A Data-driven Analysis of the Flemish Job Market

Iman Johary, Raphael Romero, Alexandru C. Mara, and Tijl De Bie

Ghent University, Belgium

Abstract. Understanding the dynamics of the labour market is crucial for policymakers, employers, and job seekers. This paper presents an analysis of the Flemish labour market using a large dataset of resumes provided by VDAB, the public employment service in Flanders, Belgium. We utilize Large Language Models (LLMs) to process the resumes and extract structured data. Additionally, we employ sentence embedding models to map the structured resumes to standardized ESCO occupation codes. As a result, we obtain a dataset that contains real-world examples of occupations and transitions between them. We further analyse the dataset by investigating the relationship between education and job applications, examining the influence of career breaks on job transitions, exploring mobility within the job market, and analysing job stability across various sectors. We quantify the impact of university degrees on careers, visualize the effect of career breaks on job transitions, and analyse the mobility of individuals within the job market. As an additional contribution, we release an aggregated dataset that includes over 600,000 ESCO occupations extracted from more than 115,000 user resumes. This work contributes to a deeper understanding of labour market dynamics and provides valuable insights for data-driven decision-making in the labour market.

Keywords: Labour market analysis · Large Language Models · ESCO classification; datasers

1 Introduction

Understanding the complex dynamics of the labour market, —which are influenced by factors such as economic conditions, technological advancements, and social trends— is crucial for policymakers, employers, and job seekers[23]. For policymakers, a global view on labour market dynamics can aid in the development of new strategies to address skill shortages and unemployment[31]. For employers, it can lead to more efficient hiring processes and higher employee retention[20]. While for job seekers, a better understanding of the labour market can lead to more employment opportunities and fulfilling career paths[2]. These significant societal and economic needs, along with increased data availability and advancements in machine learning and data analytics, are creating an ideal environment for the development of new data-driven methods for the labour market[3]. By leveraging such methods we can extract valuable insights from job market data, identify trends and patterns, and predict future workforce demands.

The rapid expansion of digital technologies has led to a significant increase in job market data. This increase impacts publicly available information regarding job vacancies including descriptions, required skills, offered benefits, etc. At the same time, it concerns (often proprietary) job seeker information such as online profiles and resumes. Though less studied in practice, the latter is particularly interesting as it offers a complementary view of the labour market by highlighting existing skills, education, and experiences. This newly generated job market data, however, is often presented as unstructured textual information. The unstructured nature, combined with a high complexity and large volume results in significant challenges for traditional data analysis approaches[29]. On the other hand, recent developments in natural language understanding, particularly in Large Language Models (LLMs), are unlocking new avenues for processing these types of information. Unlike traditional analysis methods, LLMs are particularly well suited for processing complex unstructured data and deriving meaningful insights[10].

As shown by Li et al. [18], LLMs are also effective tools for mapping labour market data onto standardized occupation and skill taxonomies. These taxonomies, including ESCO[1], ANZSCO[25], and O*NET[12], are becoming key enablers of data-driven labour market analyses. The aforementioned standards define mappings from labour market data, such as job adverts and resumes, onto hierarchical structures that aggregate similar occupations. As a result, they facilitate statistical analysis, data exploration and inference in the context of the labour market. The main differences between these standards lie in their definitions of occupation similarities and their links to related skill and knowledge taxonomies.

Leading the development of multi-lingual occupation taxonomies cross-linked with skill and knowledge ontologies is the European Commission through the "European Skills, Competences, Qualifications, and Occupations (ESCO)" standard. The goal of ESCO is to create a unified understanding of the European labour market, provide a global view on occupations and skills and promote international mobility. The ESCO standard assigns numerical codes to occupations and skills and is becoming a cornerstone for labour market analysis. This standard is currently being used by public and private institutions as well as researchers across the EU and neighbouring countries [7][32].

In this paper, we leverage recent advances in LLM and sentence embedding approaches, together with the ESCO occupation taxonomy, to provide a unique data-driven analysis of the Flemish labour market. We base this analysis on a large dataset of anonymised resumes provided by VDAB, the regional public employment service in Flanders, Belgium. Our approach involves the transformation of raw resume data into structured information, mapping of this data onto the ESCO occupation codes and a subsequent analysis of job market trends and patterns. The use of ESCO as a main building block allows us to explore the labour market at different granularities and obtain a complete view of the relevant occupations and skills in Flanders.

As an additional contribution to labour market research, we release a large dataset of real-world ESCO occupation transitions[1]. This dataset is the result of analysing over 115.000 resumes and extracting more than 600.000 ESCO occupations. We make the complete list of occupations along with transition dates and the corresponding resume identifier for each transition publicly available. Furthermore, in anticipation of advancements in LLMs technology and the consequent enhancement of our data extraction pipeline, and with the prospect of incorporating new data, we are committed to regularly updating this dataset. We hope this data will foster further research in this area and provide deeper insights into job market dynamics.

The main contributions of this paper are as follows:

- We introduce a new LLM-based pipeline for extracting labour market transitions from raw resume data.
- We conduct a global analysis of the Flemish job market based on real-world job transitions.
- We release the dataset of job transitions to encourage further research in this area.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 details the data curation process from extracting structured information to data normalization and evaluation. Section 4 provides an in-depth analysis of job market trends and patterns. Finally, Section 5 concludes the paper and suggests future work.

2 Related Work

This section provides an overview of previous work on related topics including Named Entity Recognition (NER), Job Classification, and Labour Market Analysis (LMA).

2.1 Named Entity Recognition

Named Entity Recognition (NER) is a sub-task of information extraction that aims to identify specific entities in text and classify them into predefined categories. NER has been widely used in various applications, such as information extraction[35], question answering[21], language translation[19], and more [5][24][26]. It has also been utilized for extracting information from resumes [6][22][4]. These works involve segmenting resumes into smaller, semantically similar subsections and training different NER models for each subsection. These approaches, however, generally rely on supervised models which require a large amount of labeled data.

[1] The dataset is available at: https://tinyurl.com/4pfs8hys

Large language models (LLMs) [28][14], capable of performing a wide range of NLP tasks, have also shown promising results for NER [33][11]. While their performance in this domain still falls short of that of supervised models, LLMs have the added benefit of not requiring labeled data. These models perform well in NER tasks even without examples of the entities to be extracted, a technique known as zero-shot prompting[15]. Due to the dynamic nature of job market data and the large variance in quality between resumes, developing comprehensive NER models capable of extracting multiple types of entities is challenging. LLM-based approaches can fill this gap and yet, to the best of our knowledge, no LLM-based NER models have been proposed until now.

