# Causal Synthetic Data Generation in Recruitment

Andrea Iommi[1,*], Antonio Mastropietro[1], Riccardo Guidotti[1,2], Anna Monreale[1,2] and Salvatore Ruggieri[1,2]

[1]University of Pisa, Italy
[2]ISTI-CNR Pisa, Italy

**Abstract**

The importance of Synthetic Data Generation (SDG) has increased significantly in domains where data quality is poor or access is limited due to privacy and regulatory constraints. One such domain is recruitment, where publicly available datasets are scarce due to the sensitive nature of information typically found in curricula vitae, such as gender, disability status, or age. This lack of accessible, representative data presents a significant obstacle to the development of fair and transparent machine learning models, particularly ranking algorithms that require large volumes of data to effectively learn how to recommend candidates. In the absence of such data, these models are prone to poor generalisation and may fail to perform reliably in real-world scenarios. Recent advances in Causal Generative Models (CGMs) offer a promising solution. CGMs enable the generation of synthetic datasets that preserve the underlying causal relationships within the data, providing greater control over fairness and interpretability in the data generation process. In this study, we present a specialised SDG method involving two CGMs: one modelling job offers and the other modelling curricula. Each model is structured according to a causal graph informed by domain expertise. We use these models to generate synthetic datasets and evaluate the fairness of candidate rankings under controlled scenarios that introduce specific biases.

**Keywords**

Causal Generative Models, Ranking, Recruitment, Bias simulation, Fairness evaluation

## 1. Introduction

Synthetic data generation is gaining importance, especially in contexts where data quality is low, privacy concerns are prominent, or regulatory constraints limit data availability. Poor-quality datasets, often containing missing or unrepresentative information, can significantly impair the performance of Machine Learning (ML) models [1]. In many domains, collecting real data is prohibitively expensive or logistically challenging, and ensuring coverage of all relevant scenarios is rarely straightforward [2]. Moreover, in high-risk settings such as healthcare [3], business [4], or recruitment [5], extensive preprocessing and privacy-preserving measures often degrade data utility, further motivating the need for high-quality synthetic alternatives. Synthetic Data Generators (SDGs) offer a promising solution to challenges related to data scarcity, privacy, and regulatory compliance [6]. In healthcare, for example, SDGs support disease modelling and drug discovery while preserving patient confidentiality. Models such as SynSys [7] and CorGAN [8] address data availability and privacy concerns in medical applications. Similarly, in business domains, strict privacy regulations often hinder research and development. In [4], it is demonstrated how the SDGs can be utilised to simulate financial scenarios under specific constraints.

As in other sensitive domains, the availability of accessible datasets in human recruitment is limited due to the private nature of attributes such as gender, disability, and age, which are pieces of information that candidates may be reluctant to disclose. Consequently, synthetic datasets play a crucial role in this field, enabling the training and evaluation of ranking models not only in terms of performance but also in terms of fairness within human recommendation systems. In addition to improving model

*AEQUITAS 2025: Workshop on Fairness and Bias in AI | co-located with ECAI 2025, Bologna, Italy*
*Corresponding author.
andrea.iommi@phd.unipi.it (A. Iommi); antonio.matropietro@di.unipi.it (A. Mastropietro); riccardo.guidotti@unipi.it (R. Guidotti); anna.monreale@unipi.it (A. Monreale); salvatore.ruggieri@unipi.it (S. Ruggieri)
0009-0007-8337-3695 (A. Iommi); 0000-0002-8823-0163 (A. Mastropietro); 0000-0002-2827-7613 (R. Guidotti); 0000-0001-8541-0284 (A. Monreale); 0000-0002-1917-6087 (S. Ruggieri)

effectiveness, synthetic data helps mitigate the risk of disclosing sensitive attributes, thereby addressing both ethical and legal concerns. Unfortunately, generating synthetic datasets that accurately reflect real-world data is a non-trivial task. Indeed, to be effective, synthetic data must closely replicate the underlying statistical properties of the original data. Furthermore, in socially sensitive domains, the generation process must also ensure fairness and interpretability to prevent biased outcomes.

Causal Generative Models (CGMs) can address these needs by explicitly encoding causal relationships using Structural Causal Models (SCMs) [9]. Indeed, unlike deep learning approaches such as Generative Adversarial Networks (GANs) [10] or Variational Autoencoders (VAEs) [11], which excel at capturing complex non-linear patterns, CGMs provide transparent and interpretable mechanisms grounded in causality. While deep learning models can uncover correlations, they often fail to reveal the underlying causal structure and may introduce spurious associations due to opaque training processes [12, 13]. In fact, in high-risk domains, the importance of interpretability is underscored by the European Union's AI Act, which mandates transparency and human oversight in AI systems [14]. This regulation highlights the need for transparent ML models and data generation processes that can be audited and understood by domain experts.

In this paper, we present a SDG system grounded on two CGMs, one for job offers and one for curricula, each structured according to causal graphs derived from interviews with domain experts. These graphs capture the decision-making processes underlying the creation of job offers and candidate profiles. We use the CGMs to generate synthetic datasets that simulate realistic recruitment scenarios.

As a test-bed of our SDG, we explore fairness in ranking tasks. We introduce a controlled bias by incorporating a parametric causal link between the *gender* attribute and *working hours*, simulating gender disparities as discussed in social sciences [15]. This setup enables us to assess how such bias, when propagated through the data, influences the fairness of rankings of job candidates produced by ML ranking models. In summary, the contribution of this work is threefold: *(i)* the formulation of causal graphs to model the HR domain for job offer and curriculum generation, *(ii)* a new approach to a tabular synthetic data generation, causality-grounded and intrinsically interpretable and *(iii)* a public and extendible GitHub Python repository[1] in which are deployed Causal mechanisms that work with multiple data types.

The rest of this paper is organised as follows. After reviewing works on synthetic data generation in Section 2, we briefly review the key concepts behind our proposal in Section 3. Then, in Section 4, we describe our proposal. In Section 5, we present the experimental results. Finally, Section 6 summarises our contributions and outlines potential directions for future research.

## 2. Related Work

We begin by reviewing the literature on synthetic data generation for tabular data.

Early statistical methods, such as SMOTE [16], generate synthetic samples by fitting empirical distributions and interpolating between existing data points. While effective in leveraging marginal distributions, these techniques often fall short in capturing complex feature interactions, which can result in the generation of less realistic (low-fidelity) and less representative (biased) synthetic data.

Deep learning (DL)–based models, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models [6, 1], show impressive capabilities in learning complex data distributions. However, these models also present several limitations. First, as model complexity increases, they require large volumes of training data. In data-scarce settings, these models often struggle to learn accurate distributions, resulting in poor generalisation. Second, DL–based generative models are prone to *mode collapse*. The latter is a phenomenon where the model generates samples from only a limited subset of the true distribution, namely, a subset with high probability mass. This issue, as discussed in [17], undermines the diversity and representativeness of the synthetic data, making it difficult to ensure dataset quality. Lastly, these models typically function as black boxes, offering

[1] https://github.com/jacons/CausalSDG

limited transparency into the data generation process. This lack of interpretability poses challenges in high-stakes applications, where understanding the model's behaviour is critical.

