Politecnico di Torino

Politecnico di Torino

Master's Degree Thesis

A.a. 2025/2026

Graduation Session March 2026

Auditing Bias in AI-Based Hiring Systems: A Fairness Analysis of Nationality and Gender Discrimination

Supervisors:

Riccardo Coppola

Marco Rondina

Antonio Vetro'

Candidate:

Xhoana Shkajoti

Abstract

This thesis investigates potential bias in AI-based hiring systems, focusing on gender and nationality discrimination. In particular, it audits a screening pipeline built on Sentence-BERT (SBERT), which compares CVs and job descriptions using a semantic similarity score. As these tools are increasingly used to support shortlisting decisions, they may unintentionally reflect or reinforce existing unfair patterns, leading to different outcomes for different groups. The aim of this work is to propose a clear and reproducible audit approach to evaluate fairness in a hiring-like scenario.

To do this, we set up a controlled experiment in which the synthetic candidate profiles are written consistently, so that the only meaningful changes are those related to gender and nationality. The CVs follow the same structure and are similar in length, while the relevant demographic cues vary across profiles. We also include factors that reflect real hiring contexts, such as different writing tones and multiple seniority levels, to reduce accidental noise and make the comparison across groups more reliable.

Since the system produces a similarity score for each candidate, the evaluation is carried out at multiple levels. It examines group differences in score distributions and in average scoring behaviour, and it measures selection-rate gaps under threshold decision rules. In addition, it examines ranking-based outcomes reflecting how screening is often applied in practice. The results suggest that differences that appear small at the score level can become more apparent when ranking decisions are taken into account. It highlights the value of assessing fairness from a usage perspective, because the same system can appear more or less fair depending on whether it is interpreted as a scoring tool or a shortlisting mechanism.

This work contributes to the growing field of ethical and responsible AI by providing a reproducible framework for auditing fairness in hiring-like scenarios. It supports more transparent evaluation of automated screening tools and helps identify potential risks before such systems are applied in real hiring decisions.

Future work can build on this audit by extending the set of roles, job descriptions, and demographic attributes. The same audit setup can be used to compare other model choices beyond SBERT and explore mitigation strategies that aim to improve fairness while keeping the screening system useful.

Acknowledgements

I would like to thank my supervisors for their guidance, feedback, and support throughout the development of this thesis. Their comments and suggestions were important in helping me improve the work and approach the topic with greater clarity and critical perspective. I am especially grateful for their patience, availability, and encouragement during the different stages of this research.

I would also like to thank my family for their constant support and understanding throughout my academic journey. Their encouragement and belief in me have been a source of strength, especially during the more demanding phases of this work.

Finally, I am grateful to everyone who, in different ways, supported me during this period and contributed in my academic journey.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

The use of Artificial Intelligence (AI) in recruitment has grown rapidly in recent years. Companies now rely on automated tools to parse resumes, extract skills, rank applicants, and identify the best matches for job requirements [1].

However, this technological shift raises important ethical and practical concerns. When hiring decisions depend on automated processing of textual information—such as candidate profiles, resumes, or cover letters—the underlying models may reproduce or amplify social biases encoded in their training data [2, 3, 4]. If two candidates with identical qualifications are scored differently solely because of a protected attribute, such as gender, nationality, or other characteristics covered by Article 21 of the EU Charter of Fundamental Rights[5], the recruitment system becomes a potential source of discrimination.

Algorithmic bias in hiring is both a technical and a social-regulatory issue. The EU AI Act treats AI systems used for recruitment and selection—including filtering applications or evaluating candidates—as high-risk (Art. 6(2) with Annex III, 4(a)). This classification establishes specific design and use expectations, including requirements for risk management, transparency, and human oversight [6, 7, 8]. As AI becomes increasingly embedded in human-resources workflows, understanding when and how these models behave unfairly is essential for maintaining trust in automated recruitment technologies and supporting regulatory compliance. This thesis is motivated by the need to evaluate whether widely used language- embedding models, specifically Sentence-BERT (SBERT), introduce systematic disparities when comparing candidate CVs to job descriptions.

1.2 Ethical Considerations

Because this thesis deals with fairness in a high-stakes setting, it is helpful to clarify from the start how we use two terms that will appear throughout the work: *bias* and *stereotype*. In the remainder of this thesis, we use *bias* to mean a systematic gap in the model outputs across demographic groups when candidates are comparable in terms of job-relevant content. In practical terms, this is what we observe if profiles with equivalent qualifications receive different similarity scores, different ranking positions, or different pass/fail outcomes, where the only differences in the text concern demographic cues.

