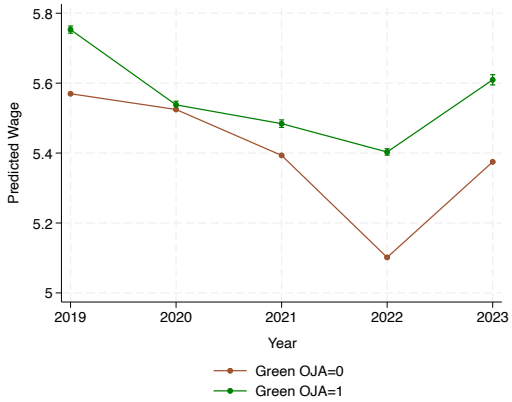
Fig. 8.2 Interaction effects: Time variable with Green OJA Dummy



(a) Years and Education



(b) Years and Experience



(c) Years and Wage

Note: authors' calculations on WIH-OJA data

8.3.3 Profiling OJAs with green skills intensity

As emphasized in the previous paragraphs, our green measure classifies OJAs into two categories—green and non-green—using a binary indicator that dichotomizes the degree of greenness. However, it is worth noting that some OJAs include more than one green skill, suggesting variation in the intensity of green content. This raises the question of whether our results also capture the *intensive margin* of greenness, beyond the simple binary distinction. In Table 8.3, we replace the binary green indicator with two alternative measures to capture the intensive margin of greenness in job postings. The first is the percentage of green skills in an OJA, calculated as the share of green skills relative to the total number of skills listed. This measure reflects the intuition that a higher concentration of green skills better identifies postings where green competencies are central, rather than peripheral, to the job profile. The second measure is a simple count of green skills, which shifts the focus to the absolute number of green competencies required, regardless of their share among the total. Together, these indicators allow us to assess whether the degree of greenness—both in relative and absolute terms—is associated with variation in job requirements. Table 8.3 shows that, also considering the intensive margin, most of the intuition is confirmed: higher greenness carries a higher education and wage premium, while experience remains negative when considering the count and becomes positive but not significant when considering the percentage of green skills.

8.3.4 Green skills and green occupations

So far we have identified green OJAs by focusing on their skill content, regardless of the type of occupation to which the OJA relates. As stressed in section 8.2.1 our approach in this regard differs significantly from most of the literature, which has focused on *occupations* that have been classified as green according to the specific tasks that compose them. In this section, we integrate both approaches by analyzing the relationship between green OJAs and green occupations. For this purpose, we adopt the OECD measure of green occupations as described by Scholl et al. [107]. This index builds on the greenness metrics originally developed by Consoli et al.

Table 8.3 Baseline Regression for Green Skills Fraction (%) and Count

	Education	Experience	Wage	Education	Experience	Wage
Green Skills (%)	0.00201***	0.000124	0.00538***	–	–	–
	(0.0000633)	(0.000117)	(0.000163)			
Green Skills Count	–	–	–	0.0200***	-0.0165***	0.0442***
				(0.000456)	(0.000836)	(0.00105)
Education	–	0.0130***	0.0680***	–	0.0130***	0.0679***
		(0.000361)	(0.000505)		(0.000361)	(0.000505)
Experience	–	–	0.0828***	–	–	0.0829***
			(0.000257)			(0.000257)
Constant	4.283***	3.344***	4.813***	4.283***	3.345***	4.813***
	(0.000206)	(0.00160)	(0.00238)	(0.000206)	(0.00160)	(0.00238)
ISCO FE	✓	✓	✓	✓	✓	✓
Sector FE	✓	✓	✓	✓	✓	✓
Isco*Sector FE	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓
Country FE	✓	✓	✓	✓	✓	✓
Observations	29236273	29236273	29236273	29236273	29236273	29236273
R^2	0.304	0.078	0.169	0.304	0.078	0.169

Source: Authors' calculation on WIH-OJA data.
Note: Each observation consists of an OJA or Green OJA. OLS regression using education, experience, and wage as the dependent variables. Robust standard errors in parentheses. *** p < 0.001, ** p < 0.01, * p < 0.05.

[34] and Vona et al. [125] who, relying on the detailed information provided by the O*NET database for the United States, define the Greenness of an occupation as the share of green-specific tasks relative to the total number of specific tasks associated with that occupation. The OECD measure is built by mapping indicators originally developed for the U.S. Standard Occupational Classification (SOC) system onto the International Standard Classification of Occupations (ISCO-08) used in Europe, at the detailed 4-digit occupational level.

Green OJAs and green occupations are clearly related: the mean greenness is significantly higher for OJAs containing green skills[13]. The Appendix (Sec. 8.6.4) provides some descriptive statistics of the distribution of green occupations in our sample.

The introduction of green occupations enables a more detailed investigation of the characteristics of green OJAs, specifically whether these features are concentrated within green occupations or extend across the broader occupational spectrum.

[13]As confirmed by a two-sample t-test (mean difference = 0.032, $t = -340$, $p < 0.001$)

We begin by distinguishing between green and brown occupations: the former are those assigned a positive greenness score by the OECD, while the latter include all remaining occupations. To investigate the transversal nature of green skills, we initially focus on brown occupations. In table 8.4 we add a brown occupation dummy and its interaction with the OJA green dummy.

Table 8.4 Regression results with OECD Brown occ. dummy

	Education	Experience	Wage
Green OJA	0.0359***	-0.0478***	0.148***
	(0.00165)	(0.00326)	(0.00418)
Brown occupation	-0.0844***	-0.0539***	-0.0351***
	(0.000895)	(0.00180)	(0.00237)
Green OJA × Brown occupation	0.00420*	0.0554***	0.0383***
	(0.00205)	(0.00400)	(0.00522)
Education		0.0178***	0.0709***
		(0.000357)	(0.000498)
Experience			0.0872***
			(0.000257)
Constant	4.346***	3.365***	4.807***
	(0.000701)	(0.00209)	(0.00297)
Isco FE (Digit 3)	✓	✓	✓
Sector FE	✓	✓	✓
Isco*Sector FE	✓	✓	✓
Time FE	✓	✓	✓
Country FE	✓	✓	✓
Observations	29233060	29233060	29233060
R^2	0.279	0.070	0.158

Source: Authors' calculation on WIH-OJA data.
Note: Each observation consists of an OJA. We define Brown occupations as the ones not classified as green in the OECD classification. OLS regression using education, experience, and wage as the dependent variable. Robust standard errors in parentheses *** p < 0.001, ** p < 0.01, * p < 0.05.

Brown occupations are characterized by a lower degree of education, experience, and wage, confirming the results obtained by Consoli et al. [34] for the US and by Bluedorn et al. [13] for a cross section of countries. Interestingly, the interaction term reveals that the premium for green skills is stronger in brown occupations. This

suggests that green skills required by the OJA carry significant value, particularly in roles not explicitly classified as green. This finding highlights the broad relevance of green skills, which are valued across the labor market and rewarded even in non-green occupations.

Next, we focus on the restricted sample of green occupations (Tables 8.5 and 8.6).[14]

Table 8.5 Regression Results with OECD Greenness Levels (Only Greenness > 0)

	Education	Experience	Wage
Green OJA	0.0381***	-0.0811***	0.199***
	(0.00236)	(0.00456)	(0.00592)
Greenness	0.227***	-0.146***	-0.111***
	(0.00598)	(0.0108)	(0.0145)
Green OJA × Greenness	-0.0130	0.113***	-0.257***
	(0.00942)	(0.0174)	(0.0220)
Education		0.0556***	0.0564***
		(0.000776)	(0.00102)
Experience			0.110***
			(0.000491)
Constant	4.432***	3.375***	5.082***
	(0.00105)	(0.00395)	(0.00546)
Isco FE (Digit 3)	✓	✓	✓
Sector FE	✓	✓	✓
Isco*Sector FE	✓	✓	✓
Time FE	✓	✓	✓
Country FE	✓	✓	✓
Observations	7433134	7433134	7433134
R^2	0.246	0.048	0.145

Source: Authors' calculation on WIH-OJA data.
Note: Each observation consists of an OJA of a Green Occupation according to the OECD Classification. OLS regression using education, experience, and wage as the dependent variable. Robust standard errors in parentheses *** p < 0.001, ** p < 0.01, * p < 0.05.

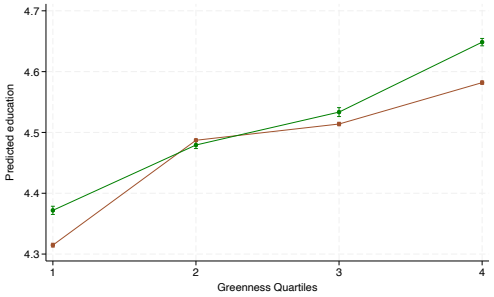
[14]The sample is consistently reduced as we limit the observations to only green occupations

The overall effect for the green OJA indicator is confirmed. However, the greeness effect is not entirely in line with the findings of Consoli et al. [34]. Apart from the obvious difference between the EU and US market, the reason is likely due to the fact that their result is a combination of the extensive (being green or not) and intensive margin (degree of greenness) of occupations. In our case, while in table 8.4 we have looked at the extensive margin, here we focus on the intensive one. When considering only OJAs without green skills, those associated with occupations of higher greenness tend to require higher education but lower experience and offer lower wages. However, the interaction term reveals that the presence of green skills in the OJA is associated with a premium in both education and experience requirements, while the wage effect is negative. To further investigate these dynamics, we divide the greenness index into quartiles and analyze the marginal effects at different levels of greenness, focusing on how they interact with the presence of green skills in the OJA.

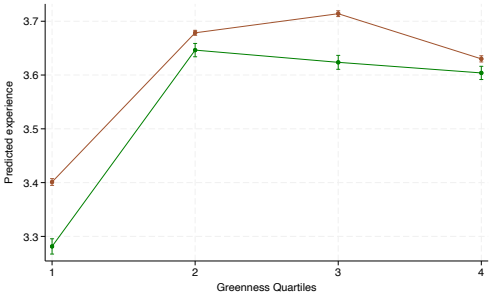
Results are shown in Table 8.6. For non-green jobs, that is for OJAs that do not mention green skills, the greenness effect is positive and generally increasing in the degree of greenness. Notably, the effect is monotonic in education requirements and wage offers. Considering the interaction effect, the premium associated with green skills within green occupations, we observe a growing relation for education requirements, and this is true in particular for occupations in the highest greenness quartile. For the intermediate quartiles (Q2 and Q3) the effects are more nuanced; however, the general finding is that higher greenness carries a higher premium in all three variables, as shown in Figure 8.3. One possible explanation for this pattern is that green know-how is increasingly seen as a standard requirement within green occupations—especially as the degree of greenness rises. In highly green occupations, where tasks are likely more standardized, the wage gap between green and brown OJAs narrows, reflecting a lower premium for green skills. Additionally, occupational composition within greenness quartiles likely contributes to these effects. Quartiles Q1 and Q4 are more polarized: Q4 is dominated by high-skilled occupations, while Q1 includes mostly elementary ones. In contrast, Q2 and Q3

encompass a broader and more heterogeneous mix of occupations across the ISCO-08 classification (see Figure 8.6 in the appendix, Sec. 8.6.4).