In contrast to deep generative models, probabilistic graphical models, such as Bayesian Networks (BNs), are better suited under certain conditions, namely when *(i)* datasets are limited in size, *(ii)* consistency is critical, meaning generated samples must adhere to domain-specific constraints, and *(iii)* strong correlations exist among features [18, 19]. BNs explicitly model the conditional dependencies between variables, allowing them to capture the joint distribution of the data more transparently. This structure enables BNs to replicate the statistical properties of an empirical dataset with greater interpretability compared to DL–based approaches. Structural Causal Models (SCMs) extend the capabilities of BNs by embedding causal semantics into the graphical structure. While BNs focus on statistical dependencies, SCMs incorporate structural equations that define how each variable is generated from its causes. This allows SCMs not only to model observational distributions but also to support interventional and counterfactual reasoning. As a result, SCMs provide a more expressive framework for generating synthetic data that is both statistically consistent and causally grounded, making them particularly valuable in domains where fairness, transparency, and causal interpretability are essential.

A growing body of research has focused on developing methods for generating synthetic data that explicitly mitigate fairness concerns while preserving utility. Specifically, [20, 21, 22] extend GAN–based methods, and [23] adopts a genetic approach. Unlike these approaches, *(i)* we focus on the specific domain of recruiting (curricula and job offer datasets), *(ii)* we adopt a SCM approach with controllable bias parameters, and *(iii)* the data generation process is fully interpretable. The focus on a specific domain permits us to derive causal dependencies among features by eliciting them from expert knowledge, thus overcoming the difficulty of discovering causal dependencies from observational data. We instead use observational data to learn the structural equations given the known causal dependencies among features. To the best of our knowledge, this is the first approach to using SCMs for SDG in the recruiting domain.

The closest works are [22] and [24]. van Breugel et al. [22] is a generic approach, which assumes a given causal graph and learns the structural equations through conditional GANs. The data generation process first intervenes on the derived SCM by eliminating dependencies that lead to unfairness in a downstream model (these dependencies may be specific to the fairness metric being considered). Subsequently, it generates data by applying the (GAN–based) structural equations of the remaining dependencies. On the other hand, our method integrates fairness constraints directly into the causal mechanisms during the generative process, permitting us to produce fairness requirements without integrating post-hoc intervention. Barbierato et al. [24] present a methodology for bias-controllable synthetic data generation using parametric causal mechanisms. Their experimental framework explores fairness metrics by systematically varying a bias parameter during the data generation process. However, their approach is limited to continuous variables and lacks the ability to learn causal mechanisms directly from data, relying instead on predefined parametric forms. In contrast, our approach supports mixed data types and utilises a learned causal mechanism fitted from observational data.

## 3. Background

To keep the paper self-contained, we provide a brief outline of the key concepts underlying our proposal.

### 3.1. The ESCO Taxonomy and the EQF Classification System

The *European Skills, Competences, Qualifications, and Occupations* [25] (ESCO, https://esco.ec.europa.eu/) is a multilingual classification system developed by the European Union (EU) to provide a standardised framework for describing skills and occupations. It aims to support better matching between individuals and job opportunities, as well as between education and labour market needs.

ESCO is organised into two main pillars: Occupations and Skills. The Occupations pillar provides a structured vocabulary for consistently describing occupations. This structure is based on hierarchical

relationships between concepts, using the *International Standard Classification of Occupations* (ISCO-08) as its foundational taxonomy. Each occupation entry includes several attributes, such as a Unique Resource Identifier (URI), a preferred term that represents the concept in a specific language, a set of non-preferred terms including synonyms, spelling variants, declensions, and abbreviations, and a textual description. Similarly, the Skills pillar provides a taxonomy for describing competencies and knowledge. It mirrors the hierarchical organisation of the Occupations pillar and shares the same set of attributes. This consistency facilitates interoperability and integration across different systems and languages. ESCO also defines associations between occupations and skills, categorising them as either essential or optional. These associations enhance the system's ability to support detailed profiling and more accurate matching in both employment and educational contexts.

The *European Qualifications Framework* (EQF, https://europass.europa.eu/en/europass-digital-tools/european-qualifications-framework) serves as a common European reference framework to facilitate the comparison of qualifications across different countries and education systems. Established by the EU, the EQF aims to promote transparency, mobility, and lifelong learning by aligning national qualification systems through a shared structure into eight levels: from Level 1, which corresponds to basic general knowledge and skills, to Level 8, which reflects the highest level of expertise, typically associated with doctoral-level qualifications. Each level is defined by a set of descriptors that express the expected learning outcomes in terms of knowledge, skills, and competence.

### 3.2. Structural Causal Models

A Structural Causal Model (SCM) [26, 9, 27] describes a data-generating process by relating random variables in cause-effect pairs. Let $\mathbf{X} = \{X_1, \ldots, X_d\}$ be $d$ observable random variables, defined by a set $\mathbf{F}$ of structural equations:

$$X_i := f_i(\mathbf{PA}_i, U_i) \qquad \text{for } i = 1, \ldots, d \tag{1}$$

where $\mathbf{U} = \{U_1, \ldots, U_d\}$ are $d$ independent exogenous (unobserved) random variables, and $\mathbf{PA}_i \subseteq \mathbf{X} \setminus \{X_i\}$ are the causal parents of $X_i$. The equations describe the causal mechanism by which an $X_i$ is generated from its causal parents and an exogenous variable $U_i$. Formally, a SCM $\mathcal{M}$ is a tuple $\mathcal{M} = \langle \mathbf{U}, P(\mathbf{U}), \mathbf{X}, \mathbf{F} \rangle$, where $P(\mathbf{U}) = \prod_i P(U_i)$ is the probability distribution of the exogenous variables. The parental relations in a SCM induce a *causal graph* $\mathcal{G}$, in which the nodes represent random variables and a directed edge $X_j \to X_i$ denotes a causal relation between $X_j \in \mathbf{PA}_i$ and $X_i$. We assume Directed Acyclic Graphs (DAGs), meaning there are no loops in $\mathcal{G}$, so the data generation process can proceed by following a topological order of the variables given the graph. Under the Markov property assumption, the induced probability on $\mathbf{X}$ can then be factorized as $P(\mathbf{X}) = \prod_i P(X_i|\mathbf{PA}_i, U_i)$.

Let us assume that we know the causal graph $\mathcal{G}$, and that we are given a dataset of i.i.d. observations. Parametric and non-parametric approaches can be used to infer the structural equations. In this work, we consider the following approach based on the type of $X_i$.

For a continuous variable $X_i$, we model the task as a regression problem of the dependent continuous variable $X_i$ given the independent variables $\mathbf{PA}_i$ and the exogenous variable $U_i$. In particular, we assume additive noise, namely, the structural equation is of the form: $X_i = f_i(\mathbf{PA}_i) + U_i$, where $f_i$ is a regressor trained from the dataset of observations, and the exogenous variable $U_i$ is assumed to be empirically distributed from the residual $X_i - f_i(\mathbf{PA}_i)$.

