I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17286

Identification of an Expanded Inventory of Green Job Titles through AI-Driven Text Mining

Michał Paliński
Güneş Aşık
Tomasz Gajderowicz
Maciej Jakubowski
Efşan Nas Özen
Dhushyanth Raju

SEPTEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17286

Identification of an Expanded Inventory of Green Job Titles through AI-Driven Text Mining

Michał Paliński
University of Warsaw

Güneş Aşık
TOBB University of Economics and Technology, Ankara

Tomasz Gajderowicz
University of Warsaw

Maciej Jakubowski
University of Warsaw

Efşan Nas Özen
World Bank, Ankara

Dhushyanth Raju
World Bank, Washington and IZA

SEPTEMBER 2024

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Identification of an Expanded Inventory of Green Job Titles through AI-Driven Text Mining*

This study expands the inventory of green job titles by incorporating a global perspective and using contemporary sources. It leverages natural language processing, specifically a retrieval-augmented generation model, to identify green job titles. The process began with a search of academic literature published after 2008 using the official APIs of Scopus and Web of Science. The search yielded 1,067 articles, from which 695 unique potential green job titles were identified. The retrieval-augmented generation model used the advanced text analysis capabilities of Generative Pre-trained Transformer 4, providing a reproducible method to categorize jobs within various green economy sectors. The research clustered these job titles into 25 distinct sectors. This categorization aligns closely with established frameworks, such as the U.S. Department of Labor's Occupational Information Network, and suggests potential new categories like green human resources. The findings demonstrate the efficacy of advanced natural language processing models in identifying emerging green job roles, contributing significantly to the ongoing discourse on the green economy transition.

JEL Classification: J23, Q52, O14

Keywords: AI, text mining, occupational classification, green jobs, green economy

Corresponding author:
Efşan Nas Özen
World Bank, Ankara
Büyükesat, Uğur Mumcu Cd.
No:88, 06700 Çankaya/Ankara
Turkey
E-mail: snasozen@worldbank.org

* We thank Berfu Çopur for research assistance with the literature review. We are also grateful to Burak Baskın, Paolo Belli, Ahmet Kurnaz, René León Solano, and Aivin Vicquierra Solatorio for useful comments.

1. Introduction

The escalating impacts of climate change underscore the urgency of a green transition—a pivotal shift toward sustainable practices that is essential for our planet's future. This transition is expected to accelerate rapidly, necessitating that policy makers analyze its impacts on national labor markets and develop effective strategies to navigate the evolving landscape. Understanding the scope and nature of green jobs is crucial for informing public policy, enabling governments and organizations to develop tailored and targeted strategies for education, training, and employment to support a sustainable economy.

Worldwide, the most widely used source of green job titles is the Green Occupations list, constructed by the U.S. Department of Labor's Occupational Information Network (O*NET) in 2009 (Dierdorff et al. 2009). O*NET's original approach involved reviewing publications covering a wide array of workplace topics pertinent to the green economy. In assessing green jobs, research predominantly employs two methods: top-down approaches, which categorize entire sectors or industries as green, and bottom-up approaches, which focus on specific occupations, defining green jobs based on the green nature of the tasks or skills associated with those roles (Valero et al. 2021). O*NET's classification is the most often used source for occupation retrieval in the bottom-up approach to green jobs analysis (OECD 2023).

The green job taxonomy developed by O*NET has been instrumental in shaping quantitative research on the green economy. In the United States, its impact is reflected in studies by researchers such as Consoli et al. (2016), Popp et al. (2020), and Vona, Marin, and Consoli (2019) and Vona et al. (2018). The taxonomy has also been adapted for various regions, including the European Union (Bowen and Hancké 2019), the Netherlands (Elliott et al. 2021), the United Kingdom (Valero et al. 2021), Organisation for Economic Cooperation and Development (OECD) member countries (OECD 2023), Viet Nam (Doan et al. 2023), and Argentina (de la Vega, Porto, and Cerimelo 2024). After classifying occupations as green, studies delve into the specific skills and tasks required for green jobs, analyze trends in green job creation and distribution, and assess the broader economic impacts, such as productivity, innovation, and growth, associated with the green transition.

However, two main issues make O*NET less relevant for worldwide use, particularly for green jobs. First, O*NET was built in 2009, with the last major revision of the taxonomy completed in 2011 (Dierdorff et al. 2011) and the associated reference book last updated in 2013 (O*NET 2013). The literature on green jobs has expanded significantly since 2009. Second, O*NET is designed for the U.S. labor market, identifying tasks within occupations based on the U.S. context. The tasks and skills required to perform these jobs depend on the production technology, which can differ significantly between the United States and other economies, such as low- and middle-income countries.

Our study aims to expand the inventory of green job titles by integrating a global perspective and incorporating contemporary sources. Our literature review comprised a search for articles published after 2008 using Scopus and Web of Science—two leading bibliographic databases that are widely used by the academic community for accessing an extensive global collection of peer-reviewed publications across various disciplines (Zhu and Liu 2020). The year 2008 marked a

critical juncture in the dialogue on green jobs, with the first explicit definition of the concept (Stanef-Puică et al. 2022).

The construction of a taxonomy like O*NET typically involves qualitative coding to identify job titles within a green context, a method that is labor-intensive and time-consuming. However, there is a growing trend toward using natural language processing (NLP), often augmented by expert review, as a potent tool for job identification and categorization across various contexts, including the green economy (Chiarello et al. 2021; Decorte et al. 2021; Li, Sun et al. 2020; Papoutsoglou et al. 2022). A significant example is the 2022 initiative by the European Commission, which employed the Bidirectional Encoder Representations from Transformers (BERT) NLP algorithm alongside manual labeling to identify green concepts (skills and knowledge) within the European Skills, Competences, Qualifications, and Occupations classification (EC 2022).

Aligned with this NLP-driven methodological evolution, our research utilized an advanced artificial intelligence (AI) pipeline, specifically the retrieval-augmented generation (RAG) model (Lewis et al. 2020), to identify green job titles in academic literature. This technology enabled the examination of a substantially larger set of literature than manual methods could accommodate. RAG stands out as an effective NLP approach, merging the benefits of retrieval- and generative-based AI models, thereby addressing prevalent issues in basic generative AI, such as hallucinations and the lack of domain-specific knowledge (Gao et al. 2023). Importantly, our approach is reproducible, allowing the list of green jobs to be updated as the literature on the green transition expands in the future.

Our search of the academic literature published between January 2009 and April 2024, when we conducted the search, ultimately yielded 1,067 articles for analysis. We found that the academic literature on the green transition has significantly expanded over the past 15 years, both in the number of articles and in the diversity of represented countries and regions. In 2009, there were only 44 articles on the green transition. By 2023, this number had increased to 162. In 2009, articles almost exclusively covered the United States, Canada, China, and countries in the European Union. By 2023, the coverage had expanded to include Europe, the Caucasus, Southeast Asia, and Africa.

We identified 695 unique green job titles from 105 articles (10 percent of the 1,067 articles). Comparing our list of green jobs with those identified in O*NET, we found that 17 percent of the job titles matched perfectly or almost perfectly with O*NET, while we also identified potentially new titles through less precise matches.

Our study demonstrates that AI-based models can address capacity challenges in identifying qualitative information from a large and expanding body of literature, despite some limitations. Future research and practice should focus on refining these AI-driven methods and integrating additional information sources to continuously update and expand the inventory of green job titles as the literature on the green transition evolves.

2. Approach

Identification of Relevant Literature

In April 2024, we conducted a search of the literature published since January 2009, using the official application programming interfaces of Scopus and Web of Science. Our search strategy involved keyword combinations previously validated in three systematic literature reviews related to green jobs (Apostel and Barslund 2024; Kozar and Sulich 2023; Stanef-Puică et al. 2022). The keyword combinations included "green job(s)," "green occupation(s)," "green employment," "sustainable job(s)," "sustainable occupation(s)," "green transition job(s)," and "green-collar job(s)." (Box 1 lists the search queries.) These keywords were searched within titles, abstracts, author keywords, and "topics" as referenced in the databases. To ensure credible results, we restricted our focus to peer-reviewed publications, specifically articles and reviews (hereafter referred to as "articles").

Our search approach differs significantly from that of O*NET. While O*NET meticulously indexed and categorized sources within its reference book, the specifics of its selection process are sparingly described. There is no detailed information on specific keywords or methods used in gathering the articles. It involved collecting and reviewing more than 60 publications, including academic journals, commissioned reports, industry white papers, and government technical reports. Additionally, O*NET conducted a substantial review of various internet sources related to the green sector's workforce (O*NET 2013).