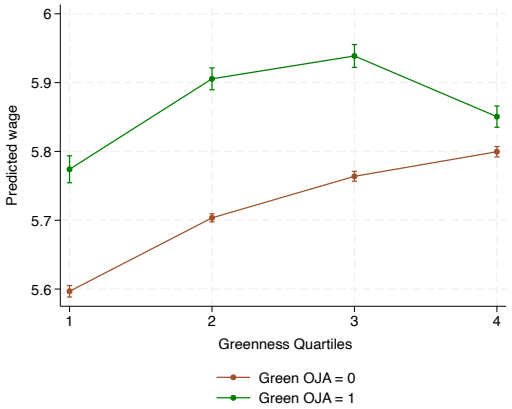
Fig. 8.3 Interaction Effects: Green OJA and Greenness Quartiles on Human Capital variables



(a) Greenness Quartiles and Education



(b) Greenness Quartiles and Experience



(c) Greenness Quartiles and Wage

8.4 Analysing green-jobs skill bundles

The analysis so far provides robust evidence that green job opportunities constitute a segment of labor market demand associated with higher wages, targeting candidates with higher educational attainment and shorter prior work experience. In this context, it becomes essential to examine the skill profile of green OJAs to better understand the specific skill mix they require compared to non-green roles. Previous contributions on this topic are framed within the task-based approach. In such a framework, the identification of green occupations is mediated by the identification

Table 8.6 Regression results with categorized greenness

	Education	Experience	Wage
Green OJA	0.0572***	-0.120***	0.177***
	(0.00322)	(0.00690)	(0.00950)
Greenness Q2	0.172***	0.277***	0.107***
	(0.00210)	(0.00436)	(0.00571)
Greenness Q3	0.199***	0.313***	0.167***
	(0.00249)	(0.00471)	(0.00642)
Greenness Q4	0.267***	0.229***	0.203***
	(0.00260)	(0.00528)	(0.00701)
Green OJA × Greenness Q2	-0.0650***	0.0876***	0.0248*
	(0.00436)	(0.00934)	(0.0124)
Green OJA × Greenness Q3	-0.0376***	0.0294**	-0.00239
	(0.00492)	(0.00952)	(0.0127)
Green OJA × Greenness Q4	0.00939*	0.0936***	-0.126***
	(0.00443)	(0.00924)	(0.0122)
Education		0.0540***	0.0552***
		(0.000777)	(0.00102)
Experience			0.110***
			(0.000491)
Constant	4.315***	3.160***	4.956***
	(0.00160)	(0.00462)	(0.00631)
Isco FE (Digit 3)	✓	✓	✓
Sector FE	✓	✓	✓
Isco*Sector FE	✓	✓	✓
Time FE	✓	✓	✓
Country FE	✓	✓	✓
Observations	7433134	7433134	7433134
R^2	0.247	0.048	0.145

Source: Authors' calculation on WIH-OJA data.
Note: Each observation consists of an OJA of a Green Occupation according to the OECD Classification. OLS regression using education, experience, and wage as the dependent variable. Robust standard errors in parentheses *** p < 0.001, ** p < 0.01, * p < 0.05.

of the green tasks they are required to perform. Once the occupation is assigned a greenness score, skills connected to these occupations are collected and analysed [118]. These studies have found that green jobs are characterized by a greater emphasis on non-routine analytical tasks [18, 34]. However, as explained in section 8.2.1 our approach diverges from the existing literature by focusing on the skills demanded in green jobs, rather than on the activities performed in green occupations. Shifting to the analysis of skills, the available evidence is much scarcer. Vona et al. [125] highlight two key sets of green skills that characterise green occupations: engineering skills for designing and producing green technologies, and managerial skills for implementing and overseeing environmental organizational practices.

In this section, we leverage information on skill sets reported in OJAs to analyze the variety and specialization of green job opportunities. Specifically, we exclude green skills from green OJAs and compare the residual skill bundle with that of non-green OJAs. The analysis focuses on the demand for key skill categories—cognitive, social and communication, digital, manual, and managerial—to assess whether green job postings differ in their broader skill profiles beyond their environmental content. The analysis is structured along two main dimensions, each capturing distinct aspects of skill requirements. First, in Section 8.4.1, we examine differences in the *variety of skill bundles*, highlighting which skill types drive the divergence and overlap in skill sets within occupations. Second, in Section 8.4.2, we focus on the *specialization of skill requirements* in green job opportunities compared to non-green ones, aiming to identify which skill groups are most distinctive of green occupations.

Table 8.7 Average Number of Skills per Group, by Green and Non-Green OJAs

OJA Type	Cognitive	Digital	Management	Manual	Soc. & Comm.	Total
Non-Green (0)	0.67	1.29	3.91	0.29	3.94	10.10
Green (1)	0.89	1.46	4.38	0.36	4.24	11.33

Note: Authors' calculations based on WIH-OJA data. Each value represents the average number of skills per skill group across green and non-green job ads.

To set the stage, Table 8.7 reports the average number of skills listed in green and non-green OJAs. Overall, the two groups exhibit a similar total number of skills, as well as a comparable distribution across skill categories. In both green and non-green

postings, management and social and communication skills have the highest average counts, suggesting their central role in job requirements regardless of environmental content.

8.4.1 Variety of the skill bundle

The first dimension of interest is skill variety, which captures the richness of the skill portfolio demanded by recruiters. Assessing the degree of overlap between the skill sets of green and non-green job opportunities helps determine whether work involving environmentally sustainable competencies is performed in a uniform way compared to jobs that do not require such skills.

We construct a measure of similarity between skill sets, computing the Jaccard distance between the skill requirements of green OJA compared with non-green OJAs within the same occupational code (ISCO-08 IV digit), defined as follows:

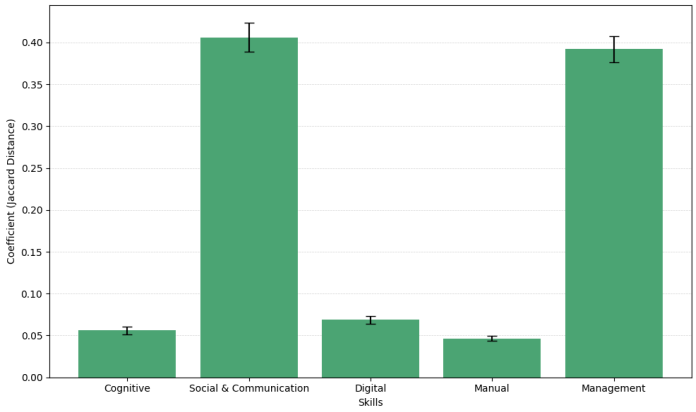
$$J_i = 1 - \frac{S_{(i,g)} \cap S_{(i,ng)}}{S_{(i,g)} \cup S_{(i,ng)}} \tag{8.2}$$

where i represents the occupation, g represents the subset of green OJAs, and ng represents non-green ones. $S_{(i,g)}$ denotes the set of skills associated with the green OJA for occupation i, and $S_{(i,ng)}$ denotes the set of skills in non-green OJA for occupation i. This formula computes the dissimilarity between the two sets $S_{(i,g)}$ and $S_{(i,ng)}$, where the numerator represents the intersection of the two sets and the denominator represents their union so that the ratio tends to 1 the more the two sets are similar. The Jaccard distance ranges from 0 to 1, with 0 indicating the sets are identical, and 1 indicating the sets have no common elements.

In the appendix section (8.6.5, we examine the overall distribution of Jaccard distances across years, countries, and occupational groups. The evidence presented in Figure 8.8a also highlights a positive relation between overlap in skill requirements and the ranking in the occupational hierarchy. Green managerial OJAs share around 60% of their skill sets with their non-green counterparts, while only 40% of skill requirements in elementary occupations are common to green and non-green job

opportunities within the same occupational group. The trend appears generally stable over time.

Fig. 8.4 Composition of Jaccard Distance



Note: authors' calculations on WIH-OJA data. The bars report the values and standard errors of a constrained regression of the overall Jaccard distance on each component considering country, sector, year and occupation fixed effects.

Following the definition given in equation (8.2), we computed the Jaccard distance in the skill bundle at occupation level between green and non-green OJAs. We have grouped the skills into 5 different mutually exclusive and exhaustive classes: Cognitive skills, Social and communication skills, Digital skills, Manual skills, Management skills, using the ESCO taxonomy as described above. Under the condition of skills belonging to one and only one of the mutually exclusive skill groups, the overall Jaccard distance has the property of additivity, meaning that it is the sum of the distances for all different skill types.

Figure 8.4 presents the coefficients from a constrained regression of the overall Jaccard distance on the five group-specific Jaccard distances, controlling for occupation, country, and year fixed effects. This specification leverages the additive structure of the Jaccard index to quantify the marginal contribution of each skill category to the overall divergence in skill requirements between green and non-green jobs.[15] The results indicate that differences in management and social and communication skills account for the largest shares of this divergence, with estimated coefficients of

[15]The constrained regression allows to estimate shares while taking into account detailed fixed effects.

0.39 and 0.40, respectively. The contributions of *cognitive*, *digital*, and *manual* skills are substantially smaller. These findings underscore the central role of higher-order interpersonal and organizational skills in the green transition.

8.4.2 Skill specialization of green jobs

Variety measures the scope and diversity of skills required to perform green jobs compared to non-green ones. However, this indicator does not capture the importance or intensity of specific skills within the job profile. To address this limitation, we introduce a measure of specialization, which reflects the relative intensity of skill demand. This allows us to compare how prominently different skill groups feature in green versus non-green job postings. For this purpose, we adopt a well-established metric in the literature—the location quotient, also known as Revealed Comparative Advantage (RCA). This class of indicators was first developed in international trade theory by [8]. The rationale is to use post-trade measures to infer the comparative advantages in the production of traded goods of a region in a multilateral context. More recently, RCA has been translated into the economic geography literature to describe regions' specialization in productive activities (see [119] for an overview of the literature and the most recent advances). In the labor market, a notable application of this concept is provided by [1] who apply it to the skill distribution using the O*NET intensity measures. In analogy with this approach, we adopt skill frequency as a measure of relevance of skill demand for each occupation and derive our measure of skill specialization based on the concept of revealed comparative advantages.