We use the term *stereotype* more cautiously. Here it is not a claim about intent, nor something we aim to prove directly; rather, it is a way to describe the kinds of associations that can be present in large text corpora and that may shape how embedding models encode relationships between words.

Concerns about fairness in these models are not limited to recruitment. Similar questions arise whenever machine learning systems support decisions in sensitive domains such as credit, healthcare, education, or legal risk assessment [9].In such settings, models may appear neutral and consistent while still producing outcomes that differ systematically across groups. One reason is that machine learning systems often learn from historical data, which may contain uneven representation or recurring associations linked to social groups. In the case of transformer-based language models, this concern is particularly relevant because they are trained on large-scale internet text and may therefore encode patterns present in that data [10]. A widely discussed example is Amazon's internal recruiting tool, which was reportedly discontinued after internal assessments suggested that it penalised CVs associated with female candidates. The case is often cited as a reminder that screening systems can inherit problematic regularities from historical data even when discrimination is not explicitly programmed as a rule [11]. For this reason, the ethical issue is not only whether a model performs well on average, but also whether its internal representations contribute to reliable and systematic differences for certain groups.If such systems are deployed without careful evaluation, they may contribute to the reproduction of existing social inequalities.[9]

Conducting fairness evaluations is one way to contribute to the responsible development of AI systems by identifying behaviours that may require mitigation, transparency mechanisms, or changes in system design.
This thesis adopts a controlled and privacy-preserving methodology to explore such issues. By relying exclusively on synthetic data, the study avoids handling real personal information, reduces the risk of harm, and ensures that no individuals are affected by the experimental outcomes. More importantly, the purpose of the research is not to construct predictive tools, but to critically assess an existing modelling approach and reflect on its broader impact. In this sense, the ethical

considerations guiding this work align with the principles emphasised in emerging regulatory frameworks, which highlight the need for robustness, transparency, and fairness in high-impact AI systems. The study aims to support these goals by contributing to a deeper understanding of how language-based models behave and by encouraging more thorough evaluation practices in sensitive application areas. The findings are intended not merely to highlight potential risks, but to support organisations, practitioners, and policymakers in developing more responsible recruitment technologies that align with regulatory standards and societal expectations.

1.3 Context of use

Modern recruitment processes depend more and more on digital platforms and Applicant Tracking Systems (ATS) to manage large volumes of applications[12, 13, 14]. In practice, an ATS supports the hiring workflow by collecting applications, tracking candidates through screening stages, and enabling recruiters to search and filter applicant pools. Many systems also include *resume parsing*, which extracts information such as education, work history, and skills from CV documents and organises it into structured candidate profiles [12].

Once candidate profiles and job descriptions are available in a structured or searchable form, ATS platforms can support early-stage screening by helping recruiters prioritise candidates who appear to best match role requirements. Historically, this has often been done through keyword-based search and rule-based filters[1]. More recently, several solutions have introduced *semantic* matching capabilities intended to retrieve candidates by meaning rather than exact term overlap. These capabilities are typically implemented through learned text representations (embeddings)[15]. For example, Greenhouse describes its Talent Matching component as using embedding representations of skills and job titles to enable semantic search [16]. Similarly, Textkernel describes semantic search and advanced matching components designed to be integrated into recruitment software workflows [17].

In operational settings, matching signals are not merely descriptive: they are often used to sort candidates for review or to narrow down the pool. As documented in the *Help Wanted* report, some hiring tools produce a form of *fit score* and explicitly rank or filter applicants based on this value [14]. This is why the behaviour of matching models is consequential: when scores are used to order candidates or support shortlisting, systematic score differences can affect who is reviewed first and who may be screened out before any human evaluation takes place.

Vendors rarely disclose the exact model architectures used in commercial matching systems. For this reason, the present work focuses on the underlying *mechanism* rather than on a specific product: representing CVs and job descriptions in a shared

vector space and comparing them through a similarity measure. Sentence-BERT (SBERT) is a suitable model for studying this mechanism because it is publicly available, and explicitly designed to produce sentence embeddings that support efficient semantic similarity comparison (e.g., via cosine similarity) [15]. In addition, SBERT-style models have been applied directly to job–resume matching in the research literature using real-world data; for instance, conSultantBERT fine-tunes a Siamese SBERT model on large-scale resume–vacancy pairs to support matching between job seekers and vacancies [18]. Taken together, these considerations motivate the use of SBERT as a transparent and reproducible proxy for analysing potential group disparities in embedding-based matching within an ATS-like screening setting.