2.2 Job Classification

The classification of job postings into standard taxonomies such as ESCO is a fundamental task in labour market analysis. Traditionally, this classification was performed through supervised algorithms leveraging exclusively job title information [7,8,9]. In contrast, Varelas et al.[32] propose an ensemble of five supervised machine learning models that utilize both job titles and job descriptions. The resulting framework is used to classify Greek job postings into ESCO categories. These supervised models, however, require large amounts of labeled data for training, which is costly to produce and, thus, rarely available.

More recently, Li et al. [18] introduced a new model for unsupervised classification of job postings based on LLMs. This model combines job titles and job descriptions into vector representations which are compared to the representations of ESCO occupations to retrieve the appropriate codes. The model shows promising results and, unlike supervised approaches, does not require large amounts of labeled data. Contrary to the setup of Li et al., this work considers the classification of occupations extracted from user resumes. Effectively, this means a significantly lower data quality as titles and descriptions are user-generated in contrast to being generated by labour-market-trained HR managers and recruiters (who usually produce job postings). Moreover, in resumes, job descriptions are frequently omitted. Therefore, a dedicated approach for the task of ESCO-classification from user-generated occupation titles had to be developed for the purposes of the present paper.

2.3 Labour Market Analysis

Labour Market Analysis (LMA) or Labour Market Intelligence (LMI) is a field that aims to analyse and understand job market data by examining trends, identifying skill shortages, predicting future job demands, and developing strategies to address unemployment and skill gaps. Most work in this field focuses on the analysis of job postings which are abundant and publicly available [13][27][34][2]. However, analysing resumes can provide a different and complementary view of the job market and reveal real-world job transitions and career paths.

Several works analyse the job market using resumes. Aljohani et al. [3] analysed the job market in Saudi Arabia using big data methods to describe future market needs and proposed a job recommender based on skill matching. Other studies aim to improve the interpretability and explainability of recommendation systems for the job market [17][30]. Vankevich et al. [31] used online-sourced resumes to gain insights into the required competencies and skills in the job market and to forecast market dynamics. In this paper, we provide a dataset of ESCO occupations extracted from real-world resumes in a similar geographic area. This allows for a more comprehensive analysis of the local labour market and comparison with other regions.

3 Dataset Curation

This section first provides a formal problem definition and description of the unstructured dataset in Sec. 3.1. In Sec.3.2, it outlines the two-phases for converting unstructured dataset into Normalized dataset. In Sec. 3.3 it describes the experimental setup and, finally, in Sec. 3.4 it presents the experimental evaluation of the methods.

3.1 Problem Definition

In this paper, we focus on the following research questions:

1. How can we extract structured information related to work experiences and education from heavily anonymised resume data?
2. How does holding a tertiary degree influence the types of jobs individuals apply for?
3. How does a career break impact subsequent job opportunities, and does this effect differ for individuals with a tertiary degree?
4. In what ways do job transitions vary among different ESCO groups?
5. Which occupations do individuals tend to remain in for extended durations?

To address the first research question, we define the necessary variables and sets. Let D^u be the *unstructured* dataset, where each resume $d^u \in D^u$ is a text document containing information about the job seeker, including work experience, education, skills, and other relevant details. Our goal is to convert each resume d^u into *structured* data d^s. The structured data d^s can be defined as:

$$d^s = (E, Q)$$

Here, E is the set of work experiences extracted from d^u and Q is the set of qualifications extracted from d^u. Each work experience within E contains information like job title, company name, start date, end date, and place of work.

We define a function, denoted as $f(\cdot)$, which takes a job title as input and returns the corresponding ESCO code. Using the function $f(\cdot)$, we can define the set of *normalized* experiences, denoted as E^n, which consists of all experiences in E where any job title j is replaced by the corresponding ESCO code $f(j)$.

Finally, the normalized dataset D^n is the set of all d^n, where:

$$D^n = \{d^n \mid d^n = (E^n, Q)\}$$

The normalized dataset D^n is the data that will be used for further analysis and to answer research questions 2 to 5.

In Summary, the first research question will be tackled in two phases: first phase is converting unstructured data $d^u \in D^u$ to structured data d^s, while the second phase focuses on converting structured data d^s to normalized data $d^n \in D^n$ by mapping job titles to ESCO codes. These two phases are referred to as *extracting structured information* and *data normalization*, respectively. To fully understand these two phase, it is important to consider the input and output of our problem. In the remainder of this subsection, we will provide a more detailed description of the unstructured input data D^u, as well as an explanation of ESCO codes and their hierarchical taxonomy structure.

Data Overview To understand the variations in resumes across different sectors, backgrounds, and individuals, we conducted a thorough data analysis on an exclusive dataset consisting of 442,555 resumes from 385,052 individuals. These resumes were gathered by VDAB, the Flemish Public Employment Service (PES), as part of their mission to assist job seekers. Prior to analysis, all resumes underwent anonymisation by removing personal details such as names and addresses. Additionally, less common words were eliminated using Term Frequency-Inverse Document Frequency (TF-IDF) to set a threshold and avoid potential disclosure of sensitive information. This process inadvertently excluded some company names, job titles, or educational institutions. However, the dataset retains substantial information suitable for analysis.

The dataset is multilingual, containing 442,555 resumes, primarily in Dutch, English, and French. Specifically, 27,194 are in English, 404,611 in Dutch, and 10,750 in French. The resumes detail job seekers' work experience, education, skills, and other pertinent information.

ESCO Taxonomy The European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy is a comprehensive taxonomy used to categorize occupations based on skill level, specialization, and job requirements. It facilitates standardized comparisons of job markets across regions and countries, aiding governments, organizations, and researchers in analysing labour markets.