For each year t, we define a set of ISCO-08 IV digit occupation $O_t = o_k, k = 1, \ldots, K_t$ and its skill set $S_t = s_j, j = 1, \ldots, J_t$ defined on the domain of observed skill in the European online labor market. We define the skill frequency as:

$$sf_{ij} = \frac{\sum_{k=1}^{n} I(o_k = o_i) \cdot I(s_i = s_j)}{\sum_{k=1}^{n} I(o_i = o_k)} \tag{8.3}$$

where I denotes the indicator function and $\sum_{k=1}^{n} I(o_k = o_i) \cdot I(s_k = s_j)$ the count of the occurrences of the skill s_j for occupation o_k. The term $\sum_{i=1}^{n} I(o_i = o_k)$ represents the

total number of observations of occupation o_k. Iterating over skills and occupations, we obtain a matrix $\underset{O_t \times S_t}{M}$ of the skill frequency for each pair of occupations $i \in O$ and skills $j \in S$. The revealed comparative advantage, for occupation o_i and skill s_j is defined as :

$$rca_{ij} = \frac{sf_{ij} / \sum_{j=1}^{J_t} sf_{ij}}{\sum_{i=1}^{I} sf_{ij} / \sum_{i=1}^{I} \sum_{j=1}^{J_t} sf_{ij}} \tag{8.4}$$

If $rca_{is} > 1$, the skill is over-represented in the occupation compared to the market, indicating a specialization in that specific skill. In the original formulation, the rca_{is} is bounded between $[0, +\infty)$ and lacks symmetry around its neutral value. To address this limitation, we use the symmetric formulation proposed by [66], known as the Symmetric Revealed Comparative Advantage (RSCA):

$$rsca_{ij} = \frac{rca_{ij} - 1}{rca_{ij} + 1}$$

This symmetric formulation maps the metric to a homogeneous interval, with values ranging between $[-1, +1]$, improving the efficiency of the estimates in a regression framework [66]. The Revealed Symmetric Comparative Advantage (RSCA) has been computed for each skill and subsequently grouped following the skill taxonomy developed in the previous section. In this way, it is possible to assess the specialization of OJAs for each skill category.

Our results, presented in Table 8.8, indicate that green OJAs exhibit a higher degree of job-specific and less common skill sets compared to non-green counterparts. In particular, cognitive and manual skills are the main drivers of specialization in green OJAs. By contrast, digital and managerial skills appear to be more widely shared across both green and non-green job postings, suggesting they are more transversal in nature.

The distinction between cognitive and manual skills highlights the unique demands of green jobs, where workers are often required to combine specialized problem-solving abilities with practical, hands-on expertise. In contrast, the widespread presence of digital and managerial skills across both green and non-

Table 8.8 RSCA Results with green and not-green occupations

	RSCA
Green OJA	0.0805***
	(0.00142)
Management skills	-0.0426***
	(0.00133)
Cognitive skills	0.0840***
	(0.00194)
Digital skills	0.0930***
	(0.00186)
Manual skills	0.143***
	(0.00208)
Green OJA × Management skills	-0.00276
	(0.00210)
Green OJA × Cognitive skills	0.0232***
	(0.00313)
Green OJA × Digital skills	-0.0351***
	(0.00293)
Green OJA × Manual skills	0.00753*
	(0.00345)
Constant	0.251***
	(0.000878)
Isco FE 3Digit	✓
Time FE	✓
Country FE	✓
Observations	926671
R^2	0.056

Source: Authors' calculation on WIH-OJA data.
Note: Each observation consists of an occupation-skill pair. OLS regression uses the RSCA measure as the dependent variable. Robust standard errors in parentheses *** p < 0.001, ** p < 0.01, * p < 0.05.

green occupations likely reflects broader labor market trends—namely, increasing digitization and the rising relevance of managerial competencies across a wide range of sectors, not limited to environmentally focused roles. This finding is consistent with part of the existing literature, such as the work by [34] and [18], who find that green occupations tend to exhibit significant differences from non-green occupations, especially in terms of higher levels of non-routine, analytical skills, including creative problem-solving. We find that the difference in skill requirement between green and non-green OJAs is present even within occupations. The emphasis on manual skills in green OJA could further signal the non-standard nature of processes to be implemented for the transition to more sustainable production methods [22]. By integrating the findings on skill variety and specialization, we observe that cognitive and manual skills contribute only marginally to the overall difference in skill bundles between green and non-green jobs in terms of variety. However, these skills are critical in defining green jobs, as their demand is more concentrated, so that green jobs feature cognitive and manual skills that are distinctly unique compared to non-green ones.

8.5 Conclusions and policy implications

The task-based literature has shed light on which components of job activities are relevant to green processes, offering valuable insights into their functional relevance within occupations. The evidence emerging from our study provides complementary knowledge on the profile of green jobs from the point of view of the required skills expressed by employers. Our analysis suggests that the labor market is segmented, with green jobs differing markedly from non-green jobs—even within narrowly defined occupational categories. Specifically, green jobs tend to require higher levels of education, while employers appear to prefer less experienced candidates, possibly due to the specificity and evolving nature of green production processes. Most notably, we find a significant wage premium associated with green skills, underscoring the market value attributed to environmental competencies.

Interestingly, jobs in occupations that do not exhibit tasks related to green processes i.e. brown occupations, tend to converge to green occupations in terms of education requirement and offered wage if they require green skills, indicating that it is the presence of green competencies—rather than the occupation itself—that drives higher job requirements and remuneration. Moreover, the relevance of green skills for occupations without green tasks highlights the advantages of a combined approach in grasping signals of ongoing changes in jobs as they emerge from the market. Another important aspect of green jobs is the specificity of their skill requirements. The observed differences in skill bundles are largely driven by social, communication, and managerial skills—components of a broader portfolio that supports workers' integration into organizational structures and their ability to confidently lead or coordinate processes. Most importantly, what makes the profiles of green jobs truly unique is the presence of two additional sets of skills: cognitive and manual. This combination suggests that workers in green roles are expected to operate simultaneously at the level of analytical thinking and planning, while also being proficient in hands-on, practical tasks. The coexistence of these skill demands reflects the idiosyncratic nature of the strategies and solutions that organizations adopt to achieve their green objectives, often requiring tailored approaches that blend conceptual understanding with operational execution.

These results hold significant implications for interventions and active labor market policies aimed at fostering the creation of green jobs. Based on our analysis, upskilling initiatives are expected to be more effective and less complex to implement than reskilling ones, largely due to the diverse range of competencies required for green jobs, which also aligns with the demand for higher levels of education in green jobs. Returns to effective interventions are expected to be higher than general training, considering the wage premium the market offers for green skills. Our findings also emphasize the need to integrate green principles in education and training paths for all students and learners. Finally, they suggest that an accomplished green approach to economic activities does not consist of introducing environmental sustainability as a goal but radically revisiting production processes.

8.6 Appendix

8.6.1 Considerations on the use of OJAs

This section illustrates some studies on the potential limitations of OJAs for the analysis of the labour demand.

Recently, experts from Cedefop assessed the representativeness of OJA data in measuring the number of vacancies across EU Member States ([90]). In their study, the authors evaluated coverage biases in online job advertisement (OJA) data at the sectoral, occupational, and geographic levels. This analysis compared job titles and occupations in OJAs with those reported in the Labour Force Survey (LFS) and Job Vacancies Survey (JVS) to identify discrepancies and representation patterns. They also examined the presence of various occupations in OJAs, focusing on trends related to labour market size across different Member States. The assessment revealed significant coverage biases in OJAs compared to traditional data sources like the LFS and JVS. High-skilled occupations and more industrialized regions were often overrepresented, while certain occupations, especially in smaller Member States, were underrepresented. However, the authors also observed a growing trend among employers to use OJAs for recruiting across a wider range of roles, including non-managerial and lower-skilled positions. This suggests a broader acceptance of OJAs as a recruitment tool and indicates a potential improvement in their coverage. The paper also highlights limitations in the methodologies of existing surveys like the LFS and JVS, such as varying response rates and inconsistent methodologies across Member States, which complicate comparisons and affect the reliability of the data.

In another work, the OECD also contributed to this research by analyzing the representativeness of web-scraped vacancy data from Lightcast for Australia, Canada, the United Kingdom, and the United States between 2015 and 2022 (Tsvetkova et al. [117]), as well as for some European countries[16] between 2019 and 2022 (Vermeulen and Amaros [122]). In both cases, the researchers compared online job postings with

[16] Austria, Belgium, Germany, the Netherlands, Portugal, Spain, and Sweden. Additionally, Eurostat data is used for Bulgaria, Hungary, and Romania

traditional data sources, such as public employment services (PES), acknowledging that each source has some disadvantages. Focusing on European data, the work shows a nearly 50% increase in online job postings across many European countries from 2018 to 2019, reflecting improvements in data collection methods. However, there is significant regional variability in the ratio of Lightcast vacancies to national sources, with countries such as Austria and Germany showing fluctuations between undercounting and overcounting, while Sweden and Belgium consistently report higher vacancy numbers. While Lightcast effectively captures overall labour market trends, discrepancies exist based on region and sector, particularly with urban areas being better represented than rural ones. The study concludes that Lightcast can complement traditional sources like Eurostat and national public employment services (PES), but the correlation with administrative sources can be weak in some regions. Analysts should be cautious in relying solely on online job postings, as they may not fully represent the labour market, and PES data also has limitations.

Awareness of the limitations discussed is crucial, and researchers should account for them when designing studies using this type of data. To ensure validity, appropriate data checks are recommended, with the nature and extent of these checks tailored to the specific requirements of each project using OJA data. A useful example comes from the OECD research referred to above, where the authors detailed the techniques used by the data provider to scrape OJA data and addressed various algorithmic challenges. For instance, they noted that job titles containing terms associated with multiple occupational groups are more likely to be misclassified by the algorithm (Tsvetkova et al. [117]). Another issue involves geographical information: advertisers may list a large nearby city as the job location while mentioning the actual suburban location elsewhere in the announcement. Correcting such errors can be essential to maintaining the validity of the results in certain applications.