1.4 Research gap

Recent research has increasingly explored embedding-based approaches for candidate–job matching, where candidate profiles and job descriptions are represented as vectors and compared through similarity scores [18, 19, 20]. In this context, SBERT provides an efficient bi-encoder architecture for producing sentence/document embeddings suitable for large-scale similarity search [15], and has been adopted in research prototypes for resume screening and candidate–job matching [18, 20].

At the same time, a substantial research body has shown that learned language representations can encode social and demographic associations present in training corpora, including gender stereotypes and other group related regularities [2, 3, 10, 21]. Hiring is a particularly sensitive context for such effects: fairness concerns in algorithmic hiring are well documented, and recent empirical work has reported demographic disparities in language-model-based resume retrieval and screening scenarios [1, 22, 23, 24, 25].

Despite these advances, important gaps remain for similarity-driven screening pipelines. First, SBERT-based recruitment studies often emphasise matching quality and efficiency, while providing limited direct evidence on whether similarity-based scoring and ranking mechanisms yield systematic group disparities in operational screening outcomes [20]. Second, many bias evaluations in NLP rely on representational or association-style tests that do not directly quantify how potential embedding bias propagates into ranking visibility and shortlist decisions in applied pipelines [10]. Third, auditing hiring systems is frequently constrained by limited access to real recruitment datasets and privacy restrictions, which motivates controlled and privacy-preserving audit designs [26, 27, 28]. Finally, existing evidence often centres on a small number of protected attributes (most commonly gender), while attributes such as nationality and the way disparities may vary across job contexts or hierarchy levels receive comparatively less systematic attention in

recruitment-oriented evaluations [22, 24].

Motivated by these gaps, this thesis conducts a controlled audit of a SBERT-based CV job matching mechanism under conditions that reflect early-stage screening. By using synthetic candidate profiles with comparable job-relevant content and systematically varying demographic cues , the study isolates demographic effects without processing real personal data [29, 28]. The analysis focuses on outcomes that matter in practice similarity scores, induced ranking positions, and threshold-based shortlisting to provide empirical evidence on how demographic disparities may emerge in embedding-based recruitment pipelines [1, 27].

1.5 Contributions of the thesis

The guiding question for this research is whether, when qualifications are kept comparable, do demographic cues affect the outcomes of the hiring process.

The contributions of this work can be summarised in three points. First, we build a synthetic and privacy-preserving dataset specifically for fairness auditing, where candidate profiles are comparable in job-relevant content and differ only in selected demographic cues. Second, we analyse the problem in a screening setting that reflects how these systems are used in practice: we consider not only differences in similarity scores, but also how such differences translate into ranking positions and threshold-based shortlisting decisions. Third, we provide a reproducible auditing framework for SBERT-based matching by clearly specifying the evaluation pipeline, the measures used to quantify group differences, and the statistical tests used to assess whether the observed patterns are consistent across job contexts.

1.6 Scope and assumptions

The scope of this thesis is deliberately focused on the CV–job description matching stage, rather than on end-to-end organisational hiring decisions. The analysis relies on synthetic profiles in order to control job-relevant content and to avoid processing real personal data. For this reason, the results should be interpreted as evidence about the behaviour of an embedding-based matching mechanism under controlled conditions, not as a claim about any specific commercial ATS product, whose internal models and deployment choices are typically not observable.

1.7 Thesis Structure

The remainder of this thesis is organised as follows.

Chapter 2 provides the theoretical background and reviews the main concepts and research contributions related to AI-based hiring systems, algorithmic bias, and embedding models used for candidate evaluation.

Chapter 3 presents the methodological framework adopted in the study, including the design of the synthetic dataset, the use of SBERT embeddings, and the procedures for computing similarity scores and assessing fairness.

Chapter 4 reports the results of the empirical analysis, illustrating the behaviour of the similarity scores across demographic groups and the outcomes of the statistical evaluation.

Chapter 5 discusses the findings and their implications, including limitations to be considered and possible future researches.

Chapter 2

Background

2.1 AI-Based Hiring Systems

The recruitment process typically involves multiple stages, ranging from the publication of a job opening to the final selection of a candidate. In contemporary labour markets, particularly within medium and large organisations, the early phases of recruitment are increasingly supported by digital platforms commonly referred to as Applicant Tracking Systems (ATS)[20]. These systems are designed to manage large volumes of applications by automating tasks such as resume collection, parsing, storage, and preliminary screening.