Identification of Green Job Titles in the Literature

We used the RAG model to simulate the work traditionally performed by research assistants who would manually tag green job titles within articles. This manual tagging process, extending over 1,000 pages across the articles in the analysis set, is resource-intensive and susceptible to errors from human oversight, cognitive biases, and heuristic shortcuts. In contrast, the RAG model offers a robust and consistent approach.

A significant advantage of using the RAG model is the reproducibility of the results. By utilizing seed parameters available in the OpenAI models, specifically the Generative Pre-trained Transformer 4 (GPT-4)-0125-preview model, we ensured that our results were replicable, providing a degree of consistency that manual tagging struggles to achieve. Although the models cannot be entirely deterministic due to their inherent stochastic nature, the use of seed parameters helps to ensure that the results are highly consistent across multiple runs (Anadkat 2023). Furthermore, the advanced natural language understanding capabilities of the GPT-4 model enabled a nuanced analysis of the context in which job titles are discussed in the articles. This is particularly vital in our analysis set, where green and nongreen jobs are often mentioned in the same articles. The model's ability to discern the context and classify job titles accordingly is a substantial improvement over older NLP approaches, such as less capable embedding models like BERT or fully supervised methods like named entity recognition (NER), which might not capture such subtleties or nuances.

Employing the RAG model, we used embedding models to identify relevant sections within articles (chunks) that discussed specific job titles. We used OpenAI's most capable text-embedding-3-large model with 3,072 dimensions in the embedding process. While chunking is often employed to circumvent the context window limitations of certain models, our application of GPT-4, which boasts an expansive context window of 128,000 tokens (comparable to 96,000 words), was not hindered by such constraints. Instead, the decision to chunk text in our analysis was dictated by the fact that chunking significantly improves the relevance of retrieved content as it decreases noise in the embedded text (Yepes et al. 2024). Next, we employed the GPT-4 model to scrutinize segments of articles where job titles were mentioned, aiming to infer from the context whether the authors classified these roles as examples of green jobs. Aware of the several competing definitions of "green jobs" in the academic literature (Stanef-Puică et al. 2022), we refrained from adhering to any singular definition. Instead, we directed the model to determine if the authors considered that these jobs were green, such as whether they were discussed within the realms of the green economy, sustainability, or climate change mitigation. We purposefully did not expose the model to any preestablished classifications of green jobs to prevent priming effects and promote an unbiased evaluation based on context. We present a more detailed description of the RAG model's pipeline stages in the appendix.

The generative capabilities of the AI were specifically harnessed in the final stage of the RAG model implementation process (figure 1). While the AI possesses extensive knowledge from its training, our model strategically refrains from using this knowledge. The model's generative functions are not employed to introduce or infer information from its training but rather to interpret and analyze the text that is presented to it. When the model identifies potential sections of the text that might discuss green jobs, it leverages its natural language understanding capabilities to analyze the given text. The goal is to ascertain whether the authors of the articles are indeed mentioning specific job titles and if these titles are discussed within the green context.

The model we used has commonalities with NER, a process in NLP that involves identifying and categorizing key information (entities) in text (Li, Shi et al. 2020). Entities could be names of people, companies, locations, and so forth. Our work parallels this approach by identifying green job titles within text. Illustrating the progression in NLP, research has shown that even the older GPT-3 model could match the performance of fully supervised NER baselines (Wang et al. 2023). Zhou et al. (2023) demonstrate that the Large Language Model Meta AI, a large language model (LLM), significantly outperforms supervised NER models, as evidenced by a substantial margin in the F1 score, a measure of a test's accuracy. This comparison spanned 43 data sets encompassing nine varied domains. Similarly, Monajatipoor et al. (2024) demonstrate that in the biomedical field, GPT-4 outperforms traditional NER models.

Empirical studies comparing GPT models to earlier text mining methods, such as BERT, remain limited. Compared to fine-tuned BERT models, GPT-3 has exhibited superior performance in text classification tasks in related contexts (Liga and Robaldo 2023; Pawar and Makwana 2022). However, GPT-4, when employed in a zero-shot setting, significantly outperformed the base BERT model but was outperformed by fine-tuned BERT models in specific tasks such as protein sequence identification (Rehana et al. 2023). Despite these findings, no studies have been identified that directly compare the performance of these models in a context similar to ours, where accurate classification is highly contingent on the surrounding context, such as distinguishing

between green and nongreen jobs. We posit that GPT might outperform BERT in this context because its more complex architecture, larger number of parameters, and ability to handle longer context lengths likely enable it to better differentiate nuanced, context-dependent information, such as classifying jobs as green.

A key feature needed in such exercises is validating the output of the model. In fields like biology, medicine, law, programming, or finance, standardized benchmarks exist to measure the efficacy of LLMs as NER tools (Zhou et al. 2023). However, for our purposes, such benchmarks are unavailable. To ensure the validity of our results, we undertook two types of checks. First, we specifically focused on articles for which the model did not identify any green job titles to check for false negatives. This situation is relatively common since authors might discuss green sectors of the economy without explicitly mentioning job titles. To this end, we randomly selected a sample of such articles to review manually, ensuring that the absence of identified green job titles was consistent with the content of the articles. Second, we conducted a review of all the articles in which the model identified job titles. This step was to check for false positives—that is, erroneously classifying nongreen job titles as green—and to detect any instances of hallucinations where the model might generate nonexistent job titles.

To gain a better understanding of the context in which green jobs are analyzed, we mapped the articles mentioning green job titles to economic activities, based on the International Standard Industrial Classification of All Economic Activities (ISIC) classification scheme. ISIC is a United Nations system for classifying economic data according to industry. For this mapping, we provided ChatGPT-3.5 with the ISIC classification, including descriptions of all activities, and prompted it to find the best top-level matches for all articles.

We also identified the geographical coverage of all the retrieved articles. For this, we prompted ChatGPT-3.5 to retrieve countries mentioned as the basis for analysis in the abstracts and titles of all the articles. If no countries were mentioned, we assumed that the article had a global perspective. Next, we used Python's pycountry package (Theune 2024) for fuzzy matching of the country names with official ISO country codes and identified the continents of the countries. This approach allowed us to illustrate the global landscape of green jobs research.

Matching of Identified Green Job Titles with O*NET

We employed embedding modeling to represent both the job titles we identified and those from O*NET as 3,072-dimensional vectors, enabling a systematic comparison. For this task, we used the text-embedding-3-large model. By utilizing cosine similarity, a recommended distance measure for this model, we identified the closest matches between our identified job titles and those in O*NET. Cases where the job titles showed only minimal similarity indicated potential new green job titles that were not yet recognized in O*NET. The matching process for this step presented a significant challenge for the model because it operated with minimal context that included only the job titles themselves. Had we been able to utilize detailed tasks and skills relevant to these jobs alongside the job descriptions from O*NET, we could have achieved a more informed and accurate matching process. However, the nature of the articles typically does not lend itself to a systematic discussion of job roles, including specific tasks and skills.

We also mapped the green job titles into major green economy sectors through clustering based on their semantic similarity. We used job title embeddings and applied Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) (McInnes, Healy, and Melville 2018), followed by Hierarchical Density-Based Spatial Clustering (HDBSCAN) (McInnes, Healy, and Astels 2017). Fine-tuning these techniques was essential to achieve meaningful results.

We configured UMAP with 10 neighbors to balance local and global structure, and a minimum distance of 0.1 to control the density of point packing, ensuring that local detail was preserved. This configuration maintains similarity among nearby points (local structure) while grouping clusters of similar points together (global structure).

For clustering, we used HDBSCAN with a minimum cluster size of 10 to ensure significance and a minimum sample size of four to define the number of points required to form a dense region.

3. Results

Green Literature

Our search yielded a total of 1,367 articles, with Scopus contributing 991 articles and Web of Science contributing 376 articles (table 1). We used the Digital Object Identifier and International Standard Serial Number to cross-check the uniqueness of the articles across the two databases. We found that 88 percent of the articles indexed in Web of Science were also indexed in Scopus. Consequently, we integrated the unique articles from Web of Science—those not found in Scopus—to arrive at our unique "analysis set" of 1,067 articles. In our ensuing analysis, we used the full texts of 567 articles for which we were able to retrieve Portable Document Format (PDF) files and the abstracts for the remaining 500 articles. While most of the publications are articles, with 915 from Scopus and 353 from Web of Science, we retrieved a diverse set of publications, including conference proceedings, book chapters, and other materials (table 2).