Importantly, the implications of these representativeness limitations depend on the research question under investigation. In the context of this thesis, and in particular in the analysis of green skills and green job profiles, the focus is not on estimating aggregate vacancy levels or comparing demand across occupations

or regions. Rather, the analysis is conducted conditionally within occupations, examining changes in the skill content of job advertisements associated with a given occupational group. By comparing skill requirements within the same occupation, the approach is less sensitive to cross-occupational or cross-sectoral coverage biases that typically affect OJA data. While the analysis still reflects only the portion of the labour market observable through online postings, the within-occupation perspective mitigates concerns related to differential representation across occupations and regions. Consequently, the results should be interpreted as capturing relative changes in skill demand within occupations, rather than providing a comprehensive or population-representative description of overall labour demand.

Beyond issues of coverage and representativeness, recent evidence suggests that online job advertisements may also exhibit systematic biases in the way skill requirements are articulated. A recent study by the Joint Research Centre (JRC)[17] shows that employers are significantly more likely to advertise vacancies online for highly skilled occupations than for elementary or routine positions, with the probability of online posting being orders of magnitude higher for managers, professionals, and technical roles. Moreover, the analysis indicates that OJAs tend to emphasise formal, standardised, and socially desirable skills—particularly advanced digital and soft skills—often beyond the actual requirements of the job. Qualitative evidence from interviews with human resource professionals further suggests that certain skill categories, especially soft skills, are included in job advertisements in a largely ritualistic manner, despite being vague or weakly operationalised, while less attractive but equally relevant job attributes, such as routineness or standardisation, are systematically underreported. These findings imply that OJA-based analyses may overstate the demand for certain skill types and underrepresent others, reinforcing the need to interpret extracted skill signals as indicative of advertised employer preferences rather than as a direct measure of task-level job requirements.

[17] https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/online-job-ads-promising-yet-biased-data-employment-and-education-policy-2024-01-24_en

8.6.2 Mappings

Table 8.9 Skill Groups and their Corresponding ESCO Skills Codes and Descriptions

Groups	ESCO Skill Codes	Description
Cognitive skills	S2: Information skills, T2: Thinking skills, T1.1: Mastering languages, T1.2: Working with numbers, T6: Life skills and competences	Skills related to problem-solving, information processing, and learning, including language mastery and numerical abilities.
Social and communication skills	T4: Social and communication skills, S3: Assisting and caring, S1: Communication collaboration and creativity	Skills referred to the ability to communicate, interact and engage with colleagues, clients, and customers.
Digital skills	S5: Working with computers, T1.3: Working with digital devices and applications	Skills that encompass a range of different abilities that allow an individual to use ICT tools at different levels.
Manual skills	T5: Physical and manual skills and competences, S6: Handling and moving, S7: Constructing, S8: Working with machinery and specialized equipment	Skills related to physical and manual labor, including handling, moving, constructing, and working with machinery.
Management skills	S4: Management skills, T3: Self-management skills and competences	Skills including leadership, organization, and decision-making.

Table 8.10 Education classes

ID	Education Level
1	Primary education
2	Lower secondary education
3	Upper secondary education
4	Post-secondary non-tertiary education
5	Short-cycle tertiary education
6	Bachelor or equivalent
7	Master or equivalent
8	Doctoral or equivalent

Table 8.11 Wage classes

ID	Wage Range
1	0 - 6,000 EUR
2	6,001 - 12,000 EUR
3	12,001 - 18,000 EUR
4	18,001 - 24,000 EUR
5	24,001 - 30,000 EUR
6	30,001 - 36,000 EUR
7	36,001 - 42,000 EUR
8	42,001 - 48,000 EUR
9	48,001 - 54,000 EUR
10	54,001 - 66,000 EUR
11	66,001 - 78,000 EUR
12	78,001 - 90,000 EUR
13	> 90,001 EUR

Table 8.12 Experience classes

ID	Experience
1	No experience
2	Up to 1 year
3	From 1 to 2 years
4	From 2 to 4 years
5	From 4 to 6 years
6	From 6 to 8 years
7	From 8 to 10 years
8	Over 10 years

Table 8.13 Sector Category Mappings

Sector Category	Code
Professional, scientific and technical activities	A
Manufacturing	B
Financial and insurance activities	C
Public administration and defence; compulsory social security	D
Administrative and support service activities	E
Human health and social work activities	F
Wholesale and retail trade; repair of motor vehicles and motorcycles	G
Accommodation and food service activities	H
Transportation and storage	I
Information and communication	J
Electricity, gas, steam and air conditioning supply	K
Other service activities	L
Construction	M
Arts, entertainment and recreation	N
Education	O
Water supply, sewerage, waste management and remediation activities	P
Real estate activities	Q
Agriculture, forestry and fishing	R
Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use	S
Mining and quarrying	T
Activities of extraterritorial organisations and bodies	U

8.6.3 ESCO Green Skills Taxonomy

The Cedefop Green Skills Taxonomy has been obtained using a data-driven framework designed to classify and analyse skills relevant to the green transition. It enhances and integrates several existing frameworks through a multi-step, bottom-up methodology. In the *first phase*, Cedefop compiled a "bag of green-related words" encompassing terms associated with green technologies, environmental tasks, tools, and competencies. This lexicon was developed from several authoritative sources, including the UN System of Environmental Economic Accounting (CEPA and CReMA classifications), the IRENA Global Renewables Outlook, LinkedIn's list of green skills, the SGG Singapore Green Economy Skills Framework, the European Commission's JRC GreenComp Framework, the U.S. O*NET Green Skills Taxonomy, and the ESCO Green Skills Taxonomy. This produced an initial list of 140 green terms in English, covering topics such as *solar energy*, *biodiversity*, *energy consumption maintenance*, and *corporate social responsibility*.

In the *second phase*, over 6 million online job advertisements (OJAs) collected by Eurostat through the WIH-OJA (Web Intelligence Hub – Online Job Advertisements) system in 2019 were used as a corpus to train a word embedding model. The model expanded the original list by identifying semantically similar terms and lexical variations, leading to an enhanced vocabulary of 182 validated green skill terms. These were then translated and validated by national experts in all WIH-OJA system languages. The enriched terms include examples like *green computing*, *ecosystems*, *environmental engineering*, *circular economy*, *hydroelectricity*, and *combined heat and power generation*.

In the final step, each term was mapped to the ESCO Green Skills Taxonomy using cosine similarity. Where no suitable match existed, the term was added as a new entry, allowing the taxonomy to remain responsive to emerging labour market trends and technologies. For further information, see Cedefop [27].

8.6.4 OECD Greenness Index: Details

Descriptive statistics

Table 8.14 Summary Statistics for Greenness by Green OJA

Green OJA	Mean	SD	Min	Max
0	0.0403011	0.1085415	0	1
1	0.0722942	0.1507054	0	1
Total	0.0418543	0.1111719	0	1

Statistics by Greenness Quartiles

Table 8.15 reports the distribution of online job advertisements (OJAs) across the four greenness quartiles based on the OECD greenness index.

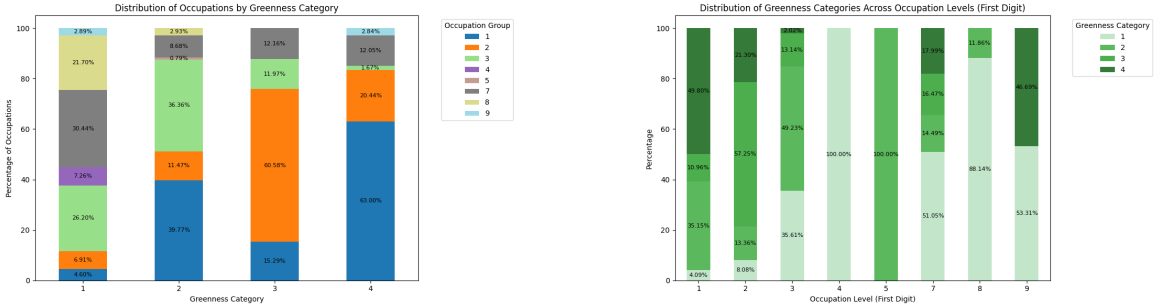
The data are relatively balanced, with each quartile representing approximately one-quarter of the total sample: 27% of the observations fall into Quartile 1 (lowest greenness), 27% into Quartile 2, 22% into Quartile 3, and 24% into Quartile 4 (highest greenness). This even split provides a solid basis for comparing job characteristics across the greenness spectrum.

Table 8.15 Distribution of Greenness Quantiles

OECD Greenness Quartiles	Frequency	Percent	Cumulative
1	2,012,167	27.07	27.07
2	2,003,910	26.96	54.03
3	1,625,019	21.86	75.89
4	1,792,052	24.11	100.00
Total	**7,433,148**	**100.00**	

Figure 8.5 offers an overview of how occupational categories differ by greenness level.

Fig. 8.5 Composition of Greenness Quartiles



(a) 1Digit Occupation by Greenness Quartiles

(b) Greenness Categories across 1D occupations

Note: authors' calculations on WIH-OJA data

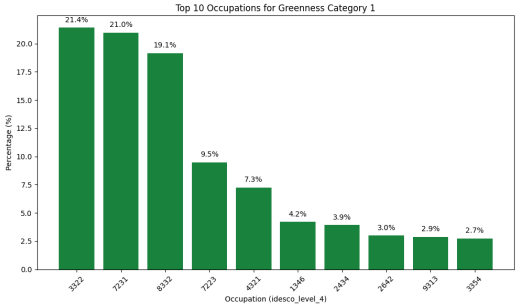
Panel (a) presents the distribution of 1-digit ISCO occupations by greenness quartile, revealing that lower quartiles are primarily composed of service, sales, and manual roles. In contrast, higher quartiles show a notable shift toward professional and managerial occupations, indicating increasing complexity and skill requirements as greenness rises. Panel (b) further illustrates this pattern by displaying the proportion of each major ISCO group within each quartile. Managers (ISCO 1) and Professionals (ISCO 2) are increasingly prevalent in quartiles 3 and 4, reinforcing the link between greener job content and more advanced occupational categories.

Figure 8.6 provides a more granular look at the top ten ESCO occupations within each greenness quartile. In quartile 1 (below the 25th percentile), occupations such as buyers (3323), commercial sales representatives (3322), and motor vehicle mechanics and repairers (7231) are the most common, reflecting operational and routine service tasks. Quartile 2 (25th to 50th percentile) begins to include more

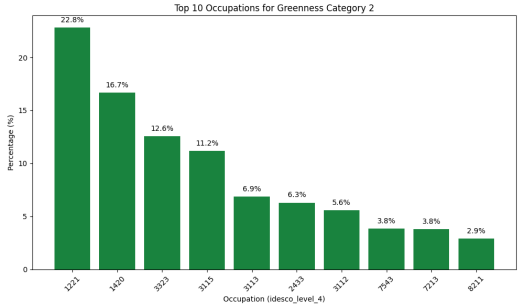
managerial roles like research and development managers (1223) and retail and wholesale trade managers (1420). As greenness increases further in Quartile 3 (50th to 75th percentile), technical and analytical occupations such as organization analysts (2421) and electronics engineers (2152) become more frequent, reflecting a stronger connection to sustainability-focused tasks. Finally, quartile 4 (above the 75th percentile) is dominated by occupations like sales and marketing managers (1221) and R&D managers (1223), roles typically associated with higher education levels and strategic responsibilities.