At a functional level, ATS platforms serve as intermediaries between candidates and recruiters. Candidates submit their curricula vitae and related documents, usually in free-text form, while recruiters define job requirements through structured or semi-structured job descriptions. The system's role is to organise this information and assist in identifying profiles that appear most relevant for a given position. Many employers now face large applicant pools per vacancy, which makes fully manual screening impractical and increases reliance on ATS and automated screening tools. For instance, Indeed Hiring Lab reports that in Germany, the average number of applications per job posting increased b in early 2026, while postings declined [30]. In France, Indeed Hiring Lab similarly reports substantial increases in applications per vacancy across several occupational groups between 2023 and 2025 [31]. Consistently, LinkedIn's Economic Graph tracks *applicants per job opening* (relative to a July 2019 baseline) across a set of labour markets that includes France, Germany, and the Netherlands, showing an increase by mid-2025 compared to pre-pandemic levels [32]. This means that manual screening may become impractical in many settings, motivating the adoption of automated decision-support tools.

Early ATS screening often relied on rule-based and keyword-driven searches. For example, documentation for commercial ATS platforms such as Oracle Taleo, SAP

SuccessFactors Recruiting and iCIMS described candidate search functions that use keyword and Boolean logic, among others, to retrieve and filter profiles [33, 34, 35]. In these systems, resumes were evaluated according to the presence or absence of predefined terms corresponding to required skills, technologies, or qualifications. While such methods enabled basic automation, they suffered from limited flexibility and robustness. Keyword-based systems are sensitive to variations in wording and formatting and often fail to capture semantic similarity between different expressions of the same concept. For example, a candidate describing experience as a "backend engineer" may be overlooked for a role seeking a "software developer," despite the functional similarity of the positions.

To address these limitations, more recent AI-based hiring systems have adopted data-driven and machine learning-based approaches. In particular, Natural Language Processing (NLP) techniques have enabled systems to move beyond surface-level term matching toward more nuanced representations of textual content.[36, 37] Resumes and job descriptions are no longer treated as collections of keywords, but as documents whose meaning can be modelled statistically.

A central development in this evolution has been the use of text embeddings. Embedding models transform textual input into numerical vector representations that encode semantic and contextual information. In an embedding space, documents that convey similar meanings are positioned closer together, even if they do not share the same vocabulary. This property makes embeddings particularly suitable for tasks such as resume–job matching, where candidates may describe their skills and experiences using diverse linguistic expressions.

In contemporary AI-assisted recruitment, a common design for resume–job matching is to treat screening as a *semantic scoring and ranking* task. In this design, resumes and job descriptions are first encoded into vector representations using pretrained language models (e.g., transformer encoders such as BERT and its variants RoBERTa, DistilBERT, MiniLM, MPNet, or sentence-embedding models such as SBERT) [36, 38, 39, 40, 41, 15]. A matching score is then computed between the candidate and job representations, and candidates are ranked according to their similarity scores.This ranking system subsequently guides the selection of candidates for more in-depth evaluation or consideration by human recruiters. This general pipeline is used in the resume job matching literature, for instance, in SBERT-based screening approaches that compute similarity-based match scores and use them to prioritise candidates [20, 19, 18].

Similar mechanisms are also described in industry documentation. For example, Greenhouse states that its Talent Matching feature uses embedding representations to enable semantic search over skills and job titles [16]. Textkernel likewise describes semantic search and advanced matching components intended to be embedded into recruitment software workflows [17], and Eightfold describes a talent matching engine that uses embeddings as part of its matching process [42]. Since vendors

typically do not disclose full architectural details, these sources are used here as evidence that embedding-based semantic matching is a practical design choice in real recruitment platforms, rather than as claims about any specific model implementation.

In practice, some recruitment systems extend this embedding-based framework by incorporating additional components, such as domain-specific knowledge bases or named entity recognition techniques[20]. These elements are used to extract structured information related to skills, roles, organisations, or experience from unstructured resume text, and to complement semantic similarity with explicit signals. Such hybrid approaches aim to improve precision and interpretability, particularly in complex hiring scenarios. Nevertheless, embedding-based similarity remains a central mechanism for candidate ranking, especially in large-scale screening contexts where efficiency and flexibility are essential.

Given, the central role of similarity scores in these systems means that the embedding model effectively acts as a gatekeeper in the hiring pipeline. Consequently, the behaviour of the underlying embedding model has a direct influence on access to employment opportunities. This makes AI-based hiring systems a particularly sensitive application domain, where model behaviour must be carefully examined and understood.

The remainder of this chapter builds on this technical overview by examining fairness considerations in recruitment, the mechanisms through which bias can arise in language models, and the role of sentence embedding architectures—