KARRIEREWEGE: A Large Scale Career Path Prediction Dataset

Elena Senger[1,2] **Yuri Campbell**[2] **Rob van der Goot**[3] **Barbara Plank**[1]

[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Fraunhofer Center for International Management and Knowledge Economy IMW, Germany
[3]Department of Computer Science, IT University of Copenhagen, Denmark
elena.senger@cis.lmu.de, yuri.campbell@imw.fraunhofer.de
robv@itu.dk, b.plank@lmu.de

Abstract

Accurate career path prediction can support many stakeholders, like job seekers, recruiters, HR, and project managers. However, publicly available data and tools for career path prediction are scarce. In this work, we introduce KARRIEREWEGE, a comprehensive, publicly available dataset containing over 500k career paths, significantly surpassing the size of previously available datasets. We link the dataset to the ESCO taxonomy to offer a valuable resource for predicting career trajectories. To tackle the problem of free-text inputs typically found in resumes, we enhance it by synthesizing job titles and descriptions resulting in KARRIEREWEGE+. This allows for accurate predictions from unstructured data, closely aligning with real-world application challenges. We benchmark existing state-of-the-art (SOTA) models on our dataset and a prior benchmark and observe improved performance and robustness, particularly for free-text use cases, due to the synthesized data.

1 Introduction

Career path prediction (also known as career trajectory prediction) is a growing field (Shreyas et al., 2024), with the potential to inform recruitment, career counseling, upskilling or reskilling, or more broadly workforce planning and workforce trends. The task is to predict future career moves based on an individual's work history, possibly using further information such as skills or education. To achieve this, robust datasets that capture detailed career histories are essential. However, the availability of large-scale benchmark career history datasets remains limited (Du et al., 2024; Decorte et al., 2023), posing a major challenge for the field.

A dataset mapped to ESCO (European Skills, Competences, Qualifications, and Occupations) Taxonomy is particularly advantageous because ESCO provides a standardized "common language" for occupations and skills across the European labor market, describing over 3,000 occupations and nearly 14,000 skills in 28 languages.[1] Since its introduction in 2017, ESCO has attracted diverse stakeholders—including employment services, job portals, educational institutions, HR departments, and international organizations.[2]

Building on these insights, we release a large, publicly available dataset mapped to ESCO occupations to address the critical need for comprehensive and standardized resources in career path prediction. By leveraging ESCO's "common language" for occupations and skills, our dataset aims to foster research and development in this growing field, paving the way for more accurate career trajectory modeling. We document the steps involved in our dataset creation process to encourage the development and evaluation of customized datasets tailored to real-world applications, ultimately promoting job mobility and fostering a more integrated and efficient labor market. Our contributions are:

- Introducing KARRIEREWEGE, a new large-scale career path prediction dataset consisting of over 500,000 career paths.[3]
- Mapping the dataset to the ESCO taxonomy, enhancing interoperability and facilitating research and real-world applications that utilize ESCO.
- Exploring data synthesis techniques by generating paraphrased titles and descriptions from the taxonomical dataset to address the real-world challenge of free text data.
- A reproduction study of Decorte et al. (2023)

[1] https://esco.ec.europa.eu/en/about-esco/what-esco

[2] https://esco.ec.europa.eu/en/about-esco/esco-stakeholders

[3] Karrierewege: https://huggingface.co/datasets/ElenaSenger/Karrierewege
Karrierewege plus: https://huggingface.co/datasets/ElenaSenger/Karrierewege_plus

Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pages 533–545
January 19–24, 2025. ©2025 Association for Computational Linguistics

to compare results on their benchmark dataset with the newly introduced KARRIEREWEGE and KARRIEREWEGE+ datasets.

2 Related Work

2.1 Datasets for Career Path Prediction

In the literature on machine learning-based career path prediction, most prior work uses large *non-public* datasets, typically sourced from major career portals such as LinkedIn (Li et al., 2017; Cerilla et al., 2023), Randstad (Schellingerhout et al., 2022), or Zippia (Vafa et al., 2024). As publicly available datasets, survey data is a popular choice (Chang et al., 2019; Vafa et al., 2024; Du et al., 2024) – see Table 1. But survey data is typically relatively small, or does not track the same individuals over a longer time span. For example, the Current Population Survey—a national U.S. labor force survey used in Chang et al. (2019)—has a panel of 54,000 respondents per year but contains a person's occupation for only two consecutive years. Other surveys (Vafa et al., 2024; Du et al., 2024) are relatively small with sizes around 9-12k respondents. A small publicly available dataset is introduced by Decorte et al. (2023). It is created using a Kaggle dataset of 2,482 anonymized English resumes. All occupations are linked to ESCO (version 1.1.2). The dataset includes both self-written job titles and synthetic descriptions (grounded on resumes), as well as standardized ESCO titles. Inspired by Decorte et al. (2023), we use their SOTA approach and compare results to their smaller dataset (see further Section 5).

Dataset	Paper	Size
Nat. Longitudinal Survey of Youth 1979	Vafa et al. (2024), Du et al. (2024)	1,200
Nat. Longitudinal Survey of Youth 1997	Vafa et al. (2024), Du et al. (2024)	9,000
Panel Study of Income Dynamics	Vafa et al. (2024),Du et al. (2024)	12,000
Current Population Survey*	Chang et al. (2019)	54,000
DECORTE	Decorte et al. (2023)	2,000
KARRIEREWEGE	our paper	500,000
KARRIEREWEGE+	our paper	100,000

Table 1: Summary of datasets used in various studies. *Only data for two consecutive years per person.

2.2 Methods for Synthetic Data Generation and LLMs in Occupations

We source the original raw data from the German Employment Agency, but it only includes standardized job titles and descriptions. To make the model more applicable to real-world scenarios, where resumes often use varied, paraphrased job titles, we generated synthetic training data. This allows for more robust career path prediction models that can handle the complexities of free-text inputs.