Overall, the figures suggest a clear gradient in occupational composition across greenness levels. Higher greenness quartiles are associated with increasing specialization, managerial oversight, and innovation-oriented tasks, whereas lower quartiles are more concentrated in routine, sales, and operational roles. This pattern aligns with our empirical findings, where greener job ads tend to be linked with higher education and wages, and slightly less required experience.

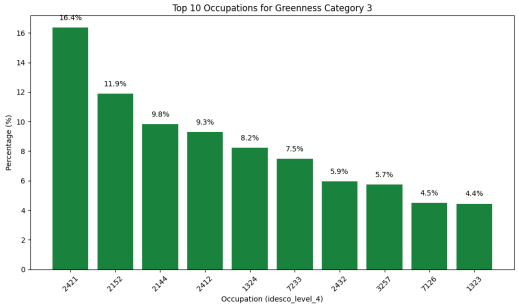
Fig. 8.6 Top 10 ESCO Occupations by Greenness Quartile



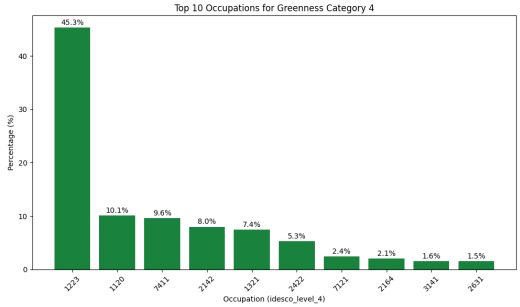
(a) Quartile 1: below 25%



(b) Quartile 2: above 25% and below 50%



(c) Quartile 3: above 50% and below 75%



(d) Quartile 4: above 75%

Note: authors' calculations on WIH-OJA data

8.6.5 Ordered logit Model

Sample extraction

To facilitate model estimation on a computationally manageable dataset while preserving the joint distribution of key categorical variables, we constructed a stratified random sample from the full dataset. Specifically, observations were grouped into strata defined by the cross-classification of `Green` status, year, country, macro-sector, and 4-digit ISCO occupation code. Within each stratum, approximately 5% of observations were randomly selected (rounded up to ensure at least one observation per cell). This proportional stratified sampling procedure ensures adequate representation of both common and rare subgroups—such as green jobs—across all dimensions, thereby reducing the risk of omitted category bias or model instability due to sparse cells, while substantially reducing the overall sample size for tractable estimation.

Table 8.16 Distribution of Green and Non-Green Jobs in Sample

Green	Frequency	Percent	Cumulative
0 (Non-Green)	1,468,332	92.51%	92.51%
1 (Green)	118,953	7.49%	100.00%
Total	1,587,285	100.00%	

Table 8.17 Summary Statistics of Observations per Stratification Cell

Variable	Obs	Mean	Std. Dev.	Min	Max
obs_per_cell	1,587,285	444.85	960.79	1	6,486

Estimates

We replicate the baseline model on the sample described above using an ordered logit specification which is more suitable to capture a categorical dependent variable. The results from the ordered logit regression, as shown in Table 8.18 are consistent with the OLS findings in Table 8.2: green OJAs are associated with higher education and salary levels, and slightly lower experience requirements. Odds ratios above 1 for education (1.109) and salary (1.172), and slightly below 1 for experience (0.976),

mirror the direction and significance of the OLS coefficients. For comparison, Table 8.19 reports the OLS coefficients obtained on the restricted sample.

Table 8.18 Ordered Logit Regression Results for *Green OJA*: Coefficients and Odds Ratios

Model	**(1) Education**	**(2) Experience**	**(3) Salary**
Log-Odds Coefficients			
Green OJA	0.103***	−0.0242***	0.159***
	(0.00563)	(0.00564)	(0.00536)
Odds Ratios (Exponentiated Coefficients)			
Green OJA	1.109***	0.976***	1.172***
	(0.00625)	(0.00550)	(0.00628)
Observations	1,587,285	1,587,285	1,587,285

Standard errors in parentheses. Coefficients are from ordered logit regressions. Odds ratios are exponentiated coefficients. Base category: Green OJA = 0.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

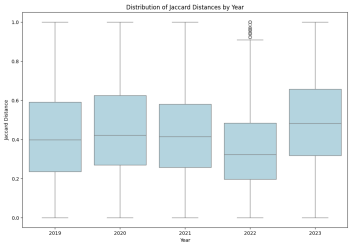
Table 8.19 Regression Results: Education, Experience, and Salary- OLS on Sample

	(1) Education	**(2) Experience**	**(3) Salary**
GREEN OJA	0.0737***	−0.0389***	0.201***
	(0.00364)	(0.00658)	(0.00920)
Education	–	0.0205***	0.0680***
		(0.00150)	(0.00212)
Experience	–	–	0.0868***
			(0.00111)
Constant	4.290***	3.316***	4.765***
	(0.00093)	(0.00670)	(0.0101)
Observations	1,587,282	1,587,282	1,587,282
R^2	0.271	0.066	0.150
Adj. R^2	0.270	0.066	0.150

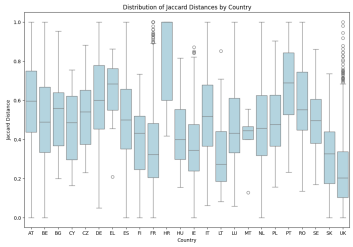
Standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Jaccard Distances distribution

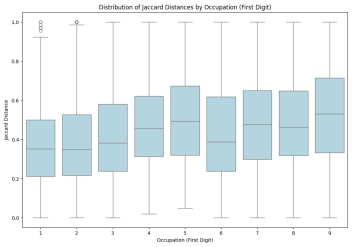
Fig. 8.7 Jaccard Distance: overall distribution by Year, Country, and 1Digit Occupation



(a) JD distribution by Year



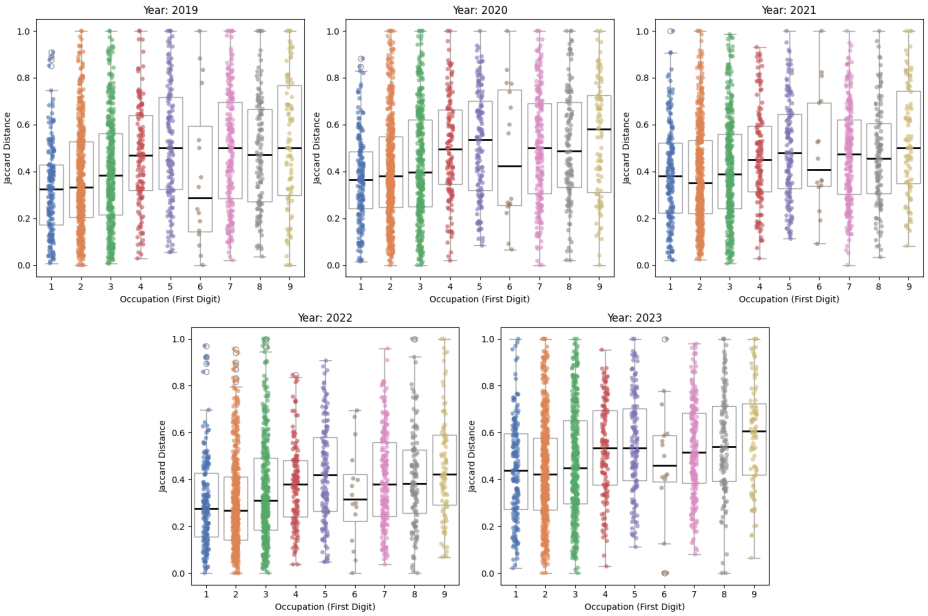
(b) JD distribution by Country



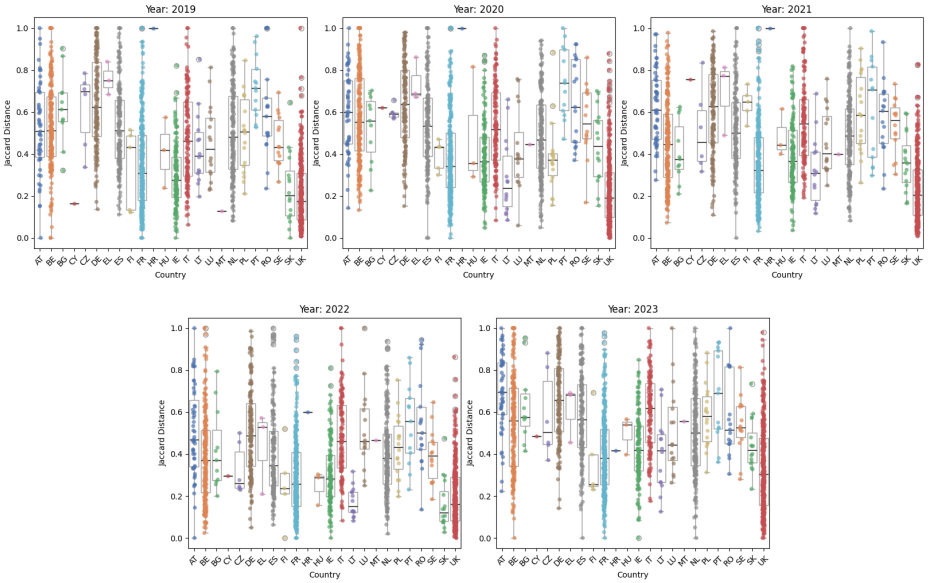
(c) JD Distribution by 1Digit ESCO

Note: authors' calculations on WIH-OJA data

Fig. 8.8 Jaccard Distance: distribution by Year for Country, and 1Digit Occupation



(a) JD distribution by 1Digit in 2018-2023



(b) JD distribution by Year and Country

Note: authors' calculations on WIH-OJA data

Conclusion and Future Work

This work began with a central challenge: the static nature of traditional skill taxonomies renders them increasingly inadequate for monitoring the dynamic, fast-evolving European labour market. In response, this work has argued for and demonstrated the value of transforming these frameworks into adaptive, data-driven infrastructures. The unifying thread of this research has been the conviction that enriched taxonomies are not merely classification systems, but essential instruments for generating timely, granular, and policy-relevant labour market intelligence. This concluding chapter synthesizes the primary contributions of the thesis, acknowledges its limitations, and outlines promising avenues for future research that build directly upon the foundations laid herein.