For $X_i$ discrete, we assume $X_i \sim Cat(\hat{f}_i(\mathbf{PA}_i))$, namely $X_i$ is categorically distributed with probabilities given by the predictions of a probabilistic ML classifier[2] $\hat{f}_i$ trained from the dataset of observations. If a variable in $\mathbf{PA}_i$ is of set type (e.g., skills required in a job offer or those possessed by a job applicant), we first one-hot encode all possible values in the set.

For $X_i$ of set type with possible values in $\{v_1, \ldots, v_m\}$, then $f_i$ samples $n \sim U(n_{min}, n_{max})$ values without replacement, where each $v_j$ is sampled with probability equal to the empirical conditional distribution $P(v_j \in X_i|\mathbf{PA}_i)$ over the dataset of observations. Here, $n_{min}$ and $n_{max}$ are user parameters.

[2]We adopt a *HistGradientBoostingClassifier* from the *scikit-learn* Python library.

# 4. Methodology

In this section, we outline: *(i)* the formulation of causal graphs to model the HR domain for job offer and curriculum generation and *(ii)* the design of a data preprocessing pipeline.

## 4.1. Causal Graphs for SDG

A central component of our approach is the causal graph representing the structure of job offers and curricula. Rather than relying on automated causal discovery from observational data, we construct this graph through expert elicitation, based on qualitative insights gathered from HR professionals. To this end, we conducted four semi-structured interviews with HR representatives from three Italian companies of varying sizes: one with over 50 employees, another with over 200 employees, and a third with more than $5,000$ employees. These organisations operate in distinct sectors and organisational scales, providing a diverse perspective on hiring practices. The interviews were designed to uncover the real-world decision-making processes behind the creation of job offers. Participants were asked to describe the key attributes considered when drafting job postings, as well as the relationships among these elements. The insights obtained were instrumental in defining the structure and dependencies encoded in the causal graph, ensuring that it reflects domain-specific knowledge and practical constraints.

**The Process Leading to Job Offers.** In the companies interviewed, the recruitment process is closely integrated with annual budget planning. Each production unit submits its needs for *occupations*, which are then reviewed and approved through an internal administrative process. These needs may arise from various factors, including employee resignations, unplanned replacement costs, increased workloads, or skill shortages within teams. A common sequence emerged across the interviews regarding how job postings are formulated. The process typically begins with the budget, which is defined during the annual planning phase. This *budget* plays a pivotal role in shaping the job offer: it determines the experience of the sought candidate (e.g., junior or senior), the nature of the employment contract, and its duration. Specifically, higher budget availability, combined with organisational needs, affects the number of *working hours* and the *contract type* (intended as contract duration: permanent or fixed-term). Under the guidance of the department head who initiated the staffing request, HR professionals identify the required hard skills. Based on these, they determine the appropriate seniority level (e.g., years of *experience*) and specify technical qualifications, such as degrees or certifications. This structured approach reflects a rationale in which workforce needs are translated into a set of competencies that candidates must possess. These competencies are typically associated with specific educational backgrounds or levels of professional experience. Moreover, defining education and experience through the lens of skills provides an initial validation mechanism: requiring a degree in a specific field, HR professionals implicitly assume that the candidate possesses the associated skills. Interestingly, some interviews revealed that the process can also operate in reverse—starting with an ideal candidate profile and subsequently adjusting it to fit within budget constraints. Thus, the budget may either serve as the starting point for defining job requirements or act as a constraint to be considered after the ideal profile has been outlined. Figure 1 shows the job posting generation process as reconstructed from the interviews.

**Causal Graphs for Job Offers and Curricula.** Job offers typically do not include all of the variables of the graph from Figure 1. For instance, the budget of companies is not disclosed. Other variables may be missing for specific data collections. In our datasets (see next section), in particular, we lack the *contract type*, the *salary*, and the *age target*. To reflect these limitations, we present a simplified version of the causal graph used in our experiments, shown in Figure 2 (left). Despite this simplification, our SDG is designed to be fully parametric with respect to the input DAG–based causal graph. This design enables the system to be applied to richer datasets than the one used in our current experiments, granting broader applicability in future scenarios where more complete data is available. The causal graph for job offers models the determination of required *education* and *experience* based on the *skills*

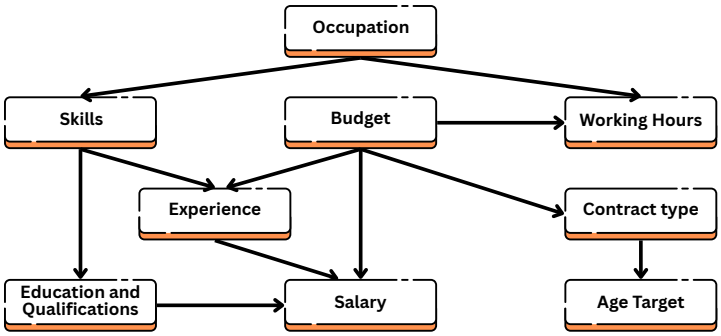


**Figure 1:** Job posting generation process as reconstructed from the interviews with HR professionals. The meaning of an arrow $A \to B$ is "$A$ determines $B$".

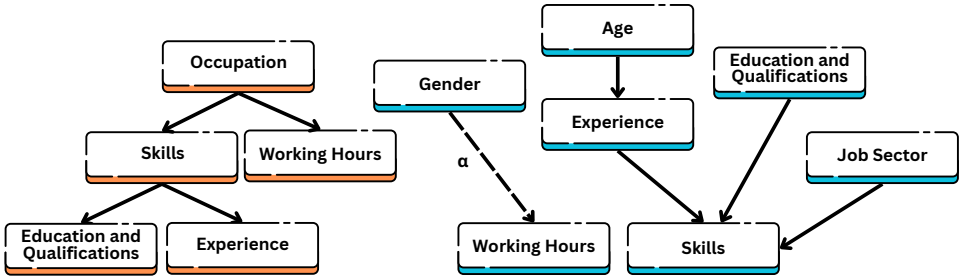


**Figure 2:** Causal graphs adopted for the experiments: left for job offers, right for curricula. The SDG developed is fully general and accepts DAG causal graphs as input. The dotted line in the curriculum's causal graph controls for bias through the parameter $\alpha$. In our experiment, all features are categorical/ordinal, except for "skills" which is a set of categorical values.

necessary for a given occupation. This reflects a form of backwards reasoning: starting from the desired skills, the HR professionals infer the *qualifications* that provide a reasonable guarantee of possessing those skills. In other words, they answer the question: *What educational background and years of experience are typically needed to ensure a candidate has the required competencies?* Conversely, for job applicants, the causal direction is reversed. A candidate's *skills* are shaped by their *education* and years of *experience*. Additionally, the *job sector* in which the candidate has specialised plays a significant role in skill development. It is important to note that *occupation* is a more specific concept than *job sector*. For example, while the *job sector* might be *Information and Communication Technology (ICT)*, the *occupation* could be *Python Developer*. In the causal graph for curricula, *age* is modelled as a determinant of the years of *experience* a candidate has acquired. This reflects the natural progression of professional development over time. The full structure of this graph is shown in Figure 2 (right).