Off-the-shelf (non-fine-tuned) large language models (LLMs) have been successfully applied to paraphrasing tasks across various domains (Jayawardena and Yapa, 2024) and have also been used to generate synthetic data in the job market, such as to create job vacancies (Li et al., 2023; Magron et al., 2024). Their effectiveness in representing occupations likely stems from the extensive training of LLMs on diverse sources of data, including occupational data, labor market news and job-related texts (Du et al., 2024). This demonstrated success in the occupational domain supports our approach of leveraging LLMs for synthesizing training data by paraphrasing job titles and generating corresponding descriptions.

3 KARRIEREWEGE

To create a large and diverse dataset for career path prediction, we sourced the data from anonymized resumes provided by the German employment agency as a basis (see Figure 1 for the dataset creation process).[4] This dataset encompasses resumes from individuals seeking employment across all industries. We note that despite of its size, the resulting dataset may still be biased—it possibly contains more resumes from industries with lower demand (where individuals are more inclined to register as unemployed) than from high-demand industries, where unemployment registration is less common. Additionally, since all resumes are from individuals seeking employment in Germany, there is a cultural bias towards that region.

3.1 Mapping to ESCO

Due to restrictions preventing the direct publication of these anonymized CVs, and in recognition of the widespread adoption of the ESCO framework, we manually mapped occupations from the German resumes to their equivalents in the German ESCO taxonomy (version 1.2.0). This mapping ensures compatibility with the widely adopted ESCO framework, enriches the dataset with additional attributes like skills and job descriptions, and enables accessibility in 28 languages. For consistency with previous work, we use the English ESCO attributes in this paper. Yet, the published dataset can be converted to any of the other languages via the unique

[4] https://www.arbeitsagentur.de/bewerberboerse/

id	order	ESCO_title	ESCO_description	new_title_oc	new_description_oc	new_title_cp	new_description_cp
0	0	Medical laboratory manager	Medical laboratory managers oversee ...	Quality Assurance Specialist	The Quality Assurance Specialist is ...	Laboratory Director	Laboratory Director: Oversees ...
0	1	Environmental health inspector	Environmental health inspectors carry ...	Pollution Control Specialist	The Pollution Control Specialist is responsible ...	Environmental Safety Specialist	Environmental Safety Specialist: Conducts ...
0	2	Environmental health inspector	Environmental health inspectors carry ...	Environmental Compliance Officer	Environmental Compliance Officer: Conducts ...	Environmental Safety Specialist	Environmental Safety Specialist: Conducts ...
1	0	Food service worker	Food service workers prepare food and ...	Concession Stand Staff	Operate a concession stand, selling ...	Culinary Service Provider	Culinary Service Provider: Provides expert culinary ...
1	1	Dietitian	Dietitians assess specific nutritional ...	Eating Disorder Specialist	The Eating Disorder Specialist provides ...	Nutritionist	Nutritionist: Helps people develop healthy ...

Table 2: Example entries from the KARRIEREWEGE+ validation dataset. Rows sharing the same *id* refer to work experiences of the same person, with *order* indicating the sequence. Titles and descriptions with the *_oc* suffix are synthesized per occupation, while those with *_cp* are synthesized per career path.

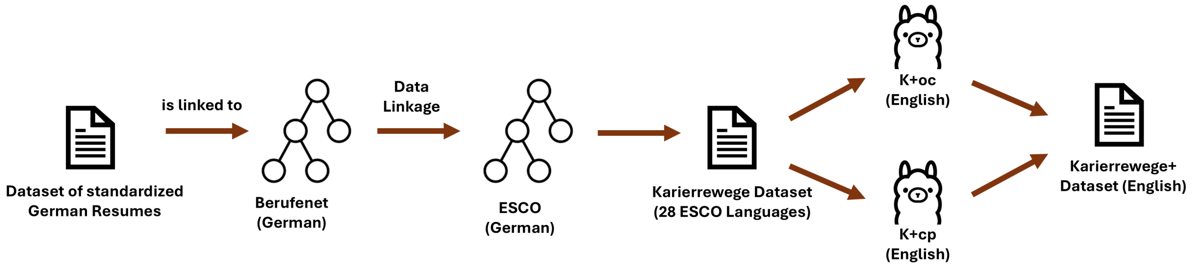
Dataset of standardized German Resumes
is linked to
Berufenet (German)
Data Linkage
ESCO (German)
Karierrewege Dataset (28 ESCO Languages)
K+oc (English)
K+cp (English)
Karierrewege+ Dataset (English)

Figure 1: Steps necessary to create KARRIEREWEGE and KARRIEREWEGE+ datasets. Titles and descriptions with the *_oc* suffix are synthesized per occupation, while those with *_cp* are synthesized per career path.

identifiers provided by ESCO.

The raw dataset contains standardized occupational titles from the German Berufenet taxonomy (version 2020).[5] Both Berufenet and ESCO occupations are mapped to ISCO-08 codes. However, due to broad categorization inherent in ISCO-08, multiple occupations share the same code, making it unsuitable for direct one-to-one mapping. Therefore, to link Berufenet and ESCO, we experimented with three methods as outlined next.

The first method involved embedding similarity, using the `distiluse-base-multilingual` model to calculate semantic similarity based on job titles and descriptions. The second method utilized the ESCO API, which returned less accurate mappings due to incomplete queries and higher-level job titles. The third method involved GPT models, specifically `GPT-3.5` and `GPT-4o-mini` for ranking mappings. Overall, `GPT-4o-mini` performed best, particularly when using English prompts, achieving around 60% correct mappings (see Table 3). However, ultimately, none of these approaches consistently produced satisfactory results. Hence, we used them only to speed up the manual linkage process conducted by a trained assistant and one of the authors.