Synthesis of Contributions and Key Findings The research presented in this thesis has produced both methodological innovations for taxonomy enrichment and substantive empirical findings regarding the European labour market. On the methodological front, three key contributions progressively enhanced the ability to keep skill taxonomies like ESCO current using real-world data:

- **Enrichment of ESCO's Green Taxonomy.** The first contribution introduced a data-driven pipeline for enriching the ESCO green skills taxonomy using distributional semantics. Leveraging millions of OJAs, the approach identified and validated new green-related terms, building a "skill pillar" and an "occupation pillar" that quantify green pervasiveness and greenness across occupations. This work provided a large-scale, empirical demonstrations of how unstructured labour market data can be used to enhance official taxonomies.

- **The TAXMAP Framework for Digital Skills.** The second methodological advance was the TAXMAP system, which integrates LLMs with human validation to enrich the ESCO digital skills taxonomy. TAXMAP formalised a hybrid Human–AI framework capable of discovering and mapping emerging digital skills from online data, achieving a balance between automation and interpretability. Its deployment within an EU-funded project demonstrated the feasibility of continuous, high-quality updates to ESCO's digital pillar.
- **The SkiLLens Framework for Emerging Skills.** The third and most comprehensive contribution was the SkiLLens framework, a multilingual pipeline for detecting, validating, and mapping new skill expressions from millions of OJAs across 28 countries. By combining embedding-based extraction, LLM-assisted validation, and large-scale expert feedback, SkiLLens scaled the enrichment process to a pan-European level.

Beyond the methodological contributions, this thesis also offered a substantive application of these enriched taxonomies. Using the data-driven green taxonomy, a large-scale analysis of nearly 29 million European job advertisements provided new evidence on the characteristics of green jobs—highlighting their wage premiums, higher educational requirements, and their diffusion across traditionally non-green occupations. These findings underscore the analytical potential of dynamic taxonomies as instruments for monitoring skill transitions and supporting policy design.

Limitations and Future Directions While the proposed methodologies mark significant progress towards adaptive, data-driven skill intelligence systems, several limitations remain and point to promising directions for future research.

First, data coverage and representativeness remain important challenges. The analyses rely primarily on acOJAs, which tend to overrepresent certain countries, sectors, and higher-skill occupations. Future work should integrate complementary data sources such as professional social networks, education curricula, or administrative datasets to mitigate sampling biases and broaden coverage. Second, although the pipelines demonstrate strong performance, the semantic robustness

of language models remains an open issue. Both static and contextual embeddings can introduce domain or language bias, and LLMs may produce inconsistent or hallucinated outputs when reasoning over complex taxonomic hierarchies. Future research should explore hybrid architectures that couple neural embeddings with symbolic or graph-based reasoning, improving interpretability and stability. Third, while `SkiLLens` successfully extends taxonomy enrichment to a multilingual setting, cross-lingual alignment and low-resource language performance can be further improved. Domain-adaptive fine-tuning and prompt optimization could enhance alignment across linguistic variants and better capture skill expressions in under-represented European languages. Finally, a crucial direction for future work is the institutional integration and sustainability of these systems. Ensuring that models like `TAXMAP` and `SkiLLens` become part of continuous institutional workflows (e.g., within CEDEFOP, EUROSTAT, or national observatories) requires not only technical readiness but also governance frameworks for quality assurance, validation, and ethical use of AI in labour market analytics.

Concluding Remarks

The research presented in this thesis demonstrates that taxonomies—once considered static reference tools—can evolve into dynamic infrastructures to understand labor markets. By progressively combining distributional semantics, LLMs, and multilingual validation, the work shows how AI can transform the way we detect, map, and interpret emerging skills. The progression from the ESCO Green enrichment, through `TAXMAP`, to `SkiLLens` mirrors a broader transition in data science for policy: from isolated analytical experiments to scalable, institutional-grade systems. As labour markets continue to evolve under the forces of digitalisation and the green transition, maintaining adaptive, transparent, and inclusive taxonomies will be key to building the evidence base for future education and employment strategies.

References

[1] Ahmad Alabdulkareem, Morgan Frank, Lijun Sun, Bedoor Alshebli, and Iyad Rahwan. Unpacking the polarization of workplace skills. *Science Advances*, 4: eaao6030, 07 2018. doi: 10.1126/sciadv.aao6030.

[2] Grigoris Antoniou and Frank van Harmelen. *A Semantic Web Primer*. MIT Press, 2nd edition, 2008.

[3] Arthur Apostel and Mikkel Barslund. Measuring and characterising green jobs: A literature review. *Energy Research & Social Science*, 111:103477, 2024. ISSN 2214-6296. doi: 10.1016/j.erss.2024.103477.

[4] Muhammad Arslan and Christophe Cruz. Semantic taxonomy enrichment to improve business text classification for dynamic environments. In *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6, 2022. doi: 10.1109/INISTA55318.2022.9894173.

[5] Deepak Suresh Asudani, Naresh Kumar Nagwani, and Pradeep Singh. Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial Intelligence Review*, pages 1–81, 2023.

[6] G. V. Auktor. Green industrial skills for a sustainable future. https://www.semanticscholar.org/paper/Green-Industrial-Skills-for-a-Sustainable-Future-Auktor/0754730859bf5d4e864d511006f3a4dc86fae4b5, 2021.

[7] David H. Autor. The "task approach" to labor markets : an overview. *Journal for Labour Market Research*, 46(3):185–199, 2013. doi: 10.1007/s12651-013-0128-z.

[8] Bela Balassa. Trade liberalisation and "revealed" comparative advantage. *The Manchester School*, 33(2):99–123, 1965. doi: 10.1111/j.1467-9957.1965.tb00050.x.

[9] N. Barbieri and D. Consoli. Regional diversification and green employment in us metropolitan areas. *Research Policy*, 48(3):693–705, 2019. doi: 10.1016/j.respol.2018.11.001.

[10] Alessandro Barducci, Simone Iannaccone, Valerio La Gatta, Vincenzo Moscato, Giancarlo Sperlì, and Sergio Zavota. An end-to-end framework for information extraction from italian resumes. *Expert Systems with Applications*, 210:118487, 2022.

[11] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

[12] Akshay Bhola, Kishaloy Halder, Animesh Prasad, and Min-Yen Kan. Retrieving skills from job descriptions: A language model based extreme multi-label classification framework. In *Proceedings of the 28th international conference on computational linguistics*, pages 5832–5842, 2020.

[13] John Bluedorn, Niels-Jakob Hansen, Diaa Noureldin, Ippei Shibata, and Marina M. Tavares. Transitioning to a greener labor market: Cross-country evidence from microdata. *Energy Economics*, 126:106836, 2023. ISSN 0140-9883. doi: https://doi.org/10.1016/j.eneco.2023.106836.

[14] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[15] F. Bontadini and F. Vona. Anatomy of green specialization: Evidence from eu production data, 1995-2015. Technical Report 21, Sciences Po Publications, 2020.

[16] Roberto Boselli, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. Using machine learning for labour market intelligence. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2017, Proc. Part III*, volume 10536 of *Lecture Notes in Computer Science*, pages 330–342, Cham, 2017. Springer. doi: 10.1007/978-3-319-71273-4_27.

[17] Roberto Boselli, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. Classifying online job advertisements through machine learning. *Future Generation Computer Systems*, 86:319–328, 2018. ISSN 0167-739X. doi: 10.1016/j.future.2018.03.032.

[18] A. Bowen, K. Kuralbayeva, and E. L. Tipoe. Characterising green employment: The impacts of 'greening' on workforce composition. *Energy Economics*, 72: 263–275, 2018. doi: 10.1016/j.eneco.2018.03.015.

[19] Ronald J. Brachman. What is-a is and isn't: An analysis of taxonomic links in semantic networks. *Computer*, 16(10):30–36, 1983.

[20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[21] Federica Calanca, Luiza Sayfullina, Lara Minkus, Claudia Wagner, and Eric Malmi. Responsible team players wanted: an analysis of soft skill requirements in job advertisements. *EPJ Data Science*, 8(1):1–20, 2019.

[22] M. Capasso, T. Hansen, J. Heiberg, A. Klitkou, and M. Steen. Green growth – a synthesis of scientific findings. *Technological Forecasting and Social Change*, 146: 390–402, 2019. doi: 10.1016/j.techfore.2019.06.013. cited By 154.

[23] Orsetta Causa, Mai Nguyen, and Elena Soldani. Lost in the green transition? measurement and stylized facts. *OECD Economics Department Working Papers*, (1796), 2024. doi: 10.1787/dce1d5fe-en.

[24] Cedefop. Green skills and environmental awareness in vocational education and training. Technical report, Publications Office of the European Union, 2012.

[25] Cedefop. *Online job vacancies and skills analysis – A Cedefop pan-European approach*. Publications Office, 2019. doi: doi/10.2801/097022.

[26] Cedefop. The green employment and skills transformation: insights from a european green deal skills forecast scenario. Publications Office of the European Union. Cedefop policy brief, 2021.

[27] Cedefop. Tracking the green transition in labour markets: Using big data to identify the skills that make jobs greener. Cedefop policy brief, Publications Office of the European Union, 2024. Available at https://www.cedefop.europa.eu.

[28] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and

Noah Fiedel. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435.

[29] G. Cicerone, A. Faggian, S. Montresor, and F. Rentocchini. Regional artificial intelligence and the geography of environmental technologies: does local ai knowledge help regional green-tech specialization? *Regional Studies*, 57(2): 330–343, 2023. doi: 10.1080/00343404.2022.2092610. cited By 45.

[30] Benjamin Clavié and Guillaume Soulié. Large language models as batteries-included zero-shot esco skills matchers. *arXiv preprint arXiv:2307.03539*, 2023.

[31] Emilio Colombo, Fabio Mercorio, and Mario Mezzanzanica. Ai meets labor market: Exploring the link between automation and skills. *Information economics and policy*, 47:27–37, 2019.

[32] Emilio Colombo, Francesco Trentini, and Alessia De Santo. Let's get green: Understanding green skills and jobs through online job advertisements. Working Paper 5137315, Social Science Research Network, 2025. URL https://ssrn.com/abstract=5137315.

[33] Kevin Connolly, Grant Allan, and Stuart McIntyre. The evolution of green jobs in scotland: A hybrid approach. *Energy Policy*, 88(C):355–360, 2016.