The graph also includes the variable *working hours*, which represents whether a candidate is looking for a part-time or full-time contract. The study presented in [15] offers valuable insights into the ways gender and societal norms influence working time patterns among men and women. It highlights that women are significantly more likely than men to engage in part-time employment and are less inclined to work overtime. Importantly, this trend is not merely a matter of personal preference but is deeply embedded in prevailing social constructs. The primary driver of this disparity lies in traditional gender roles, which continue to position women as the primary caregivers within the family. To model such a potential bias, we introduce a directed edge from *gender* to *working hours*, with the strength of this dependency modulated by a parameter $\alpha$. This parameter allows us to control the degree of gender–based bias in the data generation process and will be varied in the experimental analysis presented later in the paper. An example of a job offer and a corresponding curriculum, including the variables considered in their respective causal graphs, is provided in Table 1.

**Table 1**
An example of a generated job offer (left) and a curriculum (right).

| Features | Values |
|---|---|
| Occupation | *ICT Professional* |
| Working Hours | *Full time contract* |
| Education (EQF) | *level 6* |
| Experience | *1 - 2 years* |
| Skills | *PHP, Java, French* |

| Features | Values |
|---|---|
| Job Sector | *ICT Professional* |
| Education (EQF) | *level 6* |
| Gender | *Male* |
| Working Hours | *Full time contract* |
| Age | *20* |
| Experience | *2 years* |
| Skills | *PHP, English, Groovy* |

## 4.2. Downstream Task: Ranking

The primary downstream application of the SDG framework is to evaluate, or potentially train, a ranking model that orders applicants for a given job position. In this section, we present the end-to-end pipeline, from dataset generation to the evaluation of ranking performance and fairness metrics. To enable the assessment of gender–based (un)fairness in ranking models, our SDG approach incorporates the parameter $\alpha$ within the causal graph generating curricula see Section 4.1). This parameter allows for controlled interventions on gender-related attributes, facilitating rigorous fairness analysis in the ranking process.

**Ranking Pipeline.** Learning to Rank (LTR) methods originate from Information Retrieval [28]. They are typically classified into three main categories w.r.t. a given user query: *pointwise*, where individual documents are scored independently; *pairwise*, where the model learns to predict the relative order between pairs of documents; and *listwise*, where the entire list of documents is considered simultaneously to optimize a ranking metric. In the context of recruitment, the job offer represents a user query, and the curricula identify the documents to be ranked. Since candidates may choose which job offer to apply to, or at least to which job sector, the curricula to be ranked for a given job offer $j$ will be denoted by $\mathcal{C}_j$.

Let us consider the case of pointwise ranking here. A model is trained based on a dataset of observations consisting of pairs $(\mathbf{v}^{(j,c)}, s)$ where: $\mathbf{v}^{(j,c)}$ are fitness values of the curricula $c$ w.r.t. the job offer $j$, and $s$ is a relevance score assigned by an HR professional to the fitness values. The fitness values are numbers in the interval $[0, 1]$ that measure how much each requirement of a job offer is satisfied by a candidate, with 0 indicating no match and 1 indicating a full match. With reference to the graphs of Figure 1, a fitness value is computed for skills (resp., working hours/education and qualification/experience) of the candidate w.r.t. skills (resp., working hours/education and qualification/experience) of the job offer. For instance, the match for skills can be computed as the fraction of skills required by the job offer that the candidate possesses. Regarding the relevance score $s$, when the data is collected from past candidate selections, it is provided by the HR professional assessing the candidates. In the case of synthetically generated data, such information is lacking. To experiment with our SDG, we assume a linear model:

$$s(j, c) = \boldsymbol{w}^T \boldsymbol{v}^{(j,c)} + \beta \tag{2}$$

where $\boldsymbol{w} \in \mathbb{R}^r$ is a weight vector that encodes the importance of each fitness value and $\beta \sim N(0, \sigma^2)$ is a small noise value modelling uncertainty. Due to their interpretability, linear models are often adopted in real ranking systems. Eq. (2) provides the "ground-truth" for evaluating a learned ranking model. We emphasise that, while an actual recruitment system would utilise a ranker such as LambdaMART or ListNet with scores assigned by HR, training and evaluating such models is beyond the scope of this work. Indeed, in our experiment, we are interested in how fairness measurement changes at varying values of $\alpha$'s parameters.

For a fixed job offer $j$, the ground-truth ranking of curricula in $\mathcal{C}_j$ is the one obtained by descending value of $s(j, c)$. Candidates with the same score will have the same ranking position. Let us denote by $\pi_c^j$ the rank position of $c$. We further denote by $\hat{\pi}_c^j$ the rank position of $c$ as determined by another (e.g., a learned) ranking model.

**Performance and Fairness Metrics.** Several evaluation metrics have been considered to quantify the performance of ranking models. They measure how close $\hat{\pi}_c^j$ and $\pi_c^j$ are over all ranked candidates (Kendall's tau, Spearman's rho) or over the top $k$ ranked candidates (precision, NDCG), for a fixed job offer $j$, or for a set of job offers (average precision). In this paper, however, we are not evaluating ranking models, but rather the properties of the datasets generated by our SDG. The performance metric of a SGD is the fidelity of the generated data to the distribution of real data. However, due to the issues mentioned regarding the collection of representative real data, we opt for a "by-design" approach, assuming that the causal graphs depicted in Figure 2 are faithful to the true distribution of job offers and curricula. Conditional distribution of a variable given the parent nodes is enforced by fitting the structural equations as described in Section 3.2. Whether or not the causal graphs from Figure 2 are valid in a specific context, such as countries or industry sectors, remains to be determined context by context, possibly adapting the causal graphs in case of different/additional variables that are observable.

The other property that our SDG can model is the bias of the generated data. We link such a notion to the (un)fairness of the downstream task. Several fairness metrics have been considered in the literature [29, 30]. Since we do not consider a ranking model learned from data, we have to restrict ourselves only to metrics that regard the ground-truth ranking model from Eq. 2, possibly at the variation of the weights $\boldsymbol{w}$. We will consider the group fairness metrics [29] of demographic parity (DP) and normalised discounted difference (rND):

$$DP(j) = 1 - (\mathbb{P}(\pi_c^j \leq k |\ c \textit{ protected}, c \in \mathcal{C}_j) - \mathbb{P}(\pi_c^j \leq k |\ c \textit{ unprotected}, c \in \mathcal{C}_j)) \tag{3}$$

$DP(j)$ ranges in $[0, 2]$, with 2 denoting full unfairness against the protected group (only candidates from the protected group are selected in the top $k$ positions), 1 denoting fairness (equal probability of being chosen among protected and unprotected), and 0 denotes reverse unfairness against the unprotected group. The rND metric improves over DP by considering: (1) the deviation of the fraction of protected candidates in top-$i$ position against the proportion of protected candidates applying for the position; (2) at multiple thresholds $i$'s.

$$rND(j) = \frac{1}{Z} \sum_{i=5,10,15,\ldots}^{|\mathcal{C}_j|} \frac{1}{log_2(i)} \left| \frac{|\{\pi_c^j \leq k, c \textit{ protected}, c \in \mathcal{C}_j\}|}{|\{\pi_c^j \leq k\}|} - \frac{|\{c \textit{ protected}, c \in \mathcal{C}_j\}|}{|\mathcal{C}_j|} \right| \tag{4}$$

where $Z$ is a normalising factor (making it comparable across different $j$'s) so that $rND(j) \in [0, 1]$, where 0 means fairness (proportional representation of protected candidates in top positions).