Method	% Correct Links
Embedding Similarity Title	51.1
Embedding Similarity Description	51.2
ESCO API	30.7
GPT 3.5 DE Prompt	42.6
GPT 3.5 EN Prompt	52.9
GPT 4o mini DE Prompt	59.4
GPT 4o mini EN Prompt	60.4

Table 3: Percentage of correct links per method for mapping ESCO and Berufenet occupations.

3.2 Filtering the Data

We excluded all resumes with missing entries in the work history field and kept only those resumes with more than one and less than thirty work experiences. We also kept only resumes with a change of occupation in their career history, as we are particularly interested in learning and predicting these. Additionally, we excluded resumes that contained rare occupations (less than 10 times in the dataset). This resulted in 568,888 resumes.

4 KARRIEREWEGE+: Synthesized Data

4.1 Generating Free-Text Job Titles

To generate free-text data, we use two data syntheses methods:

KARRIEREWEGE+oc In the first approach, we use `LLAMA 3.1 8b` to generate seven alternative titles for each ESCO occupation title (K+oc). The

[5] `https://web.arbeitsagentur.de/berufenet/`

choice of seven paraphrased titles was based on empirical observations indicating that generating more than seven titles often resulted in lower quality titles. For each paraphrased title, we additionally generated a corresponding job description using the same model. The underlying hypothesis for this method is that the paraphrased titles remain closely related to the original titles while being sufficiently distinct from other ESCO titles.

KARRIEREWEGE+cp In the second approach, we directly synthesized the entire sequence of titles of a career path (K+cp). The hypothesis guiding this approach is that providing the model with the context of previous and subsequent occupation titles enables it to generate more appropriate and contextually relevant paraphrased titles. This method aims to achieve higher diversity by paraphrasing each ESCO occupation title more frequently, thereby introducing slight variations and increasing the richness of the dataset. Similar to the first approach, a corresponding job description was generated for each paraphrased title. The language model used was again the `LLAMA 3.1 8b` model. Due to the computational intensity of synthesizing individual career paths, we limited this approach to a random subset of 100,000 resumes. To maintain comparability between the two synthesis methods, we restricted the first synthesizing approach to the same number of resumes. All prompts are provided in Appendix B.

4.2 Evaluating the Quality of Paraphrased Titles and Career Paths

4.2.1 Quantitative Analysis

To evaluate the quality of the paraphrased job titles and descriptions, we followed best practices recommended by van der Lee et al. (2019) for paraphrase evaluation, i.e., to use well-defined evaluation criteria, avoid the use of smaller scales in rating (e.g., 2-point or 3-point Likert scale), employing a within-subjects design (where evaluators reviewed outputs from all systems), randomized orderings to mitigate bias from order effects and complement subjective with objective measures to provide a comprehensive evaluation. Following Jayawardena and Yapa (2024), we used BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and a 5-point Likert scale to evaluate the quality of the paraphrased job titles. We overall assessed the generated paraphrases on four key dimensions:

- **Correctness**: Measures if paraphrased titles are valid job titles and distinct from the original. Scores range from 0 (invalid or identical titles) to 5 (all titles valid and distinct).
- **Semantic Similarity**: Evaluates how well paraphrased titles capture the meaning of the original titles. Scores range from 0 (low similarity) to 5 (high similarity).
- **Diversity**: Assesses variety in the paraphrased career paths with a score of 0 indicating repetition, while 5 reflects a wide range of titles.
- **Coherence**: Measures the logical coherence of the paraphrased titles with the career path, where 0 means titles do not form a logical progression, and 5 indicates a coherent sequence.

To evaluate these dimensions, we manually labeled 100 resumes. One author, unaware of which synthesis method was used, manually evaluated each resume on the four dimensions. To further validate our findings, we used `GPT-4o mini` to evaluate the same metrics, after checking the alignment on the 100 manually labeled samples. We experimented with two prompt versions: one where the model was prompted once for all metrics, and another where the model was prompted for each metric individually (see the Appendix C for the prompts). Following the best practice of Thakur et al. (2024), we used Cohen's kappa as a measure of alignment. Cohen's kappa and the mean values for each metric indicated that prompting the model for all metrics at once resulted in closer alignment with human judgments. In general, Cohen's kappa values were relatively low, particularly for coherence, but showed stronger alignment for less subjective metrics like diversity and correctness (see Appendix D for detailed scores). The human and LLM scores revealed that K+cp achieved higher mean scores in correctness, semantic similarity, and coherence compared to K+oc. However, the K+oc outperformed K+cp in terms of diversity. These results are consistent with our expectations: the K+cp processes the entire career path, allowing for more coherent title generation, while the K+oc tends to produce greater diversity since it randomly selects from seven paraphrased options for each occupation. Overall, the Likert scale scores suggest that the K+cp yields higher-quality paraphrases.

As objective complementary measures, we used BLEU and ROUGE-L, comparing sequences of job titles and descriptions across entire career paths. For BLEU, we applied a smoothed score to account for low n-gram overlaps. In both metrics,

Category	Original Job Title	Paraphrased Job Title
1. Increasing Professionalism	Cashier Mover Bricklayer	Retail Sales Associate Professional Mover Building Specialist
2. Inaccurate or Non-existent Titles	Technical Communicator Financial Broker Bicycle Courier Printed Circuit Board Assembler	User Experience Wealth Manager: Helping clients... On-the-Go Logistics Professional PCB
3. Semantic Mismatch	City Councillor House Sitter Food Production Operator Foster Care Support Worker	Urban Planner Home Care Provider Production Line Worker Support Worker
4. Career Path as Story	Hairdresser → Sales Account Manager → Commercial Sales Rep → Hairdresser	Beauty Professional → Business Developer → Sales Specialist → Salon Owner/Operator
	Vehicle Cleaner → Factory Hand → Metal Sawing Machine Operator	Construction Laborer → Flooring Installer → Floor Covering Technician

Table 4: Examples of qualitative analysis of paraphrased job titles

K+cp consistently achieved slightly higher values, indicating better lexical similarity and sequence alignment with the original labels. However, the overall low scores suggest notable differences between both methods and the original career path (see Appendix D for detailed scores).

4.2.2 Qualitative Analysis

The paraphrased job titles reveal several interesting and distinct trends and errors. One common trend is the enhancement of professionalism, where paraphrased titles elevate the perceived professionalism of the original job titles, like "Bricklayer" becomes "Building Specialist". Another issue is the introduction of inaccurate or non-existent job titles, such as "On-the-Go Logistics Professional" where the paraphrasing becomes overly creative and diverges from widely recognized titles. Semantic mismatches also occur, for instance, when "city councillor," a political position, is paraphrased as "urban planner," a technical role focused on infrastructure. Lastly, when using the K+cp, the paraphrasing often constructs cohesive career paths, showing clear progression and skills development over time. However, when paraphrasing individual occupations without considering the full career path, this cohesive narrative is lost. Examples of these patterns are presented in Table 4.

5 Dataset Statistics and Comparison

To ensure the practical utility of KARRIEREWEGE for real-world career path prediction, we present key statistics on the dataset's characteristics—including the number of resumes, the distribution and diversity of ESCO occupations and industries, and statistics on synthesized job titles and descriptions that reflect the complexity of free-text inputs. These insights demonstrate the dataset's comprehensiveness and its suitability for advancing both academic research and industrial applications.

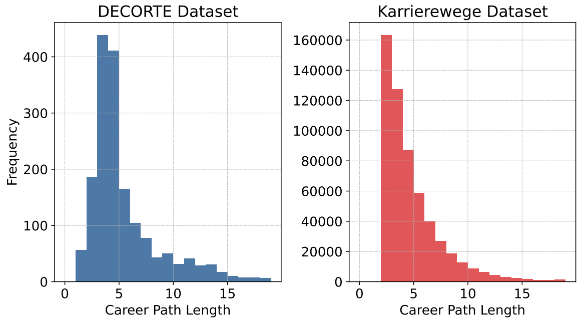

Figure 2: Work experiences per resume for the KARRIEREWEGE and DECORTE dataset.

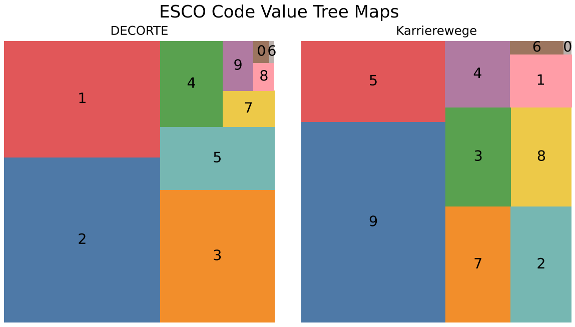

Figure 3: Tree maps on ESCO codes with one digit.

On the number and average length of resumes, the KARRIEREWEGE dataset (568,888 resumes) contains a higher proportion of resumes with fewer work experiences compared to DECORTE (2,482 resumes), which typically includes five experiences per resume (see Figure 2).

While on the distribution and diversity of occupations, KARRIEREWEGE features 1,295 unique

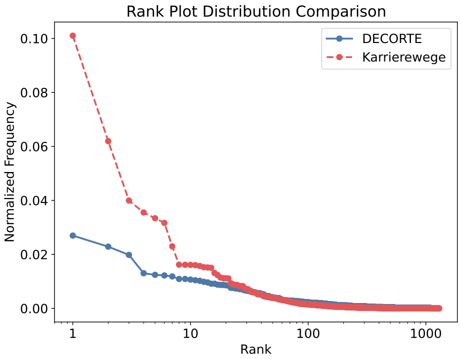

Figure 4: Rank plot of normalized frequencies of ESCO codes with full digits.

ESCO occupations, whereas DECORTE has 1,102. If aggregated by industry using ESCO taxonomy codes, as shown in Figure 3, KARRIEREWEGE covers broader economic sectors comprehensively, such as Elementary Occupations (Sector 9) and Service Workers (Sector 5), while DECORTE is concentrated in knowledge workers and managerial activities (Sectors 2 and 1). Moreover, a rank distribution plot in Figure 4 of ESCO occupations further highlights these distinctions: KARRIEREWEGE's most frequent occupations are in Sectors 9, 5, and 8, while DECORTE's are concentrated in Sectors 2 and 1. Nevertheless, despite these differences in the top occupations, both datasets show similar relative coverage of less frequent occupations, as clearly shown in the tails of both rank distributions. When considering absolute frequencies, however, KARRIEREWEGE exhibits broader overall sector coverage, even within Sectors 1 and 2. We provide a throughout presentation of these frequencies in Appendix H and the full first level ESCO classification names in Appendix G for completeness.

Regarding the statistics on the synthetic data, Figure 5 reveals that the lengths of job titles and descriptions differ depending on the data synthetization method, K+oc or K+cp. Though differences in job title lengths are minimal, apart from some outliers; K+oc generates consistently longer job descriptions (up to 800 characters), while K+cp produces shorter descriptions, with most under 400 characters. In comparison, ESCO descriptions have comparable length to the K+cp descriptions, while being in general also shorter than the ones generated with method K+oc.

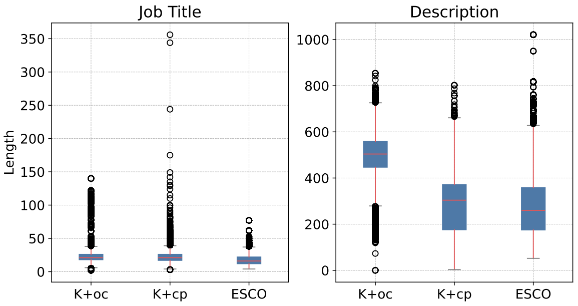

Figure 5: Length of generated job titles and job descriptions with both strategies K+oc and K+cp in comparison with ESCO.

6 Benchmark Baseline

To showcase how well KARRIEREWEGE supports existing models in realistic career prediction, we adapted a SOTA approach (Decorte et al., 2023) to benchmark its performance comprehensively. We want to showcase the effect of the dataset size and the robustness of models trained on KARRIEREWEGE+ in comparison to the smaller, less diverse DECORTE benchmark dataset.

6.1 Method

We follow the scheme called "LAST" in Decorte et al. (2023), and fine-tune a `all-mpnet-base-v2` sentence-transformer model using contrastive representation learning on pairs of text documents: one for the career path $ex_1, \cdots, ex_N$ and another for the ESCO occupation occ_N. In Decorte et al. (2023), each work experience ex_i in a career path is represented as:

```
role: <title in free-form text>
description: <description in free-form text>
```

and the career path document is composed by concatenating them with a separator token. In our adaptation, while the career document is composed similarly, a career experience ex_i is represented as:

```
esco role: <esco occupation title>
description: <esco occupation description>
```

In turn, on both approaches, the occupation document is structured always as the latter and contains data from ESCO occupation occ_N.

Finally, a linear transformation T is learned by minimizing the least squares error between transformed representations of career paths $ex_1, \cdots, ex_{N-1}$ and representations of their next ESCO occupations occ_N. Therefore, a career path prediction over the next occupation is achieved by

Metric	MRR			R@5			R@10		
Train/ Test	**DECORTE**	**K+oc**	**K+cp**	**DECORTE**	**K+oc**	**K+cp**	**DECORTE**	**K+oc**	**K+cp**
DECORTE	**0.2427**	0.1339	0.1588	**0.3418**	0.2005	0.2302	**0.4151**	0.2669	0.3076
K+oc	0.1303	**0.4312**	0.3784	0.2164	**0.5340**	0.4899	0.3091	**0.6165**	0.5784
K+cp	0.1294	0.3685	**0.4281**	0.2186	0.4693	**0.5280**	0.3235	0.5566	**0.6065**

Table 5: Cross evaluation results for MRR, R@5, and R@10 across free-text datasets.

the scoring function naturally induced by the cosine similarity between a transformed career path representation and all ESCO occupation representations. While Decorte et al. (2023) include further an ESCO skill overlap component in the scoring function, we opt to leave this out, in order to better measure the impact of only using ESCO data and synthetic free text data in our experiments. Our full experimental setup is given in Appendix E.

6.2 Results

To better understand how training data size impacts performance, we experimented with multiple dataset sizes, allowing us to assess trends in model improvement across varying scales. Models trained on Karrierewege consistently achieve higher scores compared to those trained on DECORTE, even when the dataset sizes are identical (see Table 6). This performance advantage cannot be solely attributed to validation and test set overlap, as the overlap remains negligible, or even minimal for Karrierewege+cp (see Table 10 in the Appendix). This suggests that other patterns in the data, such as the more coherent career paths, might be contributing to the improved results. Across Karrierewege, K+oc, and K+cp, a clear performance improvement is observed with the use of larger training datasets. Notably, the performance increase is most pronounced when scaling from the 2k dataset to the 100k dataset, highlighting the significant impact of additional data. However, once the dataset size reaches a substantial volume, such as 100k, the performance gains taper off, as evidenced in the smaller improvement observed when scaling from 100k to 500k in Karrierewege.

Evaluating across datasets and synthesis methods shows that models trained on Karrierewege+ datasets also performed well when tested on DECORTE, indicating that the paraphrased career paths generalize effectively across different datasets (see Table 5). This strong performance suggests that the paraphrased data captures underlying patterns and relationships between job titles, making it adaptable across various contexts.

Dataset	MRR	R@5	R@10
DECORTE 2k	0.2427	0.3418	0.4151
K+oc 2k	0.3846	0.4779	0.5423
K+oc 100k	**0.4312**	**0.5340**	**0.6165**
K+cp 2k	0.3702	0.4754	0.5568
K+cp 100k	**0.4281**	**0.5280**	**0.6065**
DECORTE ESCO 2k	0.2084	0.2813	0.3418
Karrierewege 2k	0.4232	0.5146	0.5636
Karrierewege 100k	0.4775	0.5671	0.6317
Karrierewege 500k	**0.4867**	**0.5713**	**0.6347**

Table 6: Results for different free-text and standardized resume datasets with their approximate size.

Notably, models trained on K+cp datasets generalize better than models trained on K+oc, further supporting the idea that coherent career paths play a role in improving model performance.

7 Conclusions and Further Research

We introduced Karrierewege and Karrierewege+, large-scale, publicly available datasets for career path prediction. By linking the datasets to the ESCO taxonomy and synthesizing paraphrased job titles and descriptions, we addressed the challenge of predicting career trajectories from the free-text inputs typically found in resumes. Our results demonstrate that models trained on the Karrierewege datasets, particularly the Karrierewege+cp variant, perform well on free-text data, underscoring the importance of data diversity and richness for accurate career path prediction.

While these datasets provide a strong foundation for model training and evaluation, future work could focus on expanding their scope to include more regions, and languages, further enhancing their applicability to global career path prediction. Addressing challenges such as cross-industry career transitions could also improve model robustness and generalizability.

Ethical Considerations

The use of large-scale datasets like Karrierewege carries the risk of bias amplification if the dataset overrepresents certain industries, job levels, or demographics. Biases in the dataset could inadvertently lead to discriminatory predictions, particularly when applied to automated decision-making tools used by recruiters or employment agencies. Furthermore, synthesizing data to augment or enhance the dataset introduces additional risks, as the assumptions made by the underlying models may reflect or amplify existing biases. This could result in inaccurate or skewed descriptions of career trajectories, reinforcing stereotypes or marginalizing underrepresented groups.

To address these challenges, it is crucial to continuously monitor and mitigate biases that may emerge both in the original dataset and in any synthesized data. Strategies such as bias audits, fairness metrics, and diversification of training data sources should be implemented to ensure equitable model predictions.

Acknowledgements

We thank the reviewers for their insightful feedback. We thank Jan-Peter Bergmann for computing infrastructure support and advice. ES and YC acknowledges financial support with funds provided by the German Federal Ministry for Economic Affairs and Climate Action due to an enactment of the German Bundestag under grant 46SKD127X (GENESIS). BP is supported by ERC Consolidator Grant DIALECT 101043235.

References

Micaela Tayoto Cerilla, Aaron Santillan, Carl John Vinas, and Michael B. Dela Fuente. 2023. Career path modeling and recommendations with linkedin career data and predicted salary estimations. In *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023*. OpenReview.net.

Li-Te Chang, Lisa Simon, Karthik Rajkumar, and Susan Athey. 2019. A bayesian approach to predicting occupational transitions.

Jens-Joris Decorte, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Career path prediction using resume representation learning and skill-based matching. In *RECSYS in HR 2023 : the 3rd Workshop on Recommender Systems for Human Resources (RecSys in HR 2023), Proceedings*, volume 3490, page 9. CEUR.

Tianyu Du, Ayush Kanodia, Herman Brunborg, Keyon Vafa, and Susan Athey. 2024. Labor-llm: Language-based occupational representations with large language models. *Preprint*, arXiv:2406.17972.

Lasal Jayawardena and Prasan Yapa. 2024. Parafusion: A large-scale llm-driven english paraphrase dataset infused with high-quality lexical and syntactic diversity. In *Artificial Intelligence and Big Data*, AIBD, page 219–238. Academy & Industry Research Collaboration Center.

Liangyue Li, How Jing, Hanghang Tong, Jaewon Yang, Qi He, and Bee-Chung Chen. 2017. Nemo: Next career move prediction with contextual embedding. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 505–513, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Nan Li, Bo Kang, and Tijl De Bie. 2023. Llm4jobs: Unsupervised occupation extraction and standardization leveraging large language models. *Preprint*, arXiv:2309.09708.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Antoine Magron, Anna Dai, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. 2024. JobSkape: A framework for generating synthetic job postings to enhance skill matching. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 43–58, St. Julian's, Malta. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Roan Schellingerhout, Volodymyr Medentsiy, and Maarten Marx. 2022. Explainable career path predictions using neural models. In *HR@ RecSys*.

R. Shreyas, C. Praveen, S. Shreyas, and Y.C. Kiran. 2024. A literature survey on ai driven career path prediction. *International Journal of Advanced Research in Science, Communication and Technology*, pages 578–582.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *Preprint*, arXiv:2406.12624.

Keyon Vafa, Emil Palikot, Tianyu Du, Ayush Kanodia, Susan Athey, and David M. Blei. 2024. Career: A foundation model for labor sequence data. *Preprint*, arXiv:2202.08370.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

A Prompts for linkage

To address the limited context window of the GPT models, we used ISCO-08 codes as a filtering mechanism to narrow down potential occupation matches between Berufenet and ESCO. By applying the ISCO code, the number of candidate pairs was significantly reduced, allowing them to fit within the GPT models' context window. In some cases, however, better matches were found under different ISCO codes. Therefore, ISCO filtering was only applied when necessary to reduce the number of matches. Without this, each Berufenet occupation could potentially match up to 3,039 ESCO occupations.

A.1 English Prompt

```
###CONTEXT###
You are a specialist in matching
    occupations with their ESCO labels.
    I have tried 3 different methods,
    and each method has a different
    prediction for the label. You will
    be presented with the occupation
    title, the 3 different predictions,
    and the set of candidate labels. The
     language is German.

###INSTRUCTION###
Choose the most appropriate label for an
     occupation title.
Return a JSON object with the job title
    and the selected label.

###DATA###
occupation_title: {occupation}
pred_1: {pred_1}
pred_2: {pred_2}
pred_3: {pred_3}
all_candidates: {candidates}

###OUTPUT###
```

A.2 German Prompt

```
###CONTEXT###
Du bist ein Spezialist f r das Matching
     von Berufen mit ihren ESCO-Labels.
    Ich habe 3 verschiedene Methoden
    ausprobiert, und jede Methode hat
    eine andere Vorhersage f r das
    Label. Dir werden die
    Berufsbezeichnung, die 3
    verschiedenen Vorhersagen und die
    Menge der in Frage kommenden
    Bezeichnungen vorgelegt. Die Sprache
     ist gemischt zwischen Deutsch.

###ANLEITUNG###
W hle das am besten geeignete Label
    f r eine Berufsbezeichnung.
Gib ein JSON-Objekt mit dem Berufstitel
    und dem gew hlten Label zur ck.

###DATA###
beruf_title: {occupation}
pred_1: {pred_1}
pred_2: {pred_2}
pred_3: {pred_3}
all_candidates: {candidates}

###OUTPUT###
```

B Prompts for title and description generation

B.1 Generation per career path

```
"""
    Please create a paraphrased version
        of the following career path: {
        job_list}.
    The length of the list should be the
         same as the original list.
        Please adhere to the format and
        do not add anything else.

    ###
    Format:

    'Career Path:

    [paraphrased title 1, paraphrased
        title 2, ...]'


    """
```

```
"""
    Please create a description for the
        following jobs: {job_list}.
    Please make it not longer than 4
        sentences. Please adhere to the
        format and do not add anything
        else.

    ###
    Format:

    'Descriptions:

    [Job title 1: Description of job 1,

    Job title 2: Description of job 2,

    ...]'
    """
```

B.2 Generation per occupation

```
"""
    Paraphrase the following occupation
        title: {job}.
    Please return a list of max 7
        alternative job titles.
    ###
    Example:
    occupation title: physiotherapist

    1. Physical Therapist
    2. Physiotherapy Specialist
    3. Rehabilitation Therapist
    4. Movement Therapist
    5. Injury Recovery Specialist
    6. Musculoskeletal Therapist
    7. Sports Medicine Therapist
    """
```

```
"""
    Please create a description for the
        following job: {job}.
    Please make it not longer than 4
        sentences. Please adhere to the
        format and do not add anything
        else.

    ###
    Format:

    'Description:

    Job Description of the occupation
        ...'
    """
```

C Prompts for LLM-evaluation

C.1 Prompt for all metrics at once

```
"""
    # CONTEXT
    A paraphrased career path should
        accurately reflect the skills
        and tasks of the original career
         path and it's job titles.
    You will evaluate the paraphrased
        career path on the following
        four dimensions:
    Correctness measures if the
        paraphrased titles are accurate
        representations of the original
        job titles.
    A paraphrased career path with high
        correctness means that all
        titles are job titles that
    exist in reality, but aren't just
        copies of
    the original titles. A low-
        correctness paraphrased career
        path may contain titles that are
         not job titles or are the same
        title as the original title.
    Semantic similarity assesses how
        well the paraphrased job titles
        captures the meaning of the
        original job titles in a career
        path.
    A high semantic similarity score
        means that the paraphrased
        titles accurately represent the
        skills and tasks of the original
         titles.
    Diversity measures how many unique
        job titles are present in the
        paraphrased career path.
    A high diversity score means that
        the paraphrased career path
        contains a wide range of job
        titles and does not contain the
        same title multiple times.
    Coherence evaluates how well the
        paraphrased job titles fit
        together within the paraphrased
        career path.
    A highly coherent paraphrased career
         path will have job titles that
        make sense together or form a
        logical progression.
.

    # INSTRUCTIONS
    You will be presented with a
        original career path and its
        paraphrased version.
    For each dimension, give the summary
         a score between 0 and 5. For
        correctness a score of 0 means
    the paraphrased career path contains
         only job titles that aren't
        real job titles or just copies
        of the original title, while a
        score of 5 means it
    provides only correct job titles.
    For semantic similarity a score of 0
         means that all paraphrased job
        titles are do not capture the
        meaning of the original titles
        very well, while a score of 5
        means that all paraphrased job
        titles accurately represent the
        skills and tasks of the original
         titles.
    For Diversity a score of 0 means
        that the paraphrased career path
         contains the same title
        multiple times, while a score of
         5 means that the paraphrased
        career path contains a wide
        range of job titles.
    For Coherence a score of 0 means
        that the paraphrased job titles
        do not make sense together or
        form a logical progression,
        while a score of 5 means that
        the paraphrased job titles fit
        together well in the career path
        .
    Output the Likert scores for each
        dimension as a json (key:
        dimension, value: likert-score).
    Do not add any explanation, answer
        only the Likert scores.
    Use the initial marker ```json and
        the final marker ``` to mark the
         json content.\n\n

    # EVALUATION MATERIALS
    ## Original career path
    {original_career_path\}

    ## paraphrased career path
    {paraphrased_career_path\}
```

```
    """
```

C.2 Prompts for each metric

```
"""
    # CONTEXT
    Correctness measures if the
        paraphrased titles are accurate
        representations of the original
        job titles.
    A paraphrased career path with high
        correctness means that all
        titles are job titles that
    exist in reality, but aren't just
        copies of
    the original titles. A low-
        correctness paraphrased career
        path may contain titles that are
         not job titles or are the same
        title as the original title.

    # INSTRUCTIONS
    You will be presented with a
        original career path and its
        paraphrased version.
    Give the summary a score between 0
        and 5.
    Zero means the paraphrased career
        path contains only job titles
        that aren't real job titles or
        just copies of the original
        title,
    while a score of 5 means it provides
         only correct job titles.
    Just answer with the Likert Score,
        no text please.

    # EVALUATION MATERIALS
    ## Original career path
    {original_career_path}

    ## paraphrased career path
    {paraphrased_career_path}
    """
```

```
"""
    # CONTEXT
    Semantic similarity assesses how
        well the paraphrased job titles
        captures the meaning of the
        original job titles in a career
        path.
    A high semantic similarity score
        means that the paraphrased
        titles accurately represent the
        skills and tasks of the original
         titles.

    # INSTRUCTIONS
    You will be presented with a
        original career path and its
        paraphrased version.
    Give the summary a score between 0
        and 5.
    Zero means that all paraphrased job
        titles are do not capture the
        meaning of the original titles
        very well,
    while a score of 5 means that all
        paraphrased job titles
        accurately represent the skills
        and tasks of the original titles
        .
    Just answer with the Likert Score,
        no text please.

    # EVALUATION MATERIALS
    ## Original career path
    {original_career_path}

    ## paraphrased career path
    {paraphrased_career_path}
    """
```

```
"""
    # CONTEXT
    Diversity measures how many unique
        job titles are present in the
        paraphrased career path.
    A high diversity score means that
        the paraphrased career path
        contains a wide range of job
        titles and does not contain the
        same title multiple times.

    # INSTRUCTIONS
    You will be presented with a
        paraphrased career path.
    Give the summary a score between 0
        and 5.
    Zero means that the paraphrased
        career path contains the same
        title multiple times,
    while a score of 5 means that the
        paraphrased career path contains
         a wide range of job titles.
    Just answer with the Likert Score,
        no text please.

    ## paraphrased career path
    {paraphrased_career_path}
    """
```

```
"""
    # CONTEXT
    Coherence evaluates how well the
        paraphrased job titles fit
        together in the career path.
    A highly coherent paraphrased career
         path will have job titles that
        make sense together or form a
        logical progression.

    # INSTRUCTIONS
    You will be presented with a
        paraphrased career path.
    Give the summary a score between 0
        and 5.
    Zero means that the paraphrased job
        titles do not make sense
        together or form a logical
        progression,
    while a score of 5 means that the
        paraphrased job titles fit
        together well in the career path
        .
```

Metric	gpt-4o-mini one Prompt	gpt-4o-mini
Correctness (CP)	0.367089	0.058639
Semantic Similarity (CP)	0.116162	-0.047056
Diversity (CP)	0.312162	0.264043
Coherence (CP)	0.082716	-0.013099
Correctness (Free)	0.082840	0.100846
Semantic Similarity (Free)	0.022131	-0.047251
Diversity (Free)	0.389831	0.333567
Coherence (Free)	0.022483	-0.025326

Table 7: Comparison of Kappa scores between LLM with one prompt and LLM with one prompt per metric.

```
    Just answer with the Likert Score,
        no text please.

    ## paraphrased career path
    {paraphrased_career_path}
    """
```

D Alignment Scores

Table 9 shows the Likert scale score for the human labeling as well as the `gpt-4o-mini` results. Table 7 compares Kappa scores between `gpt-4o-mini` with one prompt and `gpt-4o-mini` with one prompt per metric. Table 8 presents BLEU and ROUGE-L scores for job titles and descriptions.

Metric	Karrierewege+cp	Karrierewege+oc
ROUGE-L Score (Job Titles)	0.1618	0.1592
ROUGE-L Score (Job Descriptions)	0.0426	0.0233
BLEU Score (Job Titles)	0.0005	0.0002
BLEU Score (Job Descriptions)	0.0005	0.0003

Table 8: BLEU and ROUGE-L scores for job titles and descriptions.

E Experimental Setup

Following the recipe in (Decorte et al., 2023), we fine tune the model using Multiple Negatives Ranking Loss (MNRL), in-batch negatives and augmented career path data with all possible sub-paths of minimum length 2. This augmentation is applied after the data split. The fine-tuning process is conducted using a batch size of 16, a learning rate of $2e-5$, for up to 1 epoch for the large and 2 epochs for the small datasets, with evaluation every 1% of steps based on validation loss. The best-performing model is saved based on these evaluations.

F Overlap Test and Validation Dataset

Table 10 shows the overlap between validation and test splits for various datasets.

G First level of the ESCO classification

Table 11 shows the first categorization level of the ESCO classification with their respective codes.

H Absolute statistics of ESCO occupations distribution

As presented in Tables 12 and 13, Sectors 1 and 2 have almost two order of magnitude higher absolute frequencies in Karrierewege when compared to DECORTE data. By observing their absolute numbers, the massive difference between both datasets become clearer and highlights the potential of Karrierewege.

Metric	Manual Labeling (100)	gpt-4o-mini (100)	gpt-4o-mini one Prompt (100)	gpt-4o-mini one Prompt (all)
Mean Correctness (CP)	4.82	4.31	4.72	4.72
Mean Correctness (OCC)	4.72	4.26	4.74	4.70
Mean Semantic Similarity (CP)	4.50	3.41	4.02	4.08
Mean Semantic Similarity (OCC)	4.20	2.91	3.91	3.85
Mean Diversity (CP)	4.08	3.38	4.01	3.85
Mean Diversity (OCC)	4.43	3.88	4.33	4.22
Mean Coherence (CP)	4.01	1.58	4.04	4.12
Mean Coherence (OCC)	3.95	1.44	3.95	3.93

Table 9: Comparison of Mean Scores between Manual Labeling, the LLM-as-a-judge with one prompt per metric (gpt-4o-mini) and the LLM-as-a-judge with one one prompt for all metrics (gpt-4o-mini one Prompt).

Dataset	Length Validation	Length Test	Overlap	Overlap %
DECORTE ESCO	1,558	1,801	45	2.92%
DECORTE	1,558	1,801	0	0%
KARRIEREWEGE	667,404	658,012	56,101	8.53%
KARRIEREWEGE+oc	138,275	137,530	8,724	6.34%
KARRIEREWEGE+cp	138,275	137,530	139	0.10%

Table 10: Overlap between validation and test splits across various datasets. The percentage is calculated as the number of overlapping entries divided by the total size of the test split.

Code	First Level Occupation Category
0	Armed forces occupations
1	Managers
2	Professionals
3	Technicians and associate professionals
4	Clerical support workers
5	Service and sales workers
6	Skilled agricultural, forestry and fishery workers
7	Craft and related trades workers
8	Plant and machine operators and assemblers
9	Elementary occupations

Table 11: First level of the ESCO classification.

Code	Absolute Frequency
9	941544
5	379332
7	247981
2	230291
3	213148
8	194963
4	139843
1	106259
6	23574
0	3434

Table 12: Absolute frequency of first level ESCO occupations in KARRIEREWEGE.

Code	Absolute Frequency
2	2891
1	2036
3	1699
5	807
4	600
7	210
9	172
8	66
0	39
6	12

Table 13: Absolute frequency of first level ESCO occupations in DECORTE.