[34] Davide Consoli, Giovanni Marin, Alberto Marzucchi, and Francesco Vona. Do green jobs differ from non-green jobs in terms of skills and human capital? *Research Policy*, 45(5):1046–1060, 2016. ISSN 0048-7333. doi: 10.1016/j.respol.2016.02.007.

[35] E. Mark Curtis and Ioana Marinescu. Green energy jobs in the us: What are they, and where are they? Working Paper 30332, National Bureau of Economic Research, August 2022.

[36] E. Mark Curtis, Layla O'Kane, and R. Jisung Park. Workers and the green-energy transition: Evidence from 300 million job transitions. *Environmental and Energy Policy and the Economy*, 5:127–161, 2024. doi: 10.1086/727880.

[37] Alessio D'Amato, Andrea Pronti, Susanna Paleari, Giulio Romaldi, Stefan Speck, Simone Tagliapietra, and Roberto Zoboli. Investment needs and gaps for the sustainability transition in Europe: Rethinking the European Green Deal as an EU industrial strategy. ETC CE Report 2024/8, European Topic Centre on Circular economy and resource use, 10 2024.

[38] Simone D'Amico, Alessia De Santo, Mario Mezzanzanica, and Fabio Mercorio. Taxonomy expansion through collaborative llm mapping. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, SAC '25, page 1961–1968, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400706295. doi: 10.1145/3672608.3707906. URL https://doi.org/10.1145/3672608.3707906.

[39] Andrea De Mauro, Marco Greco, Michele Grimaldi, and Paavo Ritala. Human resources for big data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management*, 54(5):807–817, 2018.

[40] Jens-Joris Decorte, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. Design of negative sampling strategies for distantly supervised skill extraction. In *RecSys in HR*, volume 3218. CEUR, 2022.

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[42] E. C. Dierdorff, J. J. Norton, D. W. Drewes, D. Rivkin, and P. Lewis. Greening of the world of work: Implications for o*net-soc and new and emerging occupations. Technical report, O*NET Resource Center, 2009. Report No. 49.

[43] Robert Elliott, Wenjing Kuai, David Maddison, and Ceren Ozgen. Eco-innovation and (green) employment: A task-based approach to measuring the composition of work in firms. *Journal of Environmental Economics and Management*, 127:103015, 06 2024. doi: 10.1016/j.jeem.2024.103015.

[44] ESCO Publications. Green skills and knowledge concepts: Labelling the esco classification. Technical documentation, January 2022. This document is part of internal ESCO publications designed to enhance stakeholders' understanding, use and development of ESCO.

[45] Chuyu Fang, Chuan Qin, Qi Zhang, Kaichun Yao, Jingshuai Zhang, Hengshu Zhu, Fuzhen Zhuang, and Hui Xiong. Recruitpro: A pretrained language model with skill-aware prompt learning for intelligent recruitment. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3991–4002, 2023.

[46] Silvia Fareri, Nicola Melluso, Filippo Chiarello, and Gualtiero Fantoni. Skillner: Mining and mapping soft skills from any text. *Expert Systems with Applications*, 184:115544, 2021.

[47] Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

[48] Carl Benedikt Frey and Michael A. Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280, 2017. ISSN 0040-1625. doi: 10.1016/j.techfore.2016.08.019.

[49] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, September 2024. doi: 10.1162/coli_a_00524. URL https://aclanthology.org/2024.cl-3.8/.

[50] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655. PMLR, 2018.

[51] Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, and Mario Mezzanzanica. Skills2job: A recommender system that encodes job offer embeddings on graph databases. *Applied Soft Computing*, 101:107049, 2021. ISSN 1568-4946. doi: 10.1016/j.asoc.2020.107049.

[52] Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Seveso. Neo: A system for identifying new emerging occupation from job ads. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 16035–16037, 2021.

[53] Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, and Mario Mezzanzanica. Weta: Automatic taxonomy alignment via word embeddings. *Computers in Industry*, 138:103626, 2022.

[54] Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. Embeddings evaluation using a novel measure of semantic similarity. *Cognitive Computation*, 14(2):749–763, 2022.

[55] Atmika Gorti, Manas Gaur, and Aman Chadha. Unboxing occupational bias: Grounded debiasing of llms with u.s. labor data, 2024. URL https://arxiv.org/abs/2408.11247.

[56] Nidhi Goyal, Jushaan Kalra, Charu Sharma, Raghava Mutharaju, Niharika Sachdeva, and Ponnurangam Kumaraguru. Jobxmlc: Extreme multi-label classification of job skills with graph neural networks. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2181–2191, 2023.

[57] Akshay Gugnani, Vinay Kumar Reddy Kasireddy, and Karthikeyan Ponnalagu. Generating unified candidate skill graph for career path recommendation. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 328–333. IEEE, 2018.

[58] Fatih Gurcan and Nergiz Ercil Cagiltay. Big data software engineering: Analysis of knowledge domains and skill sets using lda-based topic modeling. *IEEE access*, 7:82541–82552, 2019.

[59] Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.

[60] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in NLP models as barriers for persons with disabilities. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.487. URL https://aclanthology.org/2020.acl-main.487/.

[61] Faizan Javed, Phuong Hoang, Thomas Mahoney, and Matt McNair. Large-scale occupational skills normalization for online recruitment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4627–4634, 2017.

[62] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Bin Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[63] Minhao Jiang, Xiangchen Song, Jieyu Zhang, and Jiawei Han. Taxoenrich: Self-supervised taxonomy completion via structure-semantic representations. In *Proceedings of the ACM Web Conference 2022*, pages 925–934, 2022.

[64] Kameni Florentin Flambeau Jiechieu and Norbert Tsopze. Skills prediction based on multi-label resume classification using cnn with model predictions explanation. *Neural Computing and Applications*, 33(10):5069–5087, 2021.

[65] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd edition, 2025. URL https://web.stanford.edu/~jurafsky/slp3/. Online manuscript released August 24, 2025.

[66] Keld Laursen. Revealed comparative advantage and the alternatives as measures of international specialization. *Eurasian Business Review*, 5(1):99–115, 2015. ISSN 1309-4297. doi: 10.1007/s40821-015-0017-1.

[67] Nan Li, Bo Kang, and Tijl De Bie. Skillgpt: a restful api service for skill extraction and standardization using a large language model. *arXiv preprint arXiv:2304.11060*, 2023.

[68] Shan Li, Baoxu Shi, Jaewon Yang, Ji Yan, Shuai Wang, Fei Chen, and Qi He. Deep job understanding at linkedin. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2145–2148, 2020.

[69] Zichen Liu, Hongyuan Xu, Yanlong Wen, Ning Jiang, Haiying Wu, and Xiaojie Yuan. Temp: Taxonomy expansion with dynamic margin loss through taxonomy-paths. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3854–3863, 2021.

[70] Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. *arXiv preprint arXiv:2407.06089*, 2024.

[71] Mingyu Derek Ma, Muhao Chen, Te-Lin Wu, and Nanyun Peng. Hyperexpan: Taxonomy expansion with hyperbolic representation learning. *arXiv preprint arXiv:2109.10500*, 2021.

[72] T. Maczulskij. How are green jobs created? a decomposition analysis. *Economics Letters*, 244, 2024. doi: 10.1016/j.econlet.2024.111950. cited By 1.

[73] Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent systems*, 16(2):72–79, 2001.

[74] Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. Meet-lm: A method for embeddings evaluation for taxonomic data in the labour market. *Computers in Industry*, 124:103341, 2021. ISSN 0166-3615. doi: 10.1016/j.compind.2020.103341.

[75] Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. Taxoref: Embeddings evaluation for ai-driven taxonomy refinement. In

Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III, page 612–627, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-86522-1. doi: 10.1007/978-3-030-86523-8_37. URL https://doi.org/10.1007/978-3-030-86523-8_37.

[76] Joana Elisa Maldonado, Anneleen Vandeplas, Istvan Vanyolos, Mauro Vigani, and Alessandro Turrini. Assessing Green Job Dynamics in the EU A Comparison of Alternative Methodologies. European Economy - Discussion Papers 206, Directorate General Economic and Financial Affairs (DG ECFIN), European Commission, July 2024.

[77] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, USA, 2008. ISBN 0521865719.

[78] Yuning Mao, Tong Zhao, Andrey Kan, Chenwei Zhang, Xin Luna Dong, Christos Faloutsos, and Jiawei Han. Octet: Online catalog taxonomy enrichment with self-supervision. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2247–2257, 2020.

[79] Giovanni Marin and Francesco Vona. Climate policies and skill-biased employment dynamics: Evidence from eu countries. *Journal of Environmental Economics and Management*, 98:102253, 2019. ISSN 0095-0696. doi: https://doi.org/10.1016/j.jeem.2019.102253. URL https://www.sciencedirect.com/science/article/pii/S0095069618304911.

[80] Michal Měchura. A taxonomy of bias-causing ambiguities in machine translation. In Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen, editors, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.gebnlp-1.18. URL https://aclanthology.org/2022.gebnlp-1.18/.

[81] Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1699–1710, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594109. URL https://doi.org/10.1145/3593013.3594109.

[82] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[83] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021.

[84] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2025. URL https://arxiv.org/abs/2402.06196.

[85] Jacob Mincer. Investment in human capital and personal income distribution. *Journal of Political Economy*, 66(4):281–302, 1958.

[86] Raymond J Mooney and Razvan Bunescu. Mining knowledge from text using information extraction. *ACM SIGKDD explorations newsletter*, 7(1):3–10, 2005.

[87] R. Moreno and D. Ocampo-Corrales. The ability of european regions to diversify in renewable energies: The role of technological relatedness. *Research Policy*, 51(5), 2022. doi: 10.1016/j.respol.2022.104508. cited By 28.

[88] Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. Large language models for creation, enrichment and evaluation of taxonomic graphs. *Semantic Web Journal*, 2024.

[89] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the*

European Chapter of the Association for Computational Linguistics, pages 2014–2037, 2023.

[90] Joanna Napierala, Vladimir Kvetan, Jiri Branka, and Cedefop. Assessing the representativeness of online job advertisements. *Publications Office of the European Union*, 2022. doi: 10.2801/807500.

[91] Lionel Nesta, Francesco Vona, and Francesco Nicolli. Environmental policies, competition and innovation in renewable energy. *Journal of Environmental Economics and Management*, 67(3):396–411, 2014. ISSN 0095-0696. doi: https://doi.org/10.1016/j.jeem.2014.01.001.

[92] Khanh Nguyen, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. Rethinking skill extraction in the job market domain using large language models. In Estevam Hruschka, Thom Lake, Naoki Otani, and Tom Mitchell, editors, *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 27–42, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.nlp4hr-1.3/.