**Bias Parametrization.** We describe here how the parameter $\alpha$ of the causal graph of the curricula (Figure 2 right) can be used to control the degree of bias induced in the generated datasets. Let us denote the variables *gender* and *working hours* as $G$ and $W$. $G$ is the sensitive variable, assuming the value 0 as *male* (unprotected group) or 1 for *not male* (protected group), while $W$ assume values $f$ for *full-time* or $p$ for *part-time*. When there is no edge between $G$ and $W$, the generation of working hours boils down to the empirical distribution $P(W \mid G)$ over the dataset of observations. The parameter $\boldsymbol{\alpha}$, which we assume to be a pair $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$, can then be used to shift such a distribution either for the unprotected group, through $\alpha_0$, or for the protected group, through $\alpha_1$. Exponential tilting, power transformation, or other probability-shifting/skewing methods can be used. We adopt here exponential tilting, for which the shifted conditional probabilities are: $P'(W = p \mid G = 0) = e^{-2\alpha_1}/C_1 \cdot P(W = p \mid G = 0)$, and $P'(W = f \mid G = 0) = e^{-2\alpha_1}/C_1 \cdot P(W = f \mid G = 0)$, where $C_1$ is a normalizing constant. Similarly, we shift the conditional probability of the protected group: $P'(W = p \mid G = 1) = e^{-2\alpha_0}/C_0 \cdot P(W = p \mid G = 1)$, and $P'(W = f \mid G = 1) = e^{-2\alpha_0}/C_0 \cdot P(W = f \mid G = 1)$, where $C_0$ is a normalising constant.

## 5. Experiments

**Experimental Settings.** In this section, we illustrate some experiments on the functionalities of our SDG. For each experimental parameter ($\alpha_0$ and $\alpha_1$, which will be discussed later), we conducted 10 runs,

**Table 2**
The distribution of *working hours* conditional to *gender* in the synthetic curricula datasets at the variation of $\alpha = (\alpha_0, \alpha_1)$. "Part" stands for part-time contract, and "Full" for full-time contract.

| | $\alpha_0 = -4$ | | $\alpha_0 = -1.5$ | | $\alpha_0 = 0$ | | $\alpha_0 = 1.5$ | | $\alpha_0 = 4$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Part | Full | Part | Full | Part | Full | Part | Full | Part | Full |
| male | 0.24 | 0.76 | 0.24 | 0.76 | 0.24 | 0.76 | 0.24 | 0.76 | 0.24 | 0.76 |
| not-male | 0.02 | 0.98 | 0.25 | 0.74 | 0.59 | 0.41 | 0.87 | 0.13 | 0.99 | 0.01 |
| | $\alpha_1 = -4$ | | $\alpha_1 = -1.5$ | | $\alpha_1 = 0$ | | $\alpha_1 = 1.5$ | | $\alpha_1 = 4$ | |
| | Part | Full | Part | Full | Part | Full | Part | Full | Part | Full |
| male | 0.01 | 0.99 | 0.06 | 0.94 | 0.24 | 0.76 | 0.58 | 0.42 | 0.95 | 0.05 |
| not-male | 0.60 | 0.40 | 0.60 | 0.40 | 0.59 | 0.41 | 0.60 | 0.40 | 0.61 | 0.39 |

where in each one we generated 300 job offers and 1000 curricula. For each job offer, all curricula are considered as candidates. The motivation behind the generated data size is that, in our ranking pipeline, there is no effective model training process, except for the generator. In particular, the relevance score for each candidate is calculated using Eq.2, and the ranking is produced accordingly. Hence, the number of job offers and curricula generated for each run represents a good compromise between observing the changes in metrics and the execution time.

Data is generated according to the causal graphs shown in Figure 2, with the structural equation fitted from the preprocessed datasets described in Appendix A. The share of male and not-male (women and non-binary) applicants is $50\%$-$50\%$. Regarding the ranking model pipeline, the fitness values between curriculum and job offer characteristics are defined through matching functions (MFs). In particular, the MF for *education* returns 1 if the candidate's education level is equal to or greater than the level required in the job offer; otherwise, it is 0. The MF for *experience* checks whether the candidate's experience falls within the interval required by the job offer. The MF for *skills* calculates the fraction of skills required by the job offer that the candidate possesses. Finally, the MF for *working hours* tests for equality of the form required in the job offer and the form desired by the candidate. For example, for the data from Table 1, the fitness value vector is $\boldsymbol{v}^{(j,c)} = [1.0, 1.0, 1/3, 1.0]$ respectively for the MF of the features *education, experience, skills* and *working hours*.

In the following experiments, we investigate how changing the *working hours* preferences conditional on *gender* affects the fairness metrics DP and rND. These kinds of experiments aim to simulate social norms that often say that women should prioritise family responsibilities over their careers, leading them to opt for part-time contracts that offer more free time but may limit career advancement opportunities. The generated job offers exhibit a skewed distribution of working hours: $86.6\%$ of the positions are full-time contracts, while only $13.3\%$ are part-time contracts. Also, the generated curricula have skewed distributions. Male applicants prefer full-time contracts ($76\%$), while not-male applicants prefer part-time contracts ($59\%$). We will be varying $\alpha = (\alpha_0, \alpha_1)$ to investigate shifted distributions. Table 2 presents the average shifted distributions over the 10 experimental runs for a few $\alpha_0$ and $\alpha_1$ (see Section 4.2). In the experiments, we will be varying $\alpha_0$ and $\alpha_1$ from $-4.0$ to $4.0$.

In particular, we denote by $\alpha_0$ the parameter for the "No-man" distribution and by $\alpha_1$ the one for "Man". We recall that larger $\alpha$'s result in a redistribution of probability mass towards part-time. Notice that for $\alpha_0 = -1.5$, the distributions of male and not-male are almost identical. The same occurs for $\alpha_1 = 1.5$.

We explore four weighting vectors for the ranking model of Eq. (2). All of them assign the same weights to the fitness values of *education* (0.8), *experience* (0.5), and *skills* (1.0), and set $\beta \sim N(0, 0.01^2)$, while for *working hours* we consider four cases: $0$, $0.5$, $0.8$ and $1.0$. Since *gender* can only affect the score through *working hours* (cfr. Figure 2 right), when the weight is $0$, then the scores of the ranking model are independent of the *gender*, hence the ranking model is fair.

We emphasise that the weights considered for the experiment correspond to a "rough approximation" of the general HR decision-making. Whether or not they constitute the effective importance value in a candidate evaluation is beyond the scope of this work. What matters is evaluating the fairness metrics at varying *working hours* importances and the $\alpha$'s values.

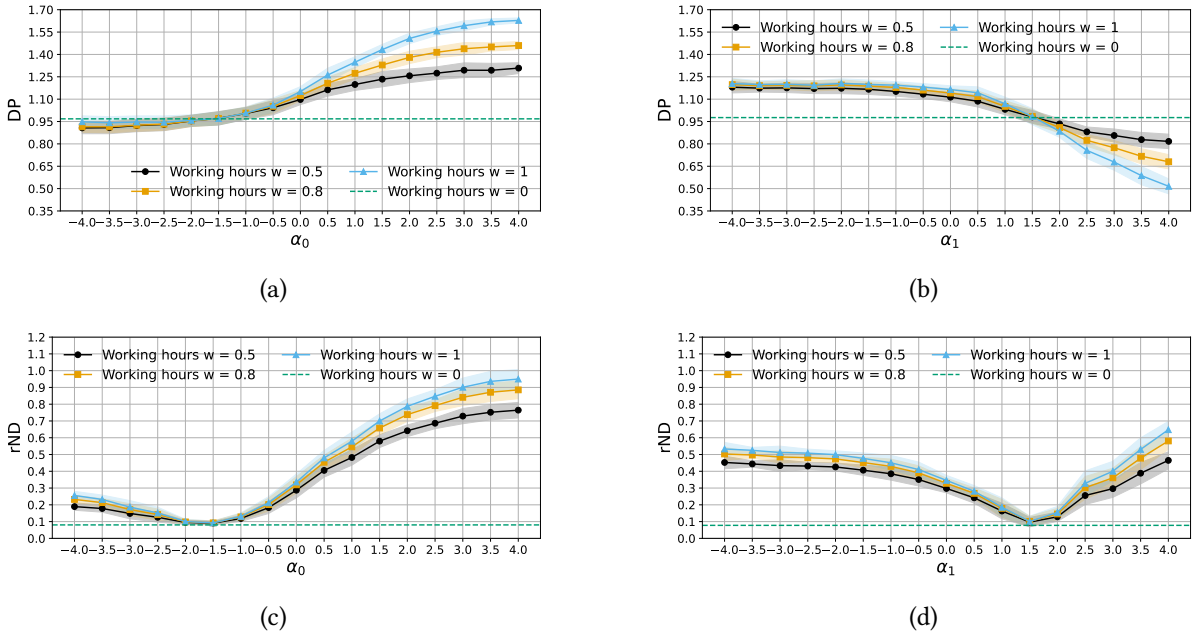


**Figure 3:** Demographic Parity difference (DP) and normalised Discounted Difference (rND) at the variation of the bias-controlling parameters $\alpha_0$ (for not-male) and $\alpha_1$ (for male). DP and rND are averaged over 10 runs, each one with 300 job offers. The series represents the ranking model of Eq. (2) for different weights to the fitness value of *working hours*. Shadows represent $\pm 1$ standard deviation over the 10 runs.

**Results.** Figure 3a illustrates the behaviour of Demographic Parity (DP) under the not-male conditional distribution shift. When the parameter $\alpha_0$ is negative, the distribution of preferences for not-male becomes skewed towards full-time employment. This shift results in a more balanced alignment between job requirements and candidate preferences. Hence, DP is close to the fair value of 1 for all weights of *working hours* in the ranking model. However, as $\alpha_0$ increases, the distribution shifts toward part-time contracts. Given the dominance of full-time job offers, this misalignment causes the ranking system to increasingly favour candidates from the non-protected group, i.e., males. As a result, DP values decline significantly, highlighting a growing disparity and reduced fairness for the protected group, except for the ranking model, where the mediating variable *working hours* has zero weight.

A reversed trend is observed in Figure 3b, which illustrates the variation of DP with respect to changes in the parameter $\alpha_1$. For negative values of $\alpha_1$, male candidates are predominantly associated with full-time contracts. As a result, they are favoured by the ranking model (except in the case where the model assigns zero weight to the *working hours* feature), leading to high DP values that indicate potential unfairness against non-male candidates. As $\alpha_1$ increases, the distribution of male candidates shifts toward part-time contract preferences. This shift reduces their alignment with the majority of job offers, which are primarily full-time, and causes the model to assign higher scores to non-male candidates. Hence, DP values fall below 1, highlighting unfairness against male candidates. Notice that fairness (DP $\approx 1$) is achieved for all ranking models when the conditional distributions of *working hours* are the same for male and not-male, namely for $\alpha_1 = 1.5$ (cfr. also Table 2).

Similar patterns can be observed for the rND metric in Figures 3c and 3d. The metric is computed by considering $i = 5, 10, 15, 20$ in Eq. (4). Unlike Demographic Parity (DP), which measures average differences in exposure between groups, rND captures absolute deviations. As a result, trends where DP falls below 1 (indicating underexposure of the protected group) correspond to elevated rND values. This reflects a greater degree of ranking disparity, as rND penalises any deviation from perfect parity, regardless of direction. As with DP, the most favourable rND values are observed at $\alpha_0 = -1.5$ and $\alpha_1 = 1.5$, corresponding to the settings where the conditional distributions of working hours are equivalent for male and non-male candidates. These configurations yield the most balanced exposure across groups, minimising both average and absolute disparities in ranking outcomes.

## 6. Conclusions

We introduced a synthetic data generation system for recruitment, based on expert-informed causal graphs that enable explicit control over bias and interpretability. Through causal modelling, we assessed the fairness impact of biased data on a linear point-wise ranking model using DP and rND metrics. Our experiments demonstrate that controlled distributional shifts in the generative process can significantly influence ranking fairness—positively or negatively—especially as the weight of bias-related features increases. To foster further research in high-risk human recommendation scenarios, we will release the system as open-source software[3].

Despite promising results, two main limitations remain: the current approach supports only tabular data and depends on the availability of both high-quality training data and well-structured causal graphs; additionally, our experiments suffered from heterogeneous and non-standardised training data, which required approximations during preprocessing and may have reduced representativeness. Future work will focus on *(i)* make comparisons with closest related works pointing out advantages and limitations between the generative methods, *(ii)* enhancing feature engineering to expand causal graphs with additional attributes, *(iii)* analysing more complex discrimination scenarios like the intersectional bias, and *(iv)* integrating into the SDG external knowledge, such as ESCO ontology, to improve the data diversity in the generated data.

## Declaration on Generative AI

During the preparation of this work, I used Grammarly in order to improve my grammar and spelling checking. After using these tools, I reviewed and edited the content as needed.

## References


[1] A. Bauer, S. Trapp, M. Stenger, R. Leppich, S. Kounev, M. Leznik, K. Chard, I. T. Foster, Comprehensive exploration of synthetic data generation: A survey, CoRR abs/2401.02524 (2024).

[2] M. Abufadda, K. Mansour, A survey of synthetic data generation for machine learning, in: ACIT, IEEE, 2021, pp. 1–7.

[3] H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, A. Bano, Synthetic data generation: State of the art in health care domain, Comput. Sci. Rev. 48 (2023) 100546.

[4] S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, M. Veloso, Generating synthetic data in finance: opportunities, challenges and pitfalls, in: ICAIF, ACM, 2020, pp. 44:1–44:8.

[5] A. Beretta, G. Ercoli, A. Ferraro, R. Guidotti, A. Iommi, A. Mastropietro, A. Monreale, D. Rotelli, S. Ruggieri, Requirements of explainable AI in algorithmic hiring, in: AIMMES, volume 3744 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024.

[6] Y. Lu, H. Wang, W. Wei, Machine learning for synthetic data generation: a review, CoRR abs/2302.04062 (2023).

[7] J. Dahmen, D. J. Cook, Synsys: A synthetic data generation system for healthcare applications, Sensors 19 (2019) 1181.

[8] A. Torfi, E. A. Fox, Corgan: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records, in: FLAIRS, AAAI Press, 2020, pp. 335–340.

[9] J. Peters, D. Janzing, B. Schölkopf, Elements of causal inference: foundations and learning algorithms, The MIT Press, 2017.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative adversarial networks, CoRR abs/1406.2661 (2014).

[11] D. P. Kingma, M. Welling, An introduction to variational autoencoders, CoRR abs/1906.02691 (2019).

[12] J. Vallverdú, Causality for Artificial Intelligence - From a Philosophical Perspective, Springer, 2024.


[3] https://github.com/jacons/CausalSGD

[13] A. Komanduri, X. Wu, Y. Wu, F. Chen, From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling, Trans. Mach. Learn. Res. 2024 (2024).
[14] G. Pavlidis, Unlocking the black box: Analysing the EU artificial intelligence act's framework for explainability in AI, CoRR abs/2502.14868 (2025).
[15] J. Mazei, N. Backhaus, A. M. Wöhrmann, C. Brauner-Sommer, J. Hüffmeier, Similar, but different: gender differences in working time arrangements and the work–life interface, Collabra: Psychology 9 (2023) 87546.
[16] K. W. Bowyer, N. V. Chawla, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, CoRR abs/1106.1813 (2011).
[17] Y. Kossale, M. Airaj, A. Darouichi, Mode collapse in generative adversarial networks: An overview, in: 2022 8th International Conference on Optimization and Applications (ICOA), IEEE, 2022, pp. 1–6.
[18] A. Schoen, G. Blanc, P. Gimenez, Y. Han, F. Majorczyk, L. Mé, A tale of two methods: Unveiling the limitations of GAN and the rise of bayesian networks for synthetic network traffic generation, in: EuroS&P Workshops, IEEE, 2024, pp. 273–286.
[19] H. M. Combrink, V. Marivate, B. Rosman, Comparing synthetic tabular data generation between a probabilistic model and a deep learning model for education use cases, CoRR abs/2210.08528 (2022).
[20] D. Xu, S. Yuan, L. Zhang, X. Wu, Fairgan: Fairness-aware generative adversarial networks, in: IEEE BigData, IEEE, 2018, pp. 570–575.
[21] M. Abroshan, A. Elliott, M. M. Khalili, Imposing fairness constraints in synthetic data generation, in: AISTATS, volume 238 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 2269–2277.
[22] B. van Breugel, T. Kyono, J. Berrevoets, M. van der Schaar, DECAF: generating fair synthetic data using causally-aware generative networks, in: NeurIPS, 2021, pp. 22221–22233.
[23] F. Mazzoni, M. M. Manerba, M. Cinquini, R. Guidotti, S. Ruggieri, Genfair: A genetic fairness-enhancing data generation framework, in: DS, volume 14276 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 356–371.
[24] E. Barbierato, M. L. D. Vedova, D. Tessera, D. Toti, N. Vanoli, A methodology for controlling bias and fairness in synthetic data generation, Applied Sciences 12 (2022) 4619.
[25] J. D. Smedt, M. le Vrang, A. Papantoniou, ESCO: towards a semantic web for the european labor market, in: LDOW@WWW, volume 1409 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015.
[26] F. Li, A forecaster's review of judea pearl's causality: Models, reasoning and inference, second edition, 2009, CoRR abs/2308.05451 (2023).
[27] A. R. Nogueira, A. Pugnana, S. Ruggieri, D. Pedreschi, J. Gama, Methods and tools for causal discovery and causal inference, WIREs Data Mining Knowl. Discov. 12 (2022).
[28] T. Liu, Learning to Rank for Information Retrieval, Springer, 2011. URL: https://doi.org/10.1007/978-3-642-14267-3. doi:10.1007/978-3-642-14267-3.
[29] E. Pitoura, , et al., Fairness in rankings and recommendations: an overview, VLDB J. 31 (2022) 431–458. URL: https://doi.org/10.1007/s00778-021-00697-y. doi:10.1007/S00778-021-00697-Y.
[30] M. Zehlike, K. Yang, J. Stoyanovich, Fairness in ranking, part I: score-based ranking, ACM Comput. Surv. 55 (2023) 118:1–118:36.
[31] G. K. Palshikar, S. Pawar, A. S. Banerjee, R. Srivastava, N. Ramrakhiyani, S. Patil, D. Thosar, J. Bhat, A. Jain, S. Hingmire, S. Chaurasia, P. Mandloi, D. Chalavadi, RINX: A system for information and knowledge extraction from resumes, Data Knowl. Eng. 147 (2023) 102202.
[32] C. Gan, T. Mori, A few-shot approach to resume information extraction via prompts, in: NLDB, volume 13913 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 445–455.

# A. Data Preprocessing for SDG

The other key input of SDG consists of two datasets of i.i.d. observations of job offers and curricula. From such a dataset, we can derive the structural equations of the SCM as described in Section 3.2. However, several preprocessing steps are necessary to ensure the datasets are of high quality and suitable for modelling. To facilitate this, our system provides a suite of APIs designed to perform various data transformations aimed at standardising and enriching the raw data. We take advantage of a Large Language Model (LLM)[4] to align the raw data with the ESCO taxonomy (see Section 3.1). The task of extracting features from curricula and job offers through LLMs is receiving increasing attention. Recent studies have addressed similar challenges, as illustrated in works such as [31, 32]. It is important to emphasise that while the ESCO and EQF are EU-centric taxonomies, they serve as tools to ensure comparability between job offers and curricula. Our proposed methodology is modular, incorporating external knowledge, such as the O*NET ontology adopted in the US.

In this section, we describe the preprocessing procedures applied to two raw data sources: a real-world dataset of 10,000 job offers ($\mathcal{J}$) and a semi-synthetic dataset of 1,020 curricula ($\mathcal{C}$). The job offers dataset was provided by a major recruiting company based in Spain. The curricula dataset was obtained by the FINDHR project (https://github.com/findhr), which combines features extracted from real-world curricula voluntarily donated for research purposes. Several challenges arise when working with these two data sources. First, the job offers are written in Spanish, whereas the curricula are in English. Second, the underlying taxonomies for education, qualifications, and related attributes differ significantly between the two datasets. The following subsections detail the preprocessing transformations applied to harmonise these datasets, organised by the features represented in the two DAGs shown in Figure 2.

**Preprocessing Education and Qualifications.** Education and qualification naming conventions vary significantly across countries and may also differ between the two datasets, $\mathcal{J}$ and $\mathcal{C}$. For instance, in Spain, educational credentials are typically expressed using the national system, including degrees such as "*Bachillerato*", "*Ingeniería Técnica*", and "*Ciclo Formativo de Grado Superior*". In contrast, English-language curricula often refer to broader categories such as "*Bachelor's Degree*" or "*Master's Degree*" to denote tertiary education levels. To enable meaningful cross-country comparisons, we adopted a mapping developed by domain experts to translate raw educational and qualification data into the European Qualifications Framework (EQF) levels (see Section 3.1). For example, "*Bachillerato*" corresponds to EQF Level 4, while "*Ingeniería Técnica*" maps to EQF Level 6, which is generally equivalent to a "*Bachelor's degree*" in other national systems.

Job offer descriptions typically specify a single education level, usually indicating the minimum required qualification for the position. In contrast, curricula often list multiple educational achievements. Thus, determining a candidate's EQF level in the $\mathcal{C}$ dataset involves several steps. We begin by applying a keyword–based approach to extract and classify educational entries. For example, a phrase such as "BSc in Computer Science" is interpreted as a Bachelor's degree. These inferred degree types are then mapped to their corresponding EQF levels. When a candidate's CV includes multiple qualifications, we select the highest EQF level as a representative indicator of their overall educational attainment.

**Preprocessing Occupation and Job Sector.** In the $\mathcal{J}$ dataset, there is a "job title" feature that captures the title associated with each job offer. However, these titles exhibit considerable variability in both structure and specificity. For example, some entries are well-defined, such as "Data Scientist" or "ICT System Architect", while others are either overly generic (e.g., "Developer") or excessively detailed (e.g., "Web Developer (PHP, JS proficiency) – full remote contract"). This inconsistency often stems from differing practices among HR departments. Some prefer concise titles with detailed descriptions, while others embed extensive information directly into the job title field. Such heterogeneity complicates the task of comparing or categorising job offers in a standardised manner. To address this, we implemented

[4]Specifically, we use *gemma2-9B*, available at https://huggingface.co/google/gemma-2-9b.

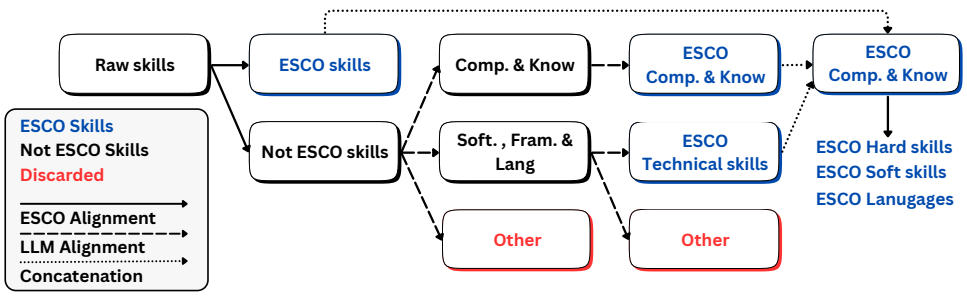


**Figure 4:** Preprocessing pipeline for skills.

a multi-phase alignment process that leverages LLMs and the ESCO taxonomy to normalise job titles across the dataset. Such an alignment process consists of three steps. *Step 1: Title Refinement via LLM.* We use an LLM to generate a cleaner, more representative job title based on the original title and the accompanying job description. This step reduces noise and ambiguity, making titles that better reflect the underlying occupation. *Step 2: ESCO Occupation Retrieval.* We query the ESCO API (https://esco.ec.europa.eu/en/use-esco/use-esco-services-api) using the LLM-generated job title. The API returns a list of relevant ESCO occupations, which serve as candidate labels for standardisation. This step is crucial, as the ESCO search engine performs more effectively when provided with well-structured input. *Step 3: Final Classification via LLM.* We use the LLM again to select the most appropriate ESCO occupation label from the list retrieved in the previous step. This ensures that each job title (i.e., *occupation* in the terminology of Figure 2) is mapped to a standardised occupational category.

For the $\mathcal{C}$ dataset, job sectors were already relatively normalised. Therefore, we applied only the second and third steps of the alignment process: querying the ESCO API and classifying the result using the LLM. Through the above pipeline, we achieved a consistent ESCO–based standardisation of *occupation* and *job sector* across both datasets, enabling reliable comparisons and downstream modelling.

**Preprocessing Skills.** The skill domains in the two datasets differ significantly, making direct comparison non-trivial. The objective of the skills alignment process is to produce an ESCO-compliant list of skills for both $\mathcal{J}$ and $\mathcal{C}$. To achieve this, we design a multi-step procedure that leverages contextual information, specifically, the ESCO *occupation* in $\mathcal{J}$ or the ESCO *job sector* in $\mathcal{C}$, previously aligned (see previous Section 4.1).

The *first step* involves separating skills that are already ESCO-compliant from those that are not. This is accomplished by matching skill terms against the ESCO taxonomy depending on the language of the dataset. Then, we search the terms among the principal concepts, known as preferred labels, and the associated synonyms provided by the ESCO APIs.

In the *second step*, we focus on the non-ESCO skills. These are further classified into three categories using the LLM: "*Competence and Knowledge*", "*Software, Frameworks, Tools and Similar*", and "*Other*". The classification is performed by using an in-context learning approach, where the LLM receives not only the skill to be classified but also contextual information such as the job sector and the surrounding skill set. This context significantly improves classification accuracy by allowing the model to consider semantic relationships beyond isolated terms. For skills categorised as "*Competence and Knowledge*", we query the ESCO API to retrieve a list of relevant ESCO skill terms. The LLM is then used to select the most semantically appropriate term based on the provided context. For "*Software, Frameworks, Tools and Similar*", we first manually selected a set of ten representative ESCO concepts. The LLM then associates each technical skill with the most relevant concept from this list. This step was necessary because the ESCO API often fails to return meaningful results for highly specialised technical terms not covered by the taxonomy. Skills classified as "*Other*" were deemed too noisy or incomplete by the LLM and were excluded from further processing.

In the *final step*, we isolate language-related skills from the remaining set and categorise all ESCO-compliant skills into "*Hard Skills*" and "*Soft Skills*" following the ESCO classification scheme. The result

is a harmonised and structured skill set for each dataset, consisting of three categories: "*Hard Skills*", "*Soft Skills*", and "*Language Skills*". Figure 4 provides a visual overview of the preprocessing pipeline for skills.

**Working Hours and Contract Type.** Job offer descriptions often include various contractual details. Based on this observation, we designed a task for the LLM to extract two specific types of information: *working hours* and *contract type*. For each job offer description, we issue two separate prompts to the LLM, each consisting of a query accompanied by the relevant context. The first prompt aims to classify the job as either a full-time or a part-time contract, based on the information provided in the job description. In cases where explicit references to working hours were absent, the LLM was instructed to infer the appropriate classification from the surrounding context. The second one focused on identifying the nature of the contract duration: "fixed-term" or "permanent". This classification followed the same procedure as the working hours task, relying on both explicit cues and contextual inference when necessary. This approach enabled the extraction of structured contractual information from unstructured job descriptions, contributing to a more comprehensive and standardised representation of job offers. Finally, we emphasise that although we extracted the "Contract Type" feature from the job descriptions, we observed that it did not align with the curriculum characteristics. Consequently, we opted not to include the feature in the experiments to preserve simplicity and interpretability in the results.