Both the number and geographical spread of the articles on green literature have expanded since 2009. In 2009, there were 44 articles, and in 2023, there were 162 articles (figure 2). Moreover, the articles in 2009 almost exclusively covered North America (table 3). However, the scope gradually diversified. Notably, by 2015, the number of articles covering European countries had surpassed those covering North American countries in cumulative terms; by 2022, the same had occurred for Asian countries.

In the early years of our analysis period, there were no articles focusing specifically on green jobs in South America and hardly any in Africa or Oceania. This situation has changed dramatically, with substantial increases in the number of articles covering these continents over time. This trend indicates that research on green jobs is becoming increasingly globalized, encompassing a broader range of geographical settings.

Green Job Titles

We initially identified 799 potential green job titles using the RAG pipeline (see file in GitHub for the list of titles). We excluded 66 titles, as they were not job titles but referred to green activities.

For example, in the study by Afolabi et al. (2018), while the model correctly identified "environmental compliance specialist" as a green job title, it also incorrectly tagged activities like "solar panel manufacturing" or "reduction of water usage on-site" as job titles. The following is the direct citation from which the model inferred these titles: "The data revealed areas that are peculiar to the provision of green jobs in the construction sector such as solar panel manufacturing (...)" (Afolabi et al. 2018, 2). It seems that the model erroneously assumed that the phrase "green jobs in the construction sector such as (...)" was introducing a list of job titles. This misinterpretation illustrates the types of heuristics the model might employ and emphasizes the importance of quality checks while working with LLMs in their current state of development. Next, we excluded 21 titles because they were too broad, such as "technicians" and "engineers." This indicates that despite instructing the model in our prompts to return only specific job titles, roles, or occupations, it erroneously produced nonspecific results in a limited number of instances. Finally, we standardized all job titles to their singular forms and removed duplicates. This process eliminated an additional 17 job titles, resulting in a total of 695 unique green job titles.

Our analysis revealed that the 695 job titles appear infrequently across the 1,067 articles, with just 105 articles (10 percent of the total analysis set) mentioning one or more green job titles. Additionally, job titles were more frequently identified in full texts, with only 19 job titles found in abstracts.

The global perspective is significant, with 28 articles referencing green job titles internationally (table 4). The United States follows closely, with mentions in 20 articles with green job titles. The European Union is well-represented with 13 mentions. Other notable countries include Brazil, China, and Spain, each with 6 mentions. In total, we found 40 countries mentioned in the articles with green job titles, although 19 countries were mentioned only once.

Over one-third of the job titles are in engineering, another one-fifth are technician-level jobs, and a significant share includes jobs in business and administration, such as policy specialists. Several other titles are in the areas of building and construction, with roles varying in terms of implied skill requirements. For example, middle-skilled job titles, such as roofers and insulation workers, constitute about 5 percent of the titles, while low-skilled construction workers are mentioned nearly as frequently in the context of green buildings.

The articles with green job titles often mention specific activities that are relevant for these jobs. Based on ISIC codes, our findings show that the public administration and defense sectors dominate, with 24 mentions, highlighting the vital role of government initiatives in promoting green jobs (table 5).[1] The manufacturing sector follows with 17 mentions, reflecting its transition toward more sustainable practices. Human health and social work activities are noted 14 times, tied to discussions about transitioning to a low-carbon economy and society in the wake of the coronavirus pandemic (Strachan, Greig, and Jones 2022). Water supply, sewerage, and waste management are referenced 12 times, emphasizing the importance of environmental management in these sectors. Electricity, gas, steam, and air conditioning supply are mentioned 11 times,

[1] For example, in the Namibian context, Wijesinghe and Thorn (2021) emphasize the critical role of municipal planners in developing urban green infrastructure in the country's capital, Windhoek, particularly in response to escalating climate change risks. Similarly, Davis (2013) discusses the necessity of establishing permanent positions for public sector compliance officers within Queensland, Australia's illegal dumping task force, to enhance sustainability.

aligning with the shift toward renewable energy sources. Other sectors, including professional, scientific, and technical activities; agriculture, forestry, and fishing; and transportation and storage, also figure prominently in our findings, illustrating the broad spectrum of economic activities impacted by the green transition.

Comparison of Identified Green Job Titles with O*NET

Our approach to identifying green jobs uses a broader definition, relying on how the authors of the articles define green jobs. The classification of many jobs as "green" can be contentious, as it is debatable when a job is truly green. Many roles are not inherently green but are in high demand within the green sector. For example, Nhamo (2010, 6) states the following:

> The greatest percentage of jobs created in renewable energy and energy efficiency are (…) conventional jobs for accountants, engineers, computer analysts, clerks, factory workers, truck drivers, and mechanics, many of whom may not even realize that they owe their livelihood to renewable energy and energy efficiency.

Our model identifies these jobs as green, mirroring the categorization that O*NET uses in the "green increased demand" category. This interpretation broadens the definition of "green" jobs to include conventional roles that are essential for supporting the green economy.

In some cases, jobs can be classified as green or nongreen depending on the production outcomes associated with the workers' roles. For instance, Wandzich and Plaza (2017) identify iron and steel workers, sheet metal workers, and welders as "green jobs associated with green activity areas," particularly in sectors like wind power generation and freight rail. Our model identifies these jobs as green as it consistently aims to align with the authors' interpretations of what constitutes a green job. Notably, the first two occupations—iron and steel workers and sheet metal workers—are also included in O*NET's list of green jobs.

One of the most debated cases involves the classification of laborers in mining. Despite initial doubts, this role was indeed cited as a green job category.[2] Notably, O*NET also includes two general occupations related to mining. This case emphasizes the significance of analyzing green jobs at a more granular level of disaggregation when possible. Such detailed scrutiny is necessary to differentiate, for example, workers engaged in environmentally sustainable mining practices from those in conventional mining.

The specific categorization of green job titles, positions, and roles into established occupations and the proposal of new occupations when no existing classification fits are beyond the scope of this study. Such detailed work requires alignment with the policies of the classification system being used to ensure relevance and accuracy. However, we provide preliminary work that compares our results with O*NET. Specifically, this comparison aims to assess whether we were able to identify novel green jobs that are not currently recognized by O*NET based on our review of the recent academic literature.

[2] See de la Vega, Patricio, Porto, and Cerimelo (2024), who label this group as potential green workers under an International Standard Classification of Occupations at the 2-digit level.

Figure 3 presents a density plot of the distribution of cosine similarity scores, representing the degree of similarity between the job titles identified by our model and their closest matches within O*NET. The scores range from 0.2 to 1.0, with 1.0 indicating perfect similarity. From the distribution, we observe two peaks, suggesting that there are two ranges where the similarity scores are most densely populated. The first, most pronounced peak occurs around 0.6 and the second is just below 1.0. The peak near 1.0 suggests a high density of job titles with very close or perfect matches to O*NET, implying that many of the identified jobs align closely with currently recognized titles. However, the presence of a peak at 0.6 indicates a substantial number of job titles with only moderate or low similarity to those in O*NET. These scores could potentially represent new, emerging roles within the green economy that are not adequately captured by the current taxonomy.

Of the green job titles that were identified, 116 (17 percent) perfectly match those within O*NET (O*NET comprises a total of 205 job titles) (table 6). These matches are observed in articles that reference O*NET explicitly, such as Vona, Marin, and Consoli (2019), and in those that do not mention O*NET, like Wandzich and Plaza (2017). The latter instances are particularly valuable as they may be considered independent confirmations of the green nature of these job titles, providing additional validity to the O*NET taxonomy.

The specific values of cosine similarity do not carry an absolute meaning and are context-dependent, particularly within the framework of specific modeling tasks. However, meaningful insights can be derived by examining the pairs of matched job titles across various ranges of cosine similarity values. Table 6 provides illustrative examples from the file available on GitHub for the five closest matches from O*NET for all the identified job titles. For values between 0.9 and 1.0, the paired job titles typically represent the same roles, with variations in wording, such as "logistical analysts" versus "logistics analysts." In the 0.8 to 0.9 range, we see similar instances, such as "zoologists and biologists" versus "zoologists and wildlife biologists," where differences might be minimal. However, this range can also include pairs where subtle yet significant differences exist, such as "hydrogeologists" versus "hydrologists," where distinctions may reflect actual differences in job scope. In the lower ranges, such as 0.4 to 0.6, the matches often involve more specific job titles being compared to broader categories, for example, "horticulturists" versus "agricultural technicians," or they may represent complete mismatches like "quality control specialist" versus "regulatory affairs specialist." Below 0.4, the differences become more pronounced, with matches like "meat scientist" versus "material scientist" suggesting only superficial or merely semantic similarities and larger potential to identify new green job titles. Additionally, some distinctions are due to regional particularities, such as the term "urban planners" used in the United States, which corresponds to "spatial planning" in European contexts.