[93] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30:6338–6347, 2017.

[94] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pages 3779–3788. PMLR, 2018.

[95] Irina Nikishina, Varvara Logacheva, Alexander Panchenko, and Natalia Loukachevitch. Studying taxonomy enrichment on diachronic WordNet versions. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3095–3106, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.276. URL https://aclanthology.org/2020.coling-main.276.

[96] OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[97] Susanna Paleari. The impact of the european green deal on eu environmental policy. *The Journal of Environment & Development*, 31(2):196–220, 2022. doi: 10.1177/10704965221082222.

[98] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[99] David J. Peters. Understanding green occupations from a task-based approach. *Applied Economic Perspectives and Policy*, 36(2):238 – 264, 2014. doi: 10.1093/aepp/ppt026.

[100] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. in naacl, 2018.

[101] Bornali Phukon, Anasua Mitra, Ranbir Sanasam, and Priyankoo Sarmah. Team: A multitask learning based taxonomy expansion approach for attach and merge. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 366–378, 2022.

[102] Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.

[103] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training, 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

[104] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[105] Andrés Rodríguez-Pose and Federico Bartalucci. The green transition and its potential territorial discontents. *Cambridge Journal of Regions, Economy and Society*, 17(2):339–358, 11 2023. ISSN 1752-1378. doi: 10.1093/cjres/rsad039.

[106] A. Santoalha, D. Consoli, and F. Castellacci. Digital skills, relatedness and green diversification: A study of european regions. *Research Policy*, 50(9), 2021. doi: 10.1016/j.respol.2021.104340. cited By 83.

[107] Nathalie Scholl, Sébastien Turban, and Peter N. Gal. The green side of productivity: An international classification of green and brown occupations. OECD Productivity Working Papers 33, OECD Publishing, May 2023.

[108] Jiaming Shen and Jiawei Han. *Taxonomy Enrichment*, pages 49–81. Springer International Publishing, Cham, 2022. ISBN 978-3-031-11405-2. doi: 10.1007/978-3-031-11405-2_4. URL https://doi.org/10.1007/978-3-031-11405-2_4.

[109] Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. Taxoexpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *Proceedings of The Web Conference 2020*. ACM, April 2020. doi: 10.1145/3366423.3380132. URL http://dx.doi.org/10.1145/3366423.3380132.

[110] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.330. URL https://aclanthology.org/2021.acl-long.330/.

[111] A. Sulich, M. Rutkowska, and Ł. Popławski. Green jobs, definitional issues, and the employment of young people: An analysis of three european union countries. *Journal of Environmental Management*, 262:110314, 2020. ISSN 0301-4797. doi: https://doi.org/10.1016/j.jenvman.2020.110314.

[112] Andrzej Sulich and Laura Soloducho-Pelc. The circular economy and green jobs creation. *Environmental Science and Pollution Research*, 29(19):14231–14247, 2022. doi: 10.1007/s11356-021-18007-3.

[113] Kunihiro Takeoka, Kosuke Akimoto, and Masafumi Oyamada. Low-resource taxonomy enrichment with pretrained language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2747–2758, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.217. URL https://aclanthology.org/2021.emnlp-main.217.

[114] Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Kai Hua, Wenpeng Hu, Zhengwei Tao, and Shuai Ma. Llms are also effective embedding models: An in-depth overview, 2025. URL https://arxiv.org/abs/2412.12591.

[115] François Torregrossa, Robin Allesiardo, Vincent Claveau, Nihel Kooli, and Guillaume Gravier. A survey on training and evaluation of word embeddings. *International Journal of Data Science and Analytics*, 11:85–103, 2021.

[116] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. URL https://api.semanticscholar.org/CorpusID:257219404.

[117] Alexandra Tsvetkova, Elettra D'Amico, Alexander Lembcke, Polina Knutsson, and Wessel Vermeulen. How well do online job postings match national sources in large English speaking countries?: Benchmarking Lightcast data against statistical sources across regions, sectors and occupations. *OECD Publishing*, (2024/01), March 2024. doi: 10.1787/c17cae09-en. URL https://ideas.repec.org/p/oec/cfeaaa/2024-01-en.html.

[118] A. Valero, J. Li, S. Muller, C. Riom, V. Nguyen-Tien, and M. Draca. Are 'green' jobs good jobs? how lessons from the experience to-date can inform labour market transitions of the future. Technical report, Grantham Research

Institute on Climate Change and the Environment and Centre for Economic Performance, 2021.

[119] Alje van Dam, Andres Gomez-Lievano, Frank Neffke, and Koen Frenken. An information-theoretic approach to the analysis of location and colocation patterns. *Journal of Regional Science*, 63(1):173–213, January 2023. doi: 10.1111/jors.12621.

[120] Anneleen Vandeplas, Istvan Vanyolos, Mauro Vigani, and Lukas Vogel. The Possible Implications of the Green Transition for the EU Labour Market. European Economy - Discussion Papers 176, Directorate General Economic and Financial Affairs (DG ECFIN), European Commission, December 2022.

[121] Nikhita Vedula, Patrick K. Nicholson, Deepak Ajwani, Sourav Dutta, Alessandra Sala, and Srinivasan Parthasarathy. Enriching taxonomies with functional domain knowledge. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 745–754, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210000. URL https://doi.org/10.1145/3209978.3210000.

[122] Wessel Vermeulen and Fernanda Gutierrez Amaros. How well do online job postings match national sources in european countries? Technical report, OECD Publishing, 2024. URL https://www.oecd-ilibrary.org/content/paper/e1026d81-en.

[123] Mikhail Vinel, Ivan Ryazanov, Dmitriy Botov, and Ivan Nikolaev. Experimental comparison of unsupervised approaches in the task of separating specializations within professions in job vacancies. In *Artificial Intelligence and Natural Language (AINL 2019)*, volume 1119 of *Communications in Computer and Information Science*, pages 99–112. Springer, Cham, 2019. doi: 10.1007/978-3-030-34518-1_7.

[124] Francesco Vona. Labour Markets and the Green Transition: a practitioner's guide to the task-based approach. JRC Research Reports JRC126681, Joint Research Centre, October 2021. URL https://ideas.repec.org/p/ipt/iptwpa/jrc126681.html.

[125] Francesco Vona, Giovanni Marin, Davide Consoli, and David Popp. Environmental regulation and green skills: An empirical exploration. *Journal of the Association of Environmental and Resource Economists*, 5(4):713–753, 2018. doi: 10.1086/698859.

[126] Francesco Vona, Giovanni Marin, and Davide Consoli. Measures, drivers and effects of green employment: Evidence from us local labor markets, 2006-2014. *Journal of Economic Geography*, 19(5):1021 – 1048, 2019. doi: 10.1093/jeg/lby038.

[127] Chengyu Wang, Xiaofeng He, and Aoying Zhou. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1190–1203, 2017.

[128] Jakob Mørup Wang and Zhiru Sun. Llm-supervised multilingual skill extraction and classification from job ads. In Ryutaro Ichise, editor, *Natural Language Processing and Information Systems*, pages 94–104, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-031-97144-0.

[129] Jingjing Wang, Changsung Kang, Yi Chang, and Jiawei Han. A hierarchical dirichlet model for taxonomy expansion for search engines. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, page 961–970, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327442. doi: 10.1145/2566486.2568037. URL https://doi.org/10.1145/2566486.2568037.

[130] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

[131] Suyuchen Wang, Ruihui Zhao, Xi Chen, Yefeng Zheng, and Bang Liu. Enquire one's parent and child before decision: Fully exploit hierarchical structure for self-supervised taxonomy expansion. In *Proceedings of the Web Conference 2021*, pages 3291–3304, 2021.

[132] Hongyuan Xu, Yunong Chen, Zichen Liu, Yanlong Wen, and Xiaojie Yuan. Taxoprompt: A prompt-based generation method with taxonomic context for self-supervised taxonomy expansion. In *IJCAI*, pages 4432–4438, 2022.

[133] Wenli Yang, Lilian Some, Michael Bain, and Byeong Kang. A comprehensive survey on integrating large language models with knowledge-based methods. *Knowledge-Based Systems*, 318:113503, 2025. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2025.113503. URL https://www.sciencedirect.com/science/article/pii/S0950705125005490.

[134] Wenli Yang, Lilian Some, Michael Bain, and Byeong Kang. A comprehensive survey on integrating large language models with knowledge-based methods. *Knowledge-Based Systems*, 318:113503, 2025. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2025.113503. URL https://www.sciencedirect.com/science/article/pii/S0950705125005490.

[135] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

[136] Yue Yu, Yinghao Li, Jiaming Shen, Haoyang Feng, Jimeng Sun, and Chao Zhang. Steam: Self-supervised taxonomy expansion with mini-paths. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020. URL https://api.semanticscholar.org/CorpusID:219792304.

[137] Qingkai Zeng, Jinfeng Lin, Wenhao Yu, Jane Cleland-Huang, and Meng Jiang. Enhancing taxonomy completion with concept generation via fusing relational representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2104–2113, 2021.

[138] Denghui Zhang, Junming Liu, Hengshu Zhu, Yanchi Liu, Lichen Wang, Pengyang Wang, and Hui Xiong. Job2vec: Job title benchmarking with collective multi-view representation learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*,

Beijing, China, 2019. ACM. doi: 10.1145/3357384.3357825. URL https://doi.org/10.1145/3357384.3357825.

[139] Jieyu Zhang, Xiangchen Song, Ying Zeng, Jiaze Chen, Jiaming Shen, Yuning Mao, and Lei Li. Taxonomy completion via triplet matching network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4662–4670, 2021.

[140] Mike Zhang, Kristian Jensen, Sif Sonniks, and Barbara Plank. Skillspan: Hard and soft skill extraction from english job postings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4962–4984, 2022.

[141] Mike Zhang, Kristian Nørgaard Jensen, and Barbara Plank. Kompetencer: Fine-grained skill classification in danish job postings via distant supervision and transfer learning. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 436–447, 2022.

[142] Mike Zhang, Rob van der Goot, and Barbara Plank. Escoxlm-r: Multilingual taxonomy-driven pre-training for the job market domain. In *The 61st Annual Meeting of the Association for Computational Linguistics*, pages 11871–11890. Association for Computational Linguistics, 2023.

[143] Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2032–2042, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714940. URL https://doi.org/10.1145/3696410.3714940.

[144] Tinghui Zhu, Jingping Liu, Jiaqing Liang, Haiyun Jiang, Yanghua Xiao, Zongyu Wang, Rui Xie, and Yunsen Xian. Towards visual taxonomy expansion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6481–6490, 2023.

[145] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books, 2016.