ESCO is structured hierarchically, starting with 10 broad groups at the first level, known as "ESCO groups". These groups encompass a wide range of general occupational categories. Below this top level, ESCO includes increasingly specific levels. The first four levels are general groupings of occupations. Second to Fourth levels provide more detailed groupings of occupations. For example, the fourth level includes 436 codes, representing specific job categories that are closely related in terms of skill and specialization. This level is referred to as "ESCO level 4". The ESCO taxonomy extends to the most detailed occupations, which are

Table 1. Overview of Prompt Structure

Role	Prompt Section	Goal	Notation
SYSTEM	Defining Role	Instructing the LLM to act as a Human Resource Assistant	-
	Explaining Input and Output	Clarifying the input format and the expected output	-
	Output Format	Detailing the expected output format	O_f
	Instructions	Providing additional instructions regarding the output	-
USER	Example Resume	Providing a one-shot example of the input	x_{1shot}
ASSISTANT	Extracted Information	Demonstrating the expected output of x_{1shot}	y_{1shot}
USER	unstructured Resume	Structuring the desired resume format	d^u

categorized from the fifth level onwards. The depth of these levels can vary, with some occupations classified as deeply as the seventh level. For example, a *software developer* (2512.4) is classified at level 5, while a specialized role like *import export specialist in electronic and telecommunications equipment* (3331.2.1.11) is classified at level 7. This paper refers to these levels as "ESCO code" and it consists of 3,039 codes.

This hierarchical structure of ESCO allows for precise and detailed classification, which is key for accurate labour market analysis and policy development.

3.2 Methods

This section describes the two necessary phase for answering first research question. The first phase is *extracting structured information* which involves the necessary steps to convert unstructured dataset $d^u \in D^u$ to structured data d^s. The second phase is *data normalization* which involves defining what the function $f(\cdot)$, which is the main component of the second phase.

Extracting Structured Information The first phase of converting unstructured dataset D^u into a normalized dataset D^n involves extracting structured data d^s from unstructured resumes d^u, which are written in various formats and have different levels of detail. The data is structured using the JSON format, making it machine-readable and suitable for further analysis. The structured data d^s consists of set of work experiences E and qualifications Q extracted from the resume d^u. Set of work experiences(E) are extracted to analyse job transitions within the dataset, while qualifications(Q) is extracted to examine the impact of having a tertiary degree on job trajectories. For each $e \in E$, we extract the job title, company name, start date, end date, and place of work.

Similarly, for $q \in Q$, we extract the degree, school name, start date, and end date.

LLMs were utilized to extract pertinent information from resumes. LLMs have the ability to identify complex relationships within lengthy texts[16] and can handle additional noise caused by pseudonymization, making them well-suited for our objectives. We provided the LLM with predefined instructions requesting the output in JSON format, in this way transforming d^u into d^s that can be processed by machines using classical data analysis pipelines.

Various strategies were explored to engineer an optimal and consistent prompt for the task. Given the extensive nature of resumes, it was necessary to use one-shot prompting to maintain the model's context length constraints, ensure satisfactory outcomes, and reduce hallucination. By refining the example prompt to address common errors encountered during zero-shot prompting, more consistent results were achieved. The experimentation involved diverse methodologies and structures for extracting structured information, ultimately leading to the realization that one-shot prompting, combined with JSON output formatting, yielded optimal results.

Table. 1 provides an overview of the prompt structure used for extracting structured information. The system message is divided into four sections: defining the role, explaining the input and output, specifying the output format, and providing instructions. The first user message includes a one-shot example of the input (x_{1shot}), and the assistant message demonstrates the expected output (y_{1shot}) for x_{ishot}. The second user message contains the unstructured resume (d^u), and in response, the LLM outputs the structured data (d^s) in the expected JSON format. The expected output format (O_f) is a JSON structure, which is commonly used to represent structured data. Fig. 1 provides a visual representation of the JSON structure that is expected from the LLM model. The structured data includes work experiences (E) and qualifications (Q), which are extracted from the resume. Work experiences encompass both work and internship details, while qualifications include information about education and certifications.

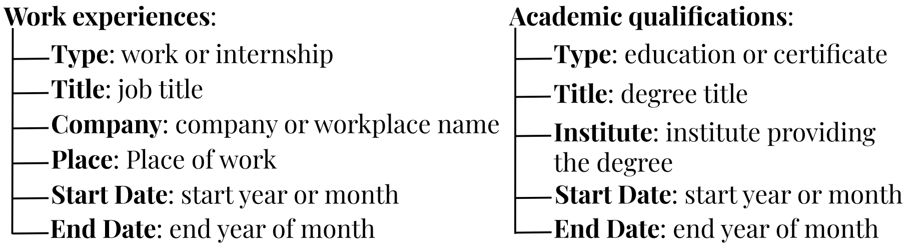
Work experiences:
Type: work or internship
Title: job title
Company: company or workplace name
Place: Place of work
Start Date: start year or month
End Date: end year of month
Academic qualifications:
Type: education or certificate
Title: degree title
Institute: institute providing the degree
Start Date: start year or month
End Date: end year of month

Fig. 1. Structured output format expected to be returned by the LLM.

Data Normalization In this section, we make use of the structured data d^s extracted from unstructured resumes d^u to map each job experience to its corresponding ESCO code using the function $f(\cdot)$. Inspired by the work of Li et al. [18], we employ a sentence embedding model for the mapping function $f(\cdot)$.

The mapping function $f(\cdot)$ is illustrated in Fig. 2. To begin, we employ a sentence embedding model to convert all available ESCO codes into embeddings, which are then stored in a vector database. These models are specifically selected for this task because they are highly proficient at capturing the semantic meaning of short sentences, making them well-suited for analyzing job titles. Next, we embed the query job title using the same sentence embedding model. We then calculate the cosine similarity between the query and all ESCO codes, and select the top ESCO codes with the highest similarity to the query.

Additionally, we employ various methods to further improve the results, as described in the subsequent subsections.

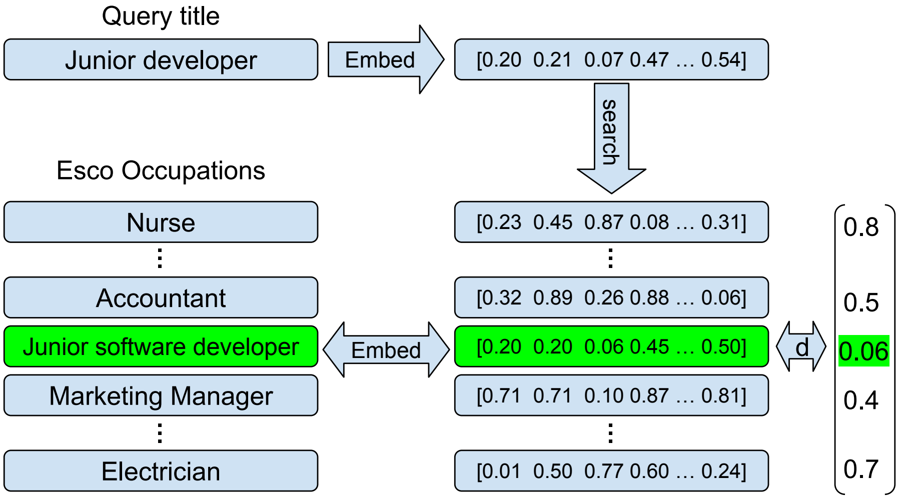

Fig. 2. High level diagram of how the function $f(\cdot)$ works. It takes a job title as input and returns the corresponding ESCO code.

Fine-tuning the Embedding Model. The choice of the sentence embedding model significantly impacts the results. However, these models are typically trained on general data and may not be specifically tailored to job market data. Fine-tuning allows the model to concentrate on the relevant aspects of job descriptions, resulting in improved performance of $f(\cdot)$.

The main hurdle in fine-tuning sentence embedding models lies in identifying the appropriate data for the task. To address this challenge, we leverage the ESCO occupation classification data, which provides a rich and structured source of job market information. By utilizing this data, we can effectively fine-tune our model to focus on job-specific content. To achieve this, contrastive learning, a self-supervised learning method, is utilized for fine-tuning. This method maximizes agreement between augmented views of the same instance while minimizing agreement between views of different instances.

In ESCO occupations, each code has a main label and possibly multiple alternative labels for the same job title, indicating their similarity. Contrastive learning is well-suited for this task, as it treats each code as a separate class. The main label acts as the centre of the class, while alternative labels represent other points in the class. This approach encourages the model to minimize the distance between job titles of the same class and maximize the distance between titles of different classes. For negative samples during contrastive learning, we utilize job titles that are not in the same ESCO group.

Heuristic methods. To further enhance our results, we employ heuristic methods alongside fine-tuning the model. These methods include translating titles to English, adding alternative labels to the database, replacing acronyms with their full versions, and reranking the results based on heuristics.

Given that most of our data was in Dutch and sentence embedding models are primarily trained on English data, we translated the query job titles to English as well as retaining the original language. Three vector databases were created, each representing the ESCO taxonomies of one of the languages in the dataset. Job titles were queried both in the original language and English. We then identified the closest match using a heuristic majority voting equation (Eq. 1), which significantly improved the results. Additionally, by adding alternative labels to the database, we increased the chances of correctly identifying the ESCO code for ambiguous job titles.

Sentence embedding models, although much smaller than large language models (LLMs), have far fewer parameters and are trained on significantly fewer tokens. Consequently, LLMs are capable of understanding a broader range of text. Sentence embedding models, however, struggle with uppercase data, leading to difficulties in recognizing acronyms. By replacing acronyms with their full versions, such as replacing "IT" with "Information Technology," we enhance the model's ability to accurately categorize job titles. We make the full list of acronyms available alongside the dataset on our online repository at https://tinyurl.com/4pfs8hys.

Finally, after incorporating alternative labels into the database and querying in two languages, multiple results were obtained for each query. To refine these results, we employed a reranking strategy designed to assign higher scores to ESCO codes with more alternative labels in the top results. Eq. 1 aims to decrease the similarity distance for codes that have a high percentage of alternative labels in the top results.

$$S^i = \sqrt{\frac{N^i_{alt} + 1}{R^i + 1}} S^i_{avg} \tag{1}$$

Where S^i represents the similarity score between the query and the ith ESCO code. N^i_{alt} denotes the number of alternative labels available for the ith ESCO code. R^i indicates the number of alternative labels that the ith ESCO code has in the top results. S^i_{avg} represents the average similarity score between the query and all alternatives of the ith ESCO code in the top results. To smoothen the effect of Eq1 and ensure consistency, we add one and take the square root. Since some ESCO codes have multiple alternative labels, we used an upper bound for N^i_{alt} and R^i to prevent the results from being biased. These values are defined as follows:

$$N^i_{alt} = \min(N^i_{alt}, N_{max}), \qquad\qquad R^i = \min(R^i, N^i_{alt})$$

Where N_{max} is a hyperparameter that is the set the upper bound for N^i_{alt} and R^i.

3.3 Experimental Setup

All experiments were conducted on a machine with 2 Nvidia A40 GPUs, 1 TB of RAM, and 48 CPU cores. We used the Huggingface Transformers library for the LLM model and the Sentence Transformer library for the sentence embedding model. We used the AdamW optimizer for fine-tuning the model and cosine similarity for calculating the similarity between the query and the ESCO codes.

We annotated two datasets for this study. The first dataset, referred to as the *Information Extraction Dataset*, consists of 200 resumes. From these resumes, we extracted various information such as job experiences, education, internships, and certifications. The annotation process, carried out by the authors with the help of close colleagues, took approximately 60 hours to complete. The annotated datasets for this study include the Information Extraction Dataset, which comprises 200 resumes. From these resumes, we extracted various information such as job experiences, education, internships, and certifications. The annotation process, conducted by the authors with the assistance of colleagues, took approximately 60 hours to complete. This dataset was used for validating and testing of extracting structured information phase. For validation, 10 percent of the dataset was randomly selected, while the remaining portion was used for testing purposes.

The second dataset that was annotated is the *ESCO Mapping Dataset* that was used for data normalization. We annotated 1,000 job experiences from the extracted data. Half of these job experiences were randomly selected, while the other half consisted of the most common job experiences. This ensured a representative sample of prevalent occupations. The annotation process for this dataset also took approximately 30 hours to complete and was conducted by the authors. We randomly selected half of the dataset for validation, and the rest was used for testing. The mapping function $f(\cdot)$ was then used to map the job titles that were not manually annotated by the researchers.

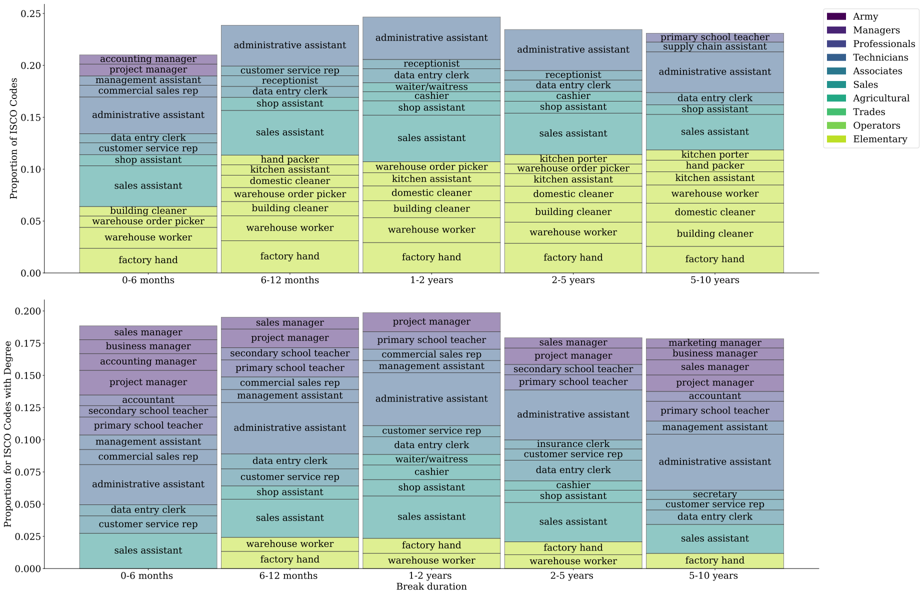

Fig. 3. This figure shows the first job after a career break for people with and without a tertiary degree. The upper part of the figure is for people without a tertiary degree, and the lower part is for people with a tertiary degree.

3.4 Results

Results are presented in three parts. The first part evaluates the performance of the extracting structured information phase. The second part evaluates the performance of the data normalization. The third part conducts an ablation study to assess the impact of different design choices on the model's performance.

Extracting Structured Information For the evaluation of this phase, we utilized the Information Extraction Dataset. We expected all JSON fields to be present in the model output, so any missing fields in the output will receive a similarity score of 0. To assess performance, we calculated similarity scores between the extracted fields and the ground truth. Each field was assigned a similarity score between 0 and 1. The average of these scores was then calculated for each resume. For text field comparisons, we employed fuzzy string matching using the FuzzyWuzzy library. For date fields, we used binary scores, assigning a score of 1 for exact matches and 0 for all other cases.

Table 2 presents the evaluation results. Accuracy refers to the average similarity scores for each resume. JSON accuracy represents the percentage of model outputs that can be converted to JSON format; outputs that cannot be converted are considered failures. Title accuracy denotes the average accuracy of all extracted titles, such as job titles and degree titles. We evaluated the Vicuna, Llama[28], and Mistral[14] models for extracting structured information.

Table 2. Comparison of models based on JSON accuracy, title accuracy, and overall accuracy.

Model	*JSON Acc*	*Title Acc*	*Overall Acc*
Vicuna-13b (one shot)	99.5	72.7	71.5
Vicuna-13b (zero shot)	100	71.7	57.9
Mistral-7b Instruct[14] (one shot)	80	68.6	63.2
Llama2-13b chat[28] (one shot)	64.5	72.2	53.0

Additionally, we assessed the impact of one-shot and zero-shot prompting on the models' performance. As illustrated in the table, the Vicuna model demonstrates superior performance in both JSON accuracy and overall accuracy. Furthermore, one-shot prompting significantly enhances the models' performance.

Given that these models require substantial computational power and considering the size of our dataset, we estimated that converting the entire dataset into structured data would require 500 GPU days with our current setup. Therefore, we limited our extraction to 115,454 resumes, which constitutes approximately 25 percent of the original dataset. We extracted educational details and job experiences to analyse the impact of a tertiary degree on job distribution. The extracted data was subsequently used for the remainder of the analysis.

Data Normalization For sentence embedding, we conducted a comprehensive analysis utilizing all 124 models available in the Sentence Transformer library. The ESCO Mapping Dataset was used for evaluating the performance of the models.

Table 3. Comparison of sentence embedding models.

Model	*ESCO level 4*	*ESCO code*	*ESCO group*
all-MiniLM-L6-v2	45.0	39.9	56.5
all-MiniLM-L12-v2	44.0	38.7	58.0
LaBSE	42.0	34.4	56.7
multi-qa-mpnet-base-dot-v1	41.0	34.4	57.0
paraphrase-multilingual-MiniLM-L12-v2	40.5	33.8	56.2

Table 3 presents the top 5 models for our task. It displays the average accuracy of the models in predicting the correct 4-digit ESCO code, ESCO code, and

ESCO group. From the table, we can observe that the all-MiniLM-L6-v2 model performs the best in terms of ESCO level 4 and ESCO group.

Ablation Study Finally, we evaluate the effect of different design choices on the performance of data normalization. We evaluate the effect of fine-tuning, heuristic methods, title translation, and alternative labels on the performance. For this evaluation, we used the ESCO Mapping Dataset. We observe the effect of each design choice on the performance of data normalization by removing that part.

Table 4. Comparison of different design choices on data normalization accuracy.

Model	***ESCO level 4***	***ESCO code***	***ESCO group***
Proposed method	46.3	41.5	58.3
W/o fine-tuning	45.0	39.9	56.5
W/o heuristic	41.2	35.4	53.4
W/o translation	38.4	32.8	52.2
W/o alternative labels	45.8	41.2	57.8
W/o smoothing	44.0	38.9	55.5

Table 4 highlights the impact of each design choice on the model's performance. Fine-tuning improved all metrics by approximately 2 percent. Heuristic methods significantly boosted performance by about 5 percent. Translating job titles to English, in addition to the original language, also had a substantial positive impact. Adding alternative labels to the database resulted in a performance increase of around 1 percent. Adding smoothing factors to the Eq1 improved the model's performance by approximately 2 percent.

4 Data Analysis

In Sec. 3, we discussed how we processed the data. In this section, we will derive insights from the data we have collected. We will address questions such as the nature of job transitions for different ESCO groups, the effect of a tertiary degree on job distribution, and the impact of career breaks on job transitions. Gaining these insights will enhance our understanding of the job market and may help policymakers, employers, and job seekers in making better decisions in the future.

Some of the following figures analyse data based on different ESCO groups. For readability, each group is annotated with a single word that best describes it. For example, "Army" represents Armed Forces occupations, "Technicians" stands for Technicians and Associate Professionals, etc. The full list of ESCO groups can be found alongside our published dataset.

In Fig. 4, we see the effect of having a tertiary degree on the distribution of jobs. This figure is weighted by job duration, so individuals who frequently change careers have the same effect as those who stay in their job for a long time. As expected, people with tertiary degrees are more likely to work in managerial and professional jobs, which typically require more education and skills. Conversely, those without a tertiary degree are more likely to work in elementary occupations.

In jobs such as service and sales workers and skilled agricultural, forestry, and fishery workers, the effect of having a tertiary degree is not significant, as these roles often require practical skills rather than theoretical knowledge. Professional and elementary occupations are the two groups most affected by having a tertiary degree.

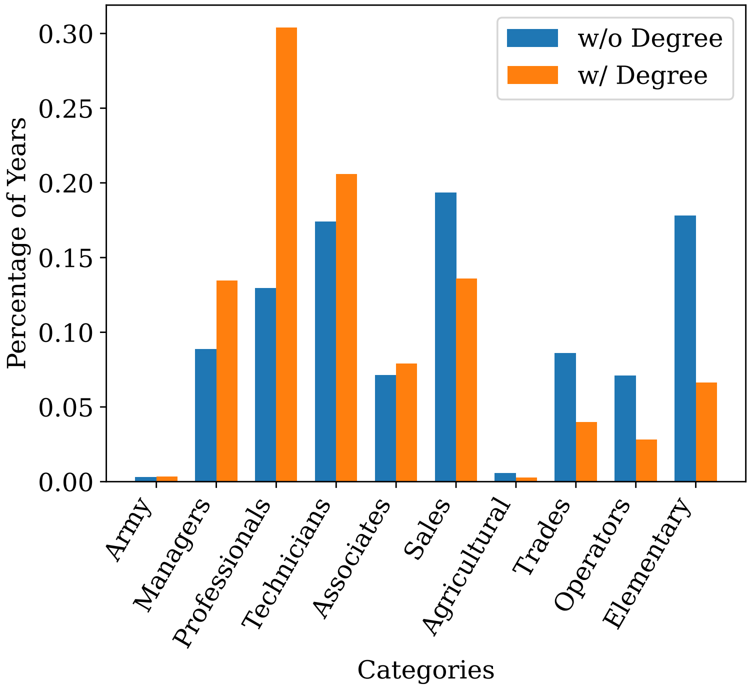

Fig. 4. Normalized effect of a tertiary degree on job categories weighted by job duration.

Fig. 3 illustrates the jobs people transition to after a career break. Each bar shows the top 13 jobs that individuals move into post-break, with bar height indicating the percentage of people transitioning to that job. The five bars on top represent all resumes with a career break, while the five bars below represent resumes with a tertiary degree and a career break. In the top figure, jobs like sales

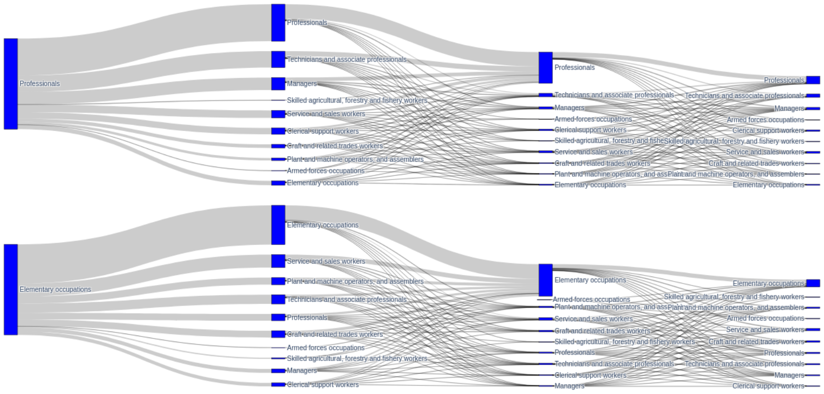

Fig. 5. Job transitions for people starting in professional and elementary occupations.

assistant, administrative assistant, factory hand, building cleaner, and warehouse worker are common post-break jobs. The first two bars indicate jobs with high career change frequency, while the other three bars represent entry-level jobs after a break lasting at least one year. As break duration increases, jobs with lower skill requirements become more common, reflecting the difficulty of finding a field-specific job after a long break.

The plot below in Fig. 3 presents the same information for individuals with a tertiary degree. Here, jobs requiring more skills are more common, suggesting that a tertiary degree helps individuals find jobs in their field even after a break. Jobs like school teacher, project manager, and sales manager remain top transition choices even after a five-year break.

Table 5. Highest median duration of jobs in years.

Occupation	**Duration (Years)**
Land-based machinery operator	5.0
ICT operations manager	5.0
Sewing machine operator	5.0
Swimming facility attendant	4.0
ICT research manager	4.0
ICT information and knowledge manager	4.0
Offset printer	4.0
Industrial quality manager	4.0
Software manager	3.0
Business manager	3.0

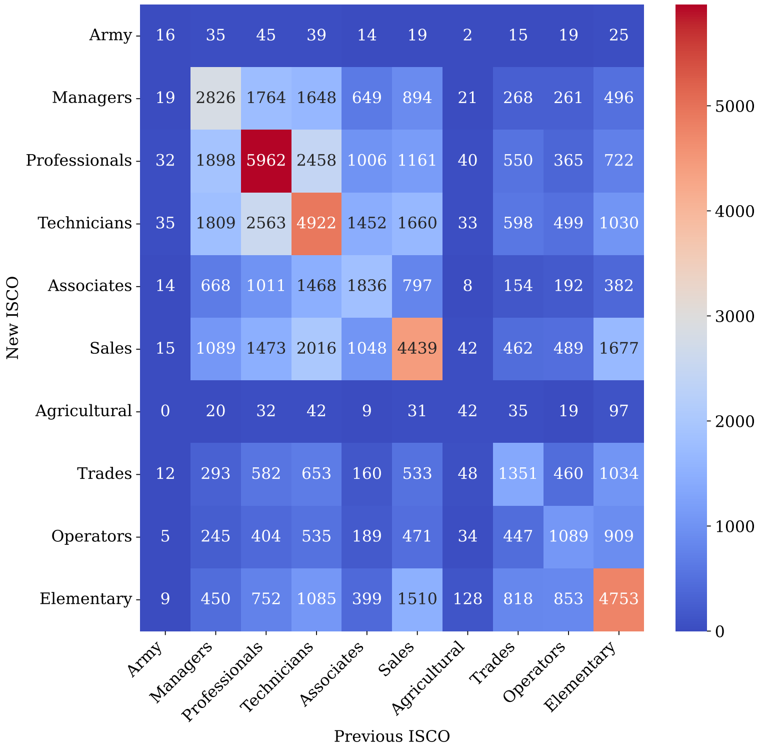

Fig. 6. Transitions between different ESCO groups.

Table 5 and Table 6 show the highest median duration of jobs in years across ESCO occupations. To remove outliers, we only consider jobs with more than 50 appearances in the dataset.

We conducted this analysis separately for all resumes and for resumes of individuals with a tertiary degree. The comparison reveals that individuals with a tertiary degree tend to stay in their jobs for shorter periods than those without a degree. This trend may be attributed to the greater opportunities available to degree holders, allowing them to find better positions more easily. Additionally, university graduates are more likely to experience vertical movement through promotions, whereas those without a tertiary degree are more likely to experience horizontal movement within the job market.

Table 6. Highest median duration of jobs in years for people with tertiary degrees.

Occupation	**Duration (Years)**
Mayor	4.0
Software Analyst	3.0
Bicycle Shop Manager	3.0
Babysitter	3.0
Research Engineer	3.0
Advanced Nurse Practitioner	3.0
Musician	3.0
ICT Project Manager	3.0
Relationship Banking Manager	2.8
Business Manager	2.6

Fig. 5 shows job transitions for individuals whose starting job was in a professional or elementary occupation. The first bar shows the starting job, and the last bar shows the fourth job in their career. As you can see in Fig.6, people are more likely to stay in the same group. Interestingly, Fig.5 shows that in our data many people who change their career once tend to return to their original group if they change careers again.

Fig. 6 shows the transition matrix between different ESCO groups. We only consider two consecutive jobs in a resume as a transition when the end date of the first job precedes the start date of the second job. There are few jobs from the Armed Forces occupations and Skilled Agricultural, Forestry, and Fishery Workers groups in our dataset. For the remaining groups, the diagonal of the matrix is more red, indicating that people are more likely to stay in the same group rather than transition to another group. Significant transitions occur between groups like Managers, Professionals, and Technicians and Associate Professionals.

5 Conclusion and Future Work

In this paper, we have conducted a comprehensive data-driven analysis of the Flemish labour market from the perspective of the available labour force. This analysis was enabled by access to a large dataset of anonymised job seeker resumes provided by VDAB, the Flemish public employment service. To process this data and extract insights, we proposed a two-step LLM-based pipeline. The first step leverages LLMs to extract structured information from unstructured text. The second step utilizes sentence embedding models to map structured data to a standard occupation taxonomy, i.e. ESCO. As a result, we obtain a curated dataset subsequently used to analyse job market trends and patterns and address key questions regarding the effects of education, career breaks, and job transitions.

Our findings confirm that having a tertiary degree significantly influences job distribution, with degree holders more likely to occupy managerial and professional roles, while non-degree holders are more prevalent in elementary occupations. Career breaks, especially longer ones, tend to push individuals towards jobs with lower skill requirements, although those with tertiary degrees face fewer difficulties in finding field-specific jobs post-break.

The complementary view of the labour market from the perspective of job seekers, is often unexplored due to a lack of data. For this reason, as an additional contribution, we release a large dataset of ESCO occupations. This dataset contains over 600.000 occupation codes aggregated into more than 115.000 user resumes. By providing aggregated and chronologically ordered occupation codes per resume, we hope to encourage future research into job market transitions. Moreover, through the use of a standard occupation taxonomy we aim to enable other researchers to compare job market trends in Flanders with other regions and countries.

This work opens multiple avenues for further research ranging from improved data extraction pipelines to more detailed analyses based on the available labour market dataset. More specifically, the accuracy and efficiency of the structured information extraction process can be significantly improved. This can be achieved through advanced prompt engineering strategies, fine-tuning techniques, and exploring newer LLMs. Additionally, larger datasets and increased computational power could lead to improved performance. Similarly, more accurate models for mapping textual data to ESCO can be developed. This can be achieved through the incorporation of additional contextual information from resumes.

The dataset we have released also has the potential to be utilized in various applications such as providing career path advice and job recommendations. By analysing historical career paths and job transitions, we can develop models that can predict the most suitable career trajectories for individuals based on their skills, education, and work experience. Furthermore, incorporating insights from our dataset can enhance job recommendation systems, enabling more personalized and accurate job matching. Additionally, further analysis of the dataset can provide deeper insights into the job market and job transitions. We hope that our work will inspire further research in this field and contribute to the development of more effective labour policies and strategies.

By continuing to refine our methods and expand the scope of our analysis, we can contribute to a deeper understanding of job market dynamics and support efforts to create more effective labour policies and strategies.

Acknowledgments. This research was funded by the ERC under the E.U.'s 7th Framework and H2020 Programmes (ERC Grant Agreement no. 615517 and 963924), the Flemish Government (AI Research Program), the BOF of Ghent University (BOF20/DOC/144 and BOF20/IBF/117), and the FWO (G0F9816N, 3G042220).

References

1. European commision: european skills, competences, qualifications and occupations dataset v1.1.1, https://esco.ec.europa.eu/en/classification
2. Alibasic, A., Upadhyay, H., Simsekler, M., Kurfess, T., Woon, W., Omar, M.: Evaluation of the trends in jobs and skill-sets using data analytics: a case study. Journal of Big Data **9**(1) (2022). https://doi.org/10.1186/s40537-022-00576-5
3. Aljohani, N.R., Aslam, M.A., Khadidos, A.O., Hassan, S.U.: A methodological framework to predict future market needs for sustainable skills management using AI and big data technologies. Applied Sciences **12**(14), 6898 (Jan 2022). https://doi.org/10.3390/app12146898, number: 14 Publisher: Multidisciplinary Digital Publishing Institute
4. Ayishathahira, C.H., Sreejith, C., Raseek, C.: Combination of neural networks and conditional random fields for efficient resume parsing. In: 2018 International CET Conference on Control, Communication, and Computing (IC4). pp. 388–393. IEEE, Thiruvananthapuram (Jul 2018)
5. Banerjee, P.S., Chakraborty, B., Tripathi, D., Gupta, H., Kumar, S.S.: A information retrieval based on question and answering and ner for unstructured information without using sql. Wireless Personal Communications **108**(3), 1909–1931 (Oct 2019). https://doi.org/10.1007/s11277-019-06501-z
6. Barducci, A., Iannaccone, S., La Gatta, V., Moscato, V., Sperlì, G., Zavota, S.: An end-to-end framework for information extraction from Italian resumes. Expert Systems with Applications **210**, 118487 (Dec 2022). https://doi.org/10.1016/j.eswa.2022.118487
7. Bethmann, A., Schierholz, M., Wenzig, K., Nester, M.: Automatic coding of occupations. using machine learning algorithms for occupation coding in several german panel surveys (Sep 2014)
8. Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., et al.: WoLMIS: a labor market intelligence system for classifying web job vacancies. Journal of Intelligent Information Systems **51**(3), 477–502 (Dec 2018). https://doi.org/10.1007/s10844-017-0488-x
9. Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M.: Classifying online job advertisements through machine learning. Future Generation Computer Systems **86**, 319–328 (Sep 2018). https://doi.org/10.1016/j.future.2018.03.035
10. Cao, L.: Data science: a comprehensive overview. ACM Comput. Surv. **50**(3) (jun 2017). https://doi.org/10.1145/3076253

11. Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., et al.: LLMs accelerate annotation for medical information extraction. In: Proceedings of the 3rd Machine Learning for Health Symposium. pp. 82–100. PMLR (Dec 2023), iSSN: 2640-3498
12. Gregory, C., Lewis, P., Frugoli, P., Nallin, A.: Updating the O*NET-SOC Taxonomy: Incorporating the 2018 SOC Structure. O*NET Resource Center, USA (2019)
13. Gurcan, F., Cagiltay, N.E.: Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. IEEE Access **7**, 82541–82552 (2019). https://doi.org/10.1109/ACCESS.2019.2924075, conference Name: IEEE Access
14. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., et al.: Mistral 7b (Oct 2023). https://doi.org/10.48550/arXiv.2310.06825, arXiv:2310.06825 [cs]
15. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: Proc. of Advances in Neural Information Processing Systems. vol. 35, pp. 22199–22213. Curran Associates, Inc. (2022)
16. Kuratov, Y., Bulatov, A., Anokhin, P., Sorokin, D., Sorokin, A., Burtsev, M.: In search of needles in a 11m haystack: recurrent memory finds what llms miss (2024)
17. Le, R., hu, W., Song, Y., Zhang, T., Zhao, D., Yan, R.: Towards effective and interpretable person-job fitting. pp. 1883–1892 (Nov 2019). https://doi.org/10.1145/3357384.3357949
18. Li, N., Kang, B., De Bie, T.: Llm4jobs: unsupervised occupation extraction and standardization leveraging large language models (Sep 2023). https://doi.org/10.48550/arXiv.2309.09708, arXiv:2309.09708 [cs]
19. Li, Z., Qu, D., Xie, C., Zhang, W., Li, Y.: Language model pre-training method in machine translation based on named entity recognition. International Journal on Artificial Intelligence Tools **29**(07n08), 2040021 (2020). https://doi.org/10.1142/S0218213020400217
20. Marín Díaz, G., Galán Hernández, J.J., Galdón Salvador, J.L.: Analyzing employee attrition using explainable ai for strategic hr decision-making. Mathematics **11**(22) (2023). https://doi.org/10.3390/math11224677
21. Mollá, D., van Zaanen, M., Smith, D.: Named entity recognition for question answering. In: Proceedings of the Australasian Language Technology Workshop 2006. pp. 51–58. Sydney, Australia (Nov 2006)
22. Ngo, T.M., Nguyen, Q.N., Nguyen, T.D., Tran, O.T.: A two-phase framework for automated information extraction from curriculum vitae. In: 2023 RIVF International Conference on Computing and Communication Technologies (RIVF). pp. 1–6 (2023). https://doi.org/10.1109/RIVF60135.2023.10471865
23. Rahhal, I., Kassou, I., Ghogho, M.: Data science for job market analysis: a survey on applications and techniques. Expert Systems with Applications **251**, 124101 (2024). https://doi.org/https://doi.org/10.1016/j.eswa.2024.124101
24. Roha, V.S., Saini, N., Saha, S., Moreno, J.G.: MOO-CMDS+NER: Named entity recognition-based extractive comment-oriented multi-document summarization. In: Advances in Information Retrieval. pp. 580–588. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-28238-6_49
25. of Statistics, A.B., Zealand, S.N.: ANZSCO - Australian and New Zealand Standard Classification of Occupations, 2021. Australian Bureau of Statistics and Statistics New Zealand, Canberra, Australia (2021), https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-classification-occupations-anzsco

26. Sufi, F.K., Razzak, I., Khalil, I.: Tracking anti-vax social movement using AI-based social media monitoring. IEEE Transactions on Technology and Society **3**(4), 290–299 (Dec 2022). https://doi.org/10.1109/TTS.2022.3192757, conference Name: IEEE Transactions on Technology and Society
27. Tavakoli, M., Hakimov, S., Ewerth, R., Kismihok, G.: A recommender system for open educational videos based on skill requirements. In: 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT). pp. 1–5. IEEE Computer Society, Los Alamitos, CA, USA (jul 2020). https://doi.org/10.1109/ICALT49669.2020.00008
28. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., et al.: Llama 2: open foundation and fine-tuned chat models (Jul 2023). https://doi.org/10.48550/arXiv.2307.09288, arXiv:2307.09288 [cs]
29. Tsai, C.W., Lai, C.F., Chao, H.C., Vasilakos, A.V.: Big data analytics: a survey. Journal of Big Data **2**(1), 21 (Oct 2015). https://doi.org/10.1186/s40537-015-0030-3
30. Upadhyay, C., Abu-Rasheed, H., Weber, C., Fathi, M.: Explainable job-posting recommendations using knowledge graphs and named entity recognition. In: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 3291–3296 (Oct 2021). https://doi.org/10.1109/SMC52423.2021.9658757, iSSN: 2577-1655
31. Vankevich, A., Kalinouskaya, I.: Ensuring sustainable growth based on the artificial intelligence analysis and forecast of in-demand skills. E3S Web of Conferences **208**, 03060 (2020). https://doi.org/10.1051/e3sconf/202020803060
32. Varelas, G., Lagios, D., Ntouroukis, S., Zervas, P., Parsons, K., Tzimas, G.: Employing natural language processing techniques for online job vacancies classification. In: Artificial Intelligence Applications and Innovations. AIAI 2022 IFIP WG 12.5 International Workshops. pp. 333–344. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-08341-9_27
33. Vijayan, A.: A prompt engineering approach for structured data extraction from unstructured text using conversational llms. In: Proceedings of the 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence. p. 183–189. ACAI '23, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3639631.3639663
34. Vinel, M., Ryazanov, I., Botov, D., Nikolaev, I.: Experimental comparison of unsupervised approaches in the task of separating specializations within professions in job vacancies. In: Artificial Intelligence and Natural Language. pp. 99–112. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-34518-1_7
35. Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., et al.: Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. Journal of Chemical Information and Modeling **59**(9), 3692–3702 (Sep 2019). https://doi.org/10.1021/acs.jcim.9b00470, publisher: American Chemical Society