Our compiled list serves as a preliminary framework for the more systematic task of assimilating identified titles into specific, regulated occupations, a process that should be conducted in accordance with the policies of the respective classification system. Our findings indicate that there are new green jobs in the literature that are not currently included in O*NET. Expanding O*NET to include these new jobs would require discussions about whether some of the identified titles are too narrow or too similar to existing categories within O*NET. For example, we have identified the job title "environmental compliance specialist," which does not currently exist in O*NET. The closest matches are "sustainability specialists," who do not primarily focus on legal aspects, and

"compliance specialists," which is a broader category. Our research also indicates emerging roles prompted by technological advancements and evolving industry needs, which are not yet reflected in O*NET. Examples include "mechanic for electric cars," "data engineer for the energy sector," "meat scientist," "decentralized-energy engineer," "specialist in ICT coordination of transport systems," and positions born from heightened environmental consciousness and regulations, such as "carbon auditor," "green architect," "organic farmer," or "ecological designer."

Finally, as an additional angle comparing our identified job titles with those in O*NET, we organize the green professions into major green economy sectors through clustering based on their semantic similarity. We identify 25 distinct clusters of job titles that can be interpreted as green economy sectors (figure 4). Our results align closely with O*NET, which identifies 12 green economy sectors where green occupations can be found (Dierdorff et al., 2009). However, some O*NET sectors are further disaggregated in our analysis (for example, renewable energy is divided into solar energy, wind energy, and so forth). This disaggregation is a result of the model parameters applied. Fine-tuning these parameters allows for merging clusters, and we designated the cluster names. Table 7 presents the correspondence between the sectors we identified using clustering and those in O*NET. There are two notable differences. First, we identified a cluster of job titles related to green human resources, which is not a distinct sector in O*NET. Second, O*NET includes a separate sector for energy trading, which contains only one occupation: energy broker. Since we set a minimum of four similar occupations to form a cluster and found fewer than four related to energy trading, we do not identify such a cluster.

4. Validation

The RAG model successfully discerned specific job titles, positions, or occupations from more general descriptions. We performed two types of robustness checks to validate our job title identification process. The first check aimed to detect potential false negatives—instances where green job titles were present in an article's text but the model overlooked them. To do this, we inspected a subset of articles where the model did not identify any green job titles. We selected 5 percent of the articles for which the full text was available but no job titles were identified, chosen through a pseudorandom process to maintain transparency.[3] In the second test, we closely reviewed one-third of all the articles where green job titles were flagged by the model to ascertain false positives—titles incorrectly marked as green when they were not mentioned in a green context—and, to ensure completeness, we verified that the model did not omit any pertinent job titles.

The first test did not reveal any omissions of green job titles. Most of the reviewed articles discussed green sectors or activities but did not mention specific job titles. For example, Wang et al. (2022) mention "ecopreneurs" as mediators linking green economy policies to sustainable economic development. However, the model did not identify ecopreneurs as a green job title. In a post-check, we utilized the chat feature of the model to inquire about this specific choice. The response suggested that this was not an oversight but a deliberate decision, aligning with the classification approach we intended to use. The model indicated that ecopreneurs is indeed mentioned as a role, though it is more of a descriptive term rather than a formal job title. Ecopreneurs are entrepreneurs who focus on ecofriendly or sustainable business practices.

[3] A pseudorandom process is one that appears random but is actually deterministic, generated by an algorithm; thus, it is reproducible, which aids in the transparency of our methodology.

The post-check process has limitations, as it was conducted in a separate model session. Consequently, the model did not retain memory of the specific decision made during the initial analysis. Instead, we recreated the same situation by inputting the same text and receiving the same result before asking for clarification. This method does not guarantee that the model's reasoning process in the post-check exactly mirrors its initial reasoning, but it provides a useful approximation of the model's decision-making logic.

Analysis of the texts from articles where green job titles were identified did not reveal any omissions of relevant green job titles. In instances where additional job titles were mentioned within these articles, they were categorized as nongreen or described as very general positions such as "managers" or "supervisors." These titles were not pertinent to our specific focus on identifying distinctly green jobs, thus confirming the effectiveness and precision of our methodology for filtering out job titles that do not meet the criteria for this exercise.

This check also revealed that the model was cautious in suggesting that a job is considered green by the authors. For instance, Vona, Marin, and Consoli (2019) provide a table of green job titles, which they describe as putatively problematic due to the mixed tasks involved. The model did not identify these job titles as green. The exclusion of such borderline cases indicates that the prompt we used set the model to err on the side of specificity, avoiding ambiguous classifications. To include these nuanced cases in future analyses, the model's prompt would need to be adjusted, potentially to broaden its criteria for what constitutes a green job. This adjustment could help to capture a wider spectrum of jobs, including those that are functionally ambiguous but still relevant to the discussion on green jobs.

5. Limitations

We discuss key limitations of the study from two perspectives: narrow and broad.

From a narrow perspective, we note two limitations. First, we had access to the full texts of only 567 of 1,067 articles (53 percent). Our results indicate that specific job titles are more frequently mentioned within the main text rather than in abstracts. We obtained articles from reputable bibliographic databases to maintain quality control and systematically documented the retrieval process. However, industry-specific publications may contain a wider range of job titles, which O*NET actively incorporates. Our method is adaptable and can be applied to such industry texts without modification.

Second, in rare instances, our model failed to identify job titles included in noneditable figures or tables, although we did not detect such cases during our validation checks. It is technically feasible to refine the method to include such content by modifying the data ingestion phase. Our model's final pipeline stage involved experimenting with various prompts to determine if jobs were mentioned in a green context.

From a broader perspective, we identify three important limitations. First, any publication bias, through researchers' own areas of interest or due to peer reviewed journal selection processes, will inevitably transfer into our study. Although one of the strengths of the proposed methodology is

that the green titles are directly extracted from articles found in two leading bibliographic databases, it is well known that there is a positive bias for research focusing on high-income countries. Gomez et al. (2022) show that academic research from higher-income countries is much more widely cited than comparable work from lower-income countries, and this bias has only grown over the past 35 years.[4]

This type of publication bias means that the emergence of green occupations may not be well captured by the proposed taxonomy, as the economic structure, technology, and occupational tasks are not necessarily the same as in lower-income countries. Additionally, informal labor markets, especially in agriculture, are prevalent in lower-income countries. While certain agricultural occupations are classified as green in this taxonomy, the methods used in lower-income countries may not necessarily be as environmentally friendly. Since our method relies on the published literature, the final green jobs list is dependent on the topics that researchers choose to study and publish, and/or on the questions the reviewers and editors deemed important. Emerging green occupations or innovative practices that have not yet attracted significant academic attention may be underrepresented or entirely absent from the database. Furthermore, peer review and publication processes can take a significant amount of time in certain fields (such as economics), which can slow down our understanding of the full spectrum of new emerging green jobs.

Second, much like O*NET's green occupations, our model categorizes jobs as green or nongreen, creating a binary taxonomy rather than a continuous spectrum of greenness based on specific tasks. Our method provides a useful list of existing and potentially emerging jobs that are likely to contribute to the green economy, but it fails to capture the complexity of certain jobs. As highlighted by Vona et al. (2018), some occupations, such as sheet metal workers, encompass both green and nongreen tasks, making it inaccurate to label them as entirely green. Another example is "lawyers with solar expertise," which our method classifies as a green job. This occupation may involve a mix of environmentally focused and conventional legal tasks, demonstrating the limitations of a binary classification system. Therefore, the proposed taxonomy in this study could overestimate the extent of greenness of labor markets compared to the continuous measures proposed by Vona et al. (2018).

Third, while our model can access and analyze the context in which green occupations are discussed within each article, the end users of the final green job titles list are often left uninformed about the context that led to the selection of the job title as green. This lack of context and absence of expert review can lead to misidentification of green job titles. Furthermore, without the nuanced understanding that experts bring, the binary classification system employed by the model might overlook the complexity of certain jobs that include both green and nongreen tasks, resulting in an oversimplified taxonomy.

6. Conclusion

Based on a search of the academic literature published between January 2009 and April 2024, we identified 1,067 articles in the green literature, consisting of 567 full texts and 500 abstracts. Of these, 105 articles contained at least one specific green job title, while other articles discussed green jobs in general without specifying particular roles. We identified 695 unique green job titles

[4] The study covers nearly 20 million articles and 150 fields.

after excluding nonspecific roles and standardizing the results. Based on our ability to match these jobs to O*NET, we demonstrate the existence of new green jobs as well as a new green economy sector.

Identifying specific green job titles within the literature is challenging. A major hurdle is the sheer volume of literature, which often necessitates stringent filtering criteria for systematic literature reviews to manage the number of articles for human researchers. Using our model has the advantage of diligently handling tasks without the fatigue and inconsistency that human annotators may experience. This results in a more reliable collection and evaluation of information about green job titles from a wide range of scientific texts. In addition, the process is reproducible, providing a clear trail of the articles behind each identified job title. This marks a significant step forward in leveraging AI to enhance and refine occupational classifications in line with the evolving landscape of the green economy.

Our results demonstrate that by employing the RAG model, we can efficiently compile a list of potential green job titles using the vast literature and potentially use them to update classification systems like O*NET. The method's satisfactory performance in matching job titles to O*NET suggests that it can also perform well in creating crosswalks between different occupational classification schemes, potentially automating a task that was traditionally performed manually.

While the current study expands the existing list of green jobs, three main directions for future research emerge. First, although we excluded synonyms among the identified job titles and merged their singular and plural versions, a separate study would be needed to delve deeper into distinguishing true new job titles from those that might be just new names for old roles. Second, it is essential to identify the greenness of these jobs. The current study is not able to determine how green a job is based on the tasks it involves. Third, matching the identified job titles to occupational classifications, such as the International Standard Classification of Occupations, is essential for integrating the updated list of green jobs into relevant labor market studies.

Tables and Figures

Table 1. Publications Return from Literature Search and Screening Process

Selection stage	Scopus	Web of Science	Total
Initial search	1,394	489	1,883
Retaining articles only	991	376	1,367
Combined set based on ISSN and DOI excluding duplicates	n.a.	n.a.	1,067
Full-text articles	n.a.	n.a.	567
Abstracts only	n.a.	n.a.	500

Note: n.a. = not applicable; DOI = Digital Object Identifier; ISSN = International Standard Serial Number.

Table 2. Distribution of Publications Initially Identified in Scopus and Web of Science, by Publication Type

Type	Scopus	Web of Science
Article	915	353
Book or book chapter	194	25
Conference paper	164	57
Review	76	31
Other	45	23

Table 3. Yearly Cumulative Number of Green Literature Articles, by Region

Publication year	Africa	Asia	Europe	North America	Oceania	South America
2009	1	4	8	21	2	0
2010	1	7	15	31	5	0
2011	4	14	26	43	6	1
2012	5	16	42	51	6	1
2013	9	21	51	63	8	5
2014	10	28	65	69	8	5
2015	13	38	85	71	11	7
2016	13	44	116	83	11	7
2017	14	54	141	91	11	9
2018	16	63	159	100	11	9
2019	17	72	172	110	15	10
2020	21	84	192	119	16	12
2021	29	110	236	128	17	13
2022	30	141	276	135	19	21
2023	38	189	321	144	21	25
2024	39	209	324	147	24	28

Note: The literature search was completed in April 2024. Consequently, the statistical information for 2024 is truncated, reflecting only the first three months of the year.

Table 4. Frequency of Mentions of Countries and Political Entities in the Articles with Identified Green Job Titles

Country/region	Number
Global	28
United States	20
European Union	13
Brazil	6
China	6
Spain	6
Czechia	4
Finland	3
India	3
Nigeria	3
Australia	2
Bulgaria	2
Canada	2
Indonesia	2
Lithuania	2
Scotland	2
South Africa	2
United Kingdom	2
Germany	2
France	2
Italy	2
Argentina	2
Belgium	1
Denmark	1
Egypt, Arab Rep.	1
Ghana	1
Greece	1
Hungary	1
Kenya	1
Malaysia	1
Malta	1
Namibia	1
Netherlands	1
Pakistan	1
Romania	1
Serbia	1
Slovak Republic	1
Korea, Rep.	1
Türkiye	1
United Arab Emirates	1
Viet Nam	1

Table 5. Frequency of Mentions of Economic Activities in the Articles with Identified Green Job Titles

ISIC code	ISIC name	Number
O	Public administration and defense	24
C	Manufacturing	17
Q	Human health and social work activities	14
E	Water supply; sewerage, waste management	12
D	Electricity, gas, steam, and air conditioning supply	11
M	Professional, scientific, and technical activities	9
N	Administrative and support service activities	4
A	Agriculture, forestry, and fishing	3
<all economy>	-	3
H	Transportation and storage	2
K	Financial and insurance activities	2
L	Real estate activities	2
P	Education	2
J	Information and communication	1
W	Activities of extraterritorial organizations and bodies	1
F	Construction	1
I	Accommodation and food service activities	1

Note: ISIC = International Standard Industrial Classification of All Economic Activities. <all economy> = no specific economic activity could be identified.

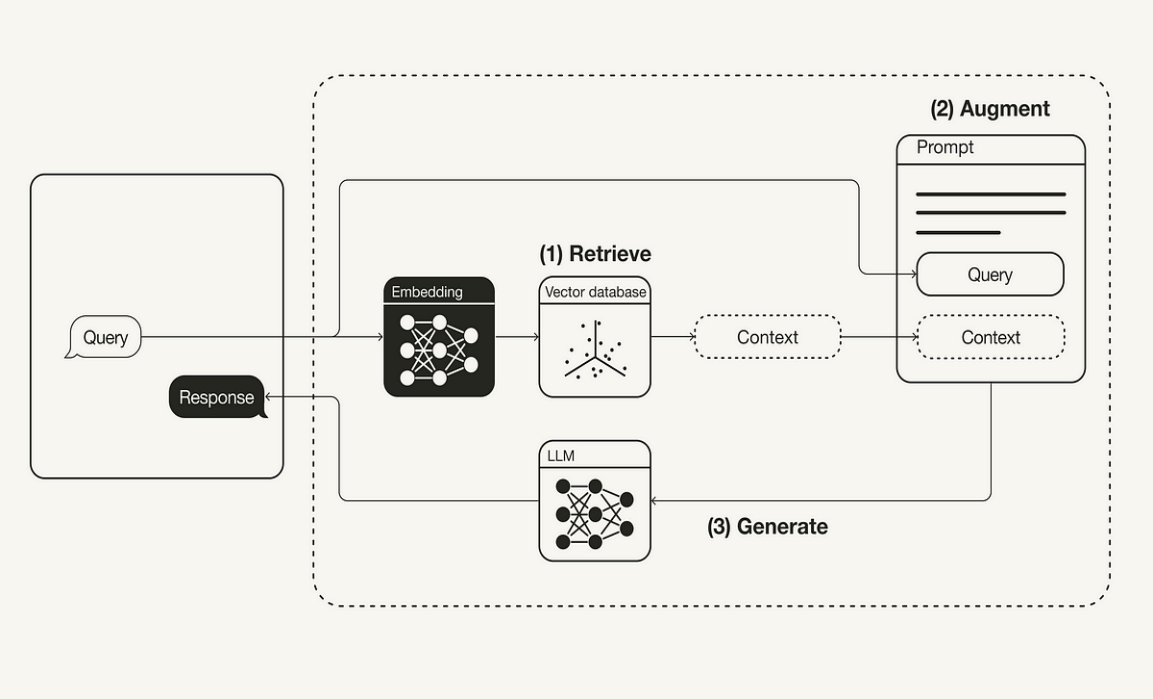
Table 6. Examples of Pairs of Matched Job Titles for Different Cosine Similarity Ranges

Green job title candidate	Closest O*NET classification match	Cosine similarity range	Percent of matched job titles
Logistical analyst	Logistics analyst	0.9–1.0	17
Electrical equipment assembler	Electrical and electronic equipment assembler		
Geological sample test technologist	Geological sample test technician		
Zoologists and biologist	Zoologist and wildlife biologist	0.8–0.9	7
Specialist in conservation of forests, technician	Forest and conservation technician		
Hydrogeologist	Hydrologist		
Field engineer (wind energy)	Wind energy engineer	0.6–0.8	42
Industrial recycling specialist	Recycling coordinator		
Ecological restoration specialist	Environmental restoration planner		
Quality control specialist	Regulatory affairs specialist	0.4–0.6	32
Urban afforestation specialist	Forest and conservation technician		
Horticulturist	Agricultural technician		
Animal caretaker	Agricultural technician	0.2–0.4	2
Arborist	Forest and conservation worker		
Meat scientist	Materials scientist		

Table 7. Comparison of O*NET Green Sectors and Identified Clusters

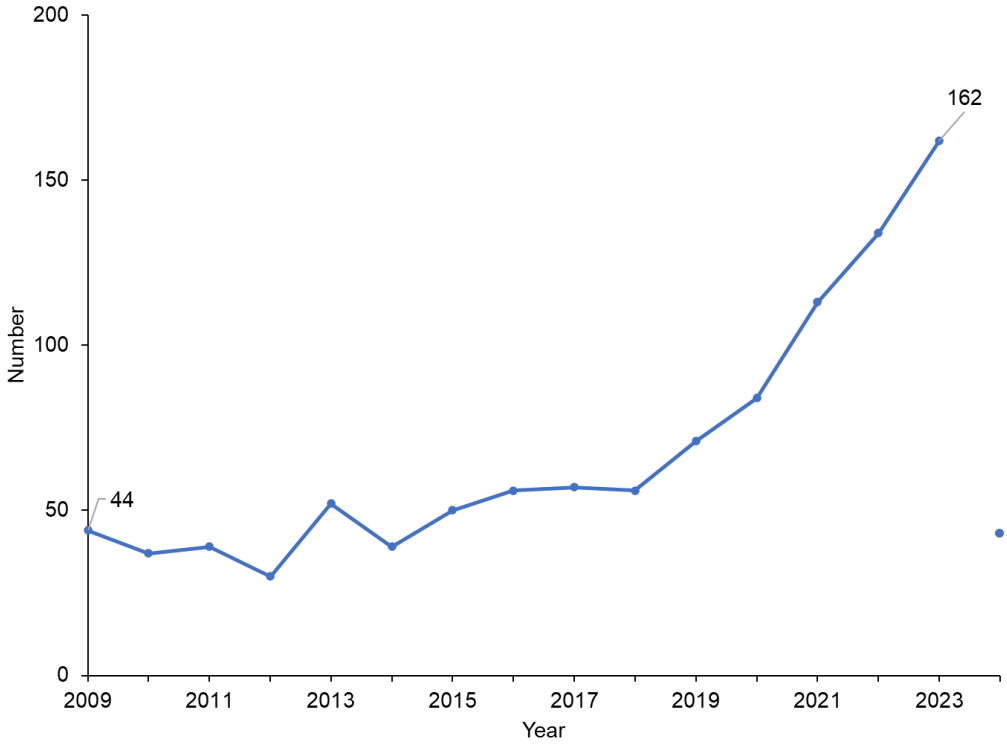
O*NET sector	Clusters identified based on semantic analysis
Agriculture and Forestry	Sustainable and Organic Agriculture, Forestry Conservation, Sustainable Landscaping
Energy and Carbon Capture and Storage	Carbon and Clean Energy Analysis
Energy Efficiency	Energy Efficiency and Retrofitting, Electrical and Civil Engineering and Maintenance
Energy Trading	-
Environment Protection	Environmental Advocacy and Restoration, Pollution Control
Governmental and Regulatory Administration	Green Regulatory Compliance, Environmental Policy and Compliance
Green Construction	Green Construction, Sustainable Infrastructure
Manufacturing	Green Manufacturing and Assembly
Recycling and Waste Reduction	Waste Management
Renewable Energy Generation	Solar Energy, Wind Energy, Geothermal Energy, Renewable Energy Engineering
Research, Design, and Consulting Services	Environmental Science and Geospatial Analysis, Environmental Technology and Health, Sustainable Development Strategy, Sustainability Management and Consulting
Transportation	Transportation, Logistics and Supply Chain Management
-	Green human resources Management

Figure 1. RAG Model Pipeline



Source: Ganesh 2024.
Note: LLM = large language model; RAG = retrieval-augmented generation.

Figure 2. Green Literature Articles over Time, 2009-2024



Note: The literature search was conducted in April 2024. Consequently, the statistical information for 2024 is truncated, reflecting only the first three months of the year.

Figure 3. Distribution of Cosine Similarity Scores between Identified Green Job Titles and O*NET

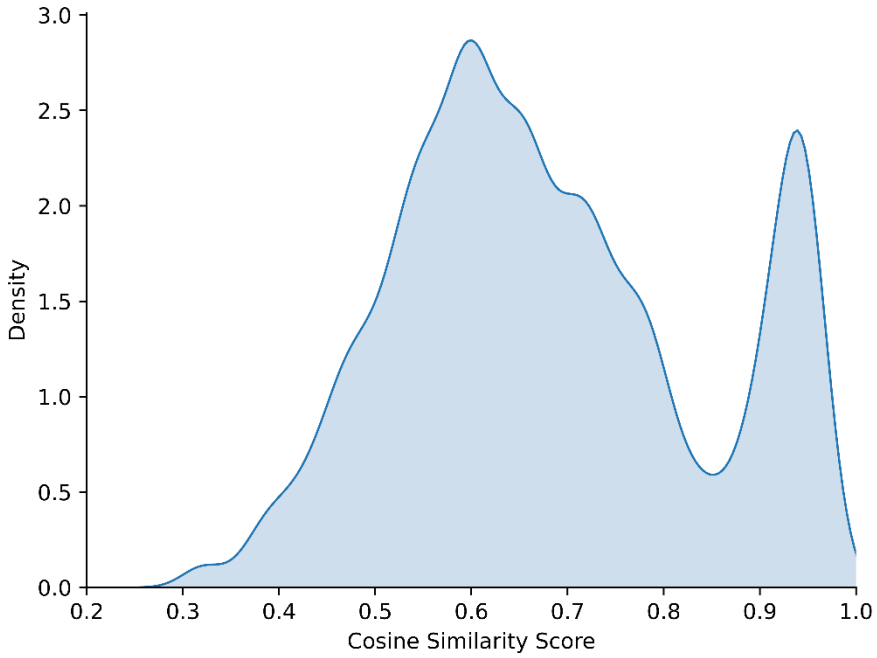
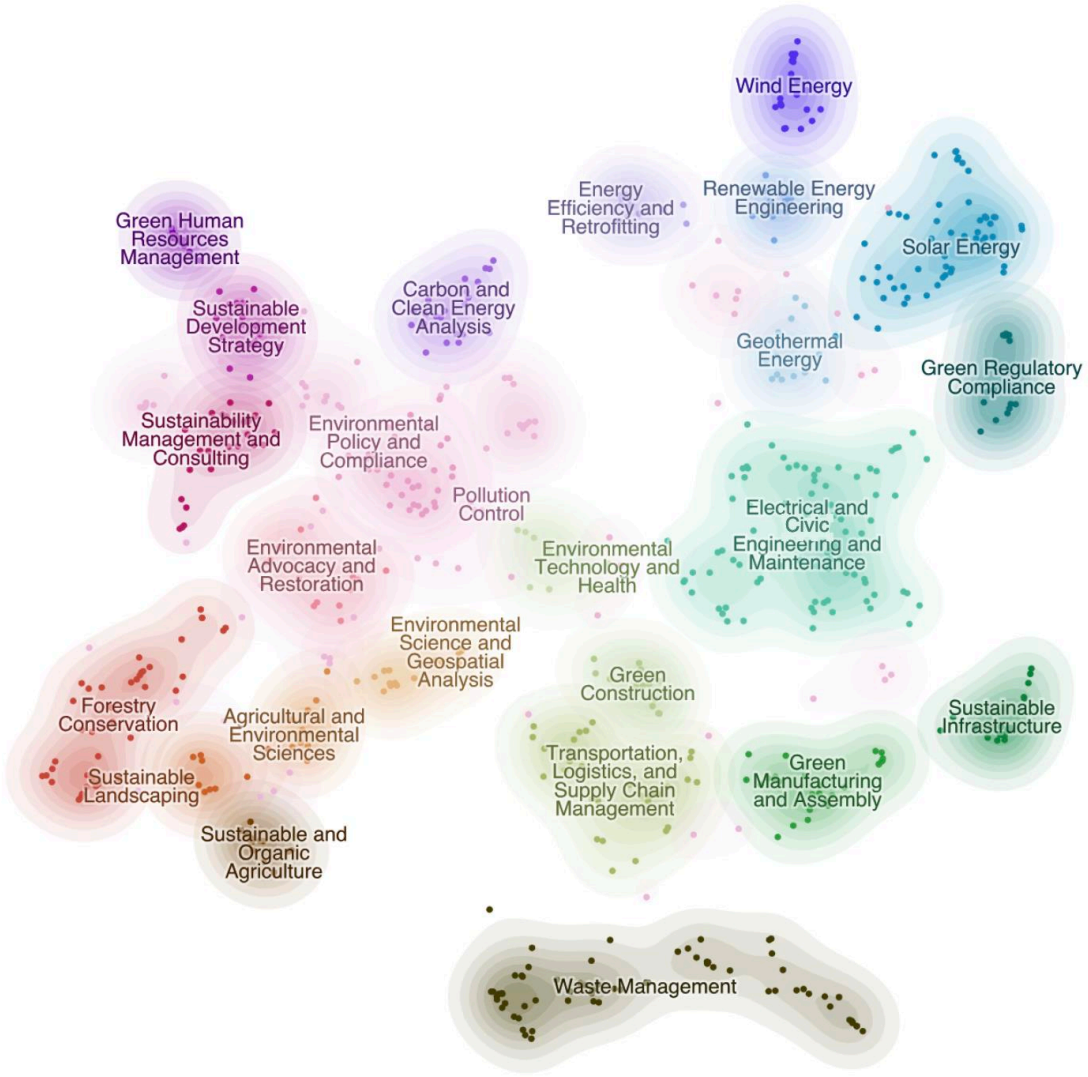


Figure 4. Clusters of Green Job Titles indicating Major Green Economy Areas



Box 1. Search Queries

Scopus query: TITLE-ABS-KEY ("green job" OR "sustainable job" OR "green occupation" OR "sustainable occupation" OR "green work" OR "sustainable work" OR "green employment" OR ("green transition" AND "job")) AND PUBYEAR > 2008 AND PUBYEAR < 2025 AND LANGUAGE (english)

Web of Science query: (AB=("green job" OR "green occupation" OR "green employment" OR "sustainable job" OR "sustainable occupation") OR TI=("green job" OR "green occupation" OR "green employment" OR "sustainable job" OR "sustainable occupation") OR AK=("green job" OR "green occupation" OR "green employment" OR "sustainable job" OR "sustainable occupation" ("green transition" AND "job")) OR TS=("green job")) AND PY=(2008-2024) AND LA=(english)

References

Afolabi, Abimbola O., Rasheed A. Ojelabi, P. Funmilayo Tunji-Olayeni, Olabosipo I. Fagbenle, and T. Oluwaseun Mosaku. 2018. "Survey Datasets on Women Participation in Green Jobs in the Construction Industry." *Data in Brief* 17: 856–62.

Anadkat, Shashank. 2023. "How to Make Your Completions Outputs Consistent with the New Seed Parameter." OpenAI Cookbook, November 2023 (accessed April 14, 2024), https://cookbook.openai.com/examples/reproducible_outputs_with_the_seed_parameter.

Apostel, Amelie, and Mikkel Barslund. 2024. "Measuring and Characterising Green Jobs: A Literature Review." *Energy Research & Social Science* 111: 103477.

Bowen, Alex, and Bob Hancké. 2019. *The Social Dimensions of "Greening the Economy."* Brussels, Belgium: European Commission Directorate-General for Employment, Social Affairs and Inclusion.

Chiarello, Francesco, Gualtiero Fantoni, Terence Hogarth, Vincenzo Giordano, Ludmila Baltina, and Ilaria Spada. 2021. "Towards ESCO 4.0–Is the European Classification of Skills in Line with Industry 4.0? A Text Mining Approach." *Technological Forecasting and Social Change* 173: 121177.

Consoli, Davide, Giovanni Marin, Alberto Marzucchi, and Francesco Vona. 2016. "Do Green Jobs Differ from Non-Green Jobs in Terms of Skills and Human Capital?" *Research Policy* 45(5): 1046–60.

Davis, Georgina. 2013. "Counting (Green) Jobs in Queensland's Waste and Recycling Sector." *Waste Management & Research* 31 (9): 902–9.

de la Vega, Patricio, Nicolas Porto, and Marcos Cerimelo. 2024. "Going Green: Estimating the Potential of Green Jobs in Argentina." *Journal for Labour Market Research* 58 (1): 1–19.

Decorte, Jeroen J., Jelmer Van Hautte, Thomas Demeester, and Chris Develder. 2021. "Jobbert: Understanding Job Titles through Skills." arXiv preprint arXiv:2109.09605.

Dierdorff, Erich C., Jonathan J. Norton, Donald W. Drewes, Christina M. Kroustalis, David Rivkin, and Philip Lewis. 2009. "Greening of the World of Work: Implications for O*NET-SOC and New and Emerging Occupations." National Center for O*NET Development, U.S. Department of Labor, Washington, DC.

Dierdorff, Erich C., Jonathan J. Norton, Corey M. Gregory, David Rivkin, and Philip Lewis. 2011. "Greening of the World of Work: Revisiting Occupational Consequences." National Center for O*NET Development, U.S. Department of Labor, Washington, DC.

Doan, Dung, Trang Luu, Nga Thi Nguyen, and Abla Safir. 2023. "Green Jobs: Upskilling and Reskilling Vietnam's Workforce for a Greener Economy." World Bank, Washington, DC.

EC (European Commission). 2022. “Green Skills and Knowledge Concepts: Labelling the ESCO Classification.” Publications Office of the European Union, Luxembourg.

Elliott, Robert J., Wei Kuai, David Maddison, and Ceren Ozgen. 2021. “Eco-innovation and Employment: A Task-based Analysis.” Discussion Paper 14028, Institute of Labor Economics, Bonn, Germany.

Ganesh, Rohan. 2024. “Understanding RAG and Vector DB.” https://ai-for-production.blog/2024/01/24/understanding-rag-and-vector-db/.

Gao, Yikun, Yue Xiong, Xin Gao, Kelei Jia, Jing Pan, Yunzhe Bi, and Hui Wang. 2023. “Retrieval-Augmented Generation for Large Language Models: A Survey.” arXiv preprint arXiv:2312.10997.

Gomez, Clara J., Alexandra C. Herman, and Paolo Parigi. 2022. “Leading Countries in Global Science Increasingly Receive More Citations than Other Countries Doing Similar Research.” *Nature Human Behaviour* 6 (7): 919–29.

Kozar, Łukasz J., and Adam Sulich. 2023. “Green Jobs: Bibliometric Review.” *International Journal of Environmental Research and Public Health* 20 (4): 2886.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, and Douwe Kiela. 2020. “Retrieval-augmented Generation for Knowledge-Intensive NLP Tasks.” *Advances in Neural Information Processing Systems* 33: 9459–74.

Li, Jing, Aixin Sun, Jiawei Han, and ChengXiang Li. 2020. “A Survey on Deep Learning for Named Entity Recognition.” *IEEE Transactions on Knowledge and Data Engineering* 34 (1): 50–70.

Li, Shuang, Bolei Shi, Jian Yang, Jun Yan, Sheng Wang, Fei Chen, and Qiang He. 2020. “Deep Job Understanding at LinkedIn.” In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2145–48. New York: Association for Computing Machinery.

Liga, Davide, and Livio Robaldo. 2023. “Fine-tuning GPT-3 for Legal Rule Classification.” *Computer Law & Security Review* 51: 105864.

McInnes, Leland, John Healy, and Steve Astels. 2017. “HDBSCAN: Hierarchical Density Based Clustering.” *Journal of Open Source Software* 2 (11): 205.

McInnes, Leland, John Healy, and James Melville. 2018. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” arXiv preprint arXiv:1802.03426.

Monajatipoor, Mehran, Jing Yang, Justin Stremmel, Mahtab Emami, Fatemeh Mohaghegh, Mehrdad Rouhsedaghat, and Kai-Wei Chang. 2024. “LLMs in Biomedicine: A Study on Clinical Named Entity Recognition.” arXiv preprint arXiv:2404.07376.

Nhamo, Godwell. 2010. “Green Economies and Green Jobs: Implications for South Africa.” *WIT Transactions on Ecology and the Environment* 131: 257–68.

O*NET (Occupational Information Network). 2013. *Greening of the World of Work: O*NET Project’s Book of References*. Washington, DC: National Center for O*NET Development, U.S. Department of Labor. https://www.onetcenter.org/reports/GreenRef.html.

OECD (Organisation for Economic Co-operation and Development). 2023. *Job Creation and Local Economic Development 2023: Bridging the Great Green Divide*. Paris: OECD.

Papoutsoglou, Maria, E. S. Rigas, Georgia M. Kapitsaki, Lefteris Angelis, and Johannes Wachs. 2022. “Online Labour Market Analytics for the Green Economy: The Case of Electric Vehicles.” *Technological Forecasting and Social Change* 177: 121517.

Pawar, Chandrashekhar S., and Ashwin Makwana. 2022. “Comparison of BERT-Base and GPT-3 for Marathi Text Classification.” *In Futuristic Trends in Networks and Computing Technologies: Select Proceedings of Fourth International Conference on FTNCT 2021*, edited by Chandrakant Nalbalwar, Shailendra Singh, and Sonal Jain, [229–39]. Singapore: Springer Nature Singapore.

Popp, David, Francesco Vona, Giovanni Marin, and Ziqiao Chen. 2020. “The Employment Impact of Green Fiscal Push: Evidence from the American Recovery Act.” Working Paper 27321, National Bureau of Economic Research, Cambridge, MA.

Rehana, Hasin, Nur Bengisu Çam, Mert Basmaci, Jie Zheng, Christianah Jemiyo, Yongqun He, Arzucan Özgür, and Junguk Hur. 2023. “Evaluation of GPT and BERT-Based Models on Identifying Protein-Protein Interactions in Biomedical Text.” *ArXiv*.

Stanef-Puică, Mădălina-Roxana, Lucia Badea, Geanina L. Serban-Oprescu, Alin-Tiberiu Serban-Oprescu, Laura Gabriela Frâncu, and Andreea Cretu. 2022. “Green Jobs—A Literature Review.” *International Journal of Environmental Research and Public Health* 19: 7998.

Strachan, Sara, Alexander Greig, and Alison Jones. 2022. “Going Green Post COVID-19: Employer Perspectives on Skills Needs.” *Local Economy*, 37 (6): 481–506.

Theune, Christian. 2024. *pycountry* (version 24.6.1) [Computer software]. https://pypi.org/project/pycountry/.

Valero, Ana, Jing Li, Samuel Muller, Cristina Riom, Vu Nguyen-Tien, and Mirko Draca. 2021. “Are ‘Green’ Jobs Good Jobs? How Lessons from the Experience To-Date Can Inform Labour Market Transitions of the Future.” Grantham Research Institute on Climate Change and the Environment and Centre for Economic Performance, London School of Economics and Political Science, London.

Vona, Francesco, Giovanni Marin, and Davide Consoli. 2019. “Measures, Drivers and Effects of Green Employment: Evidence from US Local Labor Markets, 2006–2014.” *Journal of Economic Geography* 19 (5): 1021–48.

Vona, Francesco, Giovanni Marin, Davide Consoli, and David Popp. 2018. “Environmental Regulation and Green Skills: An Empirical Exploration.” *Journal of the Association of Environmental and Resource Economists* 5 (4): 713–53.

Wandzich, Daniela E., and Grażyna A. Płaza. 2017. “New and Emerging Risks Associated with ‘Green’ Workplaces.” *Workplace Health & Safety* 65 (10): 493–500.

Wang, Shuang, Xiaoxu Sun, Xin Li, Renjie Ouyang, Fangxiang Wu, Tingting Zhang, and Guoxin Wang. 2023. “GPT-NER: Named Entity Recognition via Large Language Models.” arXiv preprint arXiv:2304.10428.

Wang, Xiaoxiang, Muhammad U. Javaid, Sidra Bano, Hasnat Younas, Abdurrahman Jan, and Ahmad A. Salameh. 2022. “Interplay among Institutional Actors for Sustainable Economic Development—Role of Green Policies, Ecopreneurship, and Green Technological Innovation.” *Frontiers in Environmental Science* 10: 956824.

Wijesinghe, Amayaa, and Jessica P. R. Thorn. 2021. “Governance of Urban Green Infrastructure in Informal Settlements of Windhoek, Namibia.” *Sustainability* 13 (16): 8937.

Yepes, Andrés J., Yiyang You, Jaroslaw Milczek, Sergio Laverde, and Linwei Li. 2024. “Financial Report Chunking for Effective Retrieval Augmented Generation.” arXiv preprint arXiv:2402.05131.

Zhou, Wenxuan, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. “Universalner: Targeted Distillation from Large Language Models for Open Named Entity Recognition.” arXiv preprint arXiv:2308.03279.

Zhu, Jing, and Wei Liu. 2020. “A Tale of Two Databases: The Use of Web of Science and Scopus in Academic Papers.” *Scientometrics* 123: 321–35.

Appendix: Details on the RAG Model Implementation

We implemented the retrieval-augmented generation (RAG) model through the following pipeline stages:

1. **Ingestion.** The initial stage involves parsing Portable Document Format (PDF) documents to extract text. Given the diverse formatting and structures of academic articles, this step is crucial to ensure accurate text extraction for further processing.

2. **Content filtering.** Large language models (LLMs) have limited context windows and perform better when noisy information is filtered out. It is essential to focus on the relevant parts of the articles by removing sections that are unlikely to contain the targeted information, such as the references section. This step is especially useful when working with smaller models (pre-GPT-4) and speeds up the retrieval process.

3. **Text chunking.** Chunking longer texts into smaller sections is necessary due to the LLM's context window limitation. Simple chunking strategies, such as dividing the text based on a fixed number of characters with some overlap, often lead to suboptimal results. These methods can disrupt the narrative or logical flow of the text, making it harder for the model to understand the context and extract meaningful information. To improve on simple chunking, we applied a method that identifies distinct sections within the articles. This approach leverages a set of rules related to font styles, numbering, and the presence of newlines, which are common indicators of new sections in academic articles. By identifying these sections, we can chunk the text in a way that maintains its logical and narrative structure, enhancing the model's ability to interpret and analyze the content effectively.

4. **Embeddings.** Once data ingestion and preliminary text processing are complete, the next step is to transform the textual data into a form the model can efficiently process. This transformation involves generating embeddings, where the text data are converted into high-dimensional vectors. These embeddings capture the semantic and syntactic essence of the text in a numerical format, allowing machine learning models to understand and process the text. We utilized OpenAI's latest model, text-embedding-3-large, which embeds texts using 3,072 dimensions.

5. **Storing embeddings in vector databases.** After the embeddings are generated, they need to be stored in a manner that facilitates quick and efficient retrieval. This is where vector databases come into play. Unlike traditional databases optimized for scalar data, vector databases are specifically designed to handle high-dimensional vectorized data. By storing the embeddings in vector databases, we ensure that the system can rapidly perform search and retrieval operations, which is crucial for the RAG process.

6. **LLMs as the generative component.** LLMs are at the heart of the RAG model, serving as the generative component. These models have been trained on extensive corpora of text, which equips them with a nuanced understanding of language and the ability to generate coherent, contextually relevant text. In the RAG model, the LLM operates by taking the

input query and the context information retrieved from the vector databases to generate responses. This capability is particularly valuable in our case for discerning the context in which job titles are mentioned and determining their relevance to the green economy.

The querying stage is where the interaction between the user's input and the stored embeddings unfolds. When we input a query, the RAG model leverages the embeddings stored in the vector databases to find relevant information. It does this by converting the query into a vector and comparing it with the stored vectors using similarity metrics (cosine similarity in our case). The most relevant embeddings (and thereby the associated text or information) are retrieved from the database. The LLM then uses this retrieved information, combined with the original query, to generate an informed response or classify the text accurately, such as identifying green job titles within the provided text.

Our prompt for the job identification was: "Return all green job titles, positions, or occupations mentioned in the article's section. Return only specific job titles, positions, or occupations, e.g., 'Electrical Engineer' mentioned in the context of the green economy, sustainability, countering climate change, etc. If none are mentioned explicitly, return <No results>. Separate results with commas."

As for the model's parameters, we used a temperature of 1 and a top-p of 1. The temperature parameter controls the randomness of the model's predictions, with a value of 1 providing a balanced approach between creativity and determinism. The top-p parameter, also known as nucleus sampling, determines the diversity of predictions by considering the smallest set of tokens whose cumulative probability meets or exceeds the top-p value. By setting both parameters to 1, we chose values that allow for a fully deterministic generation process, ensuring stable and controlled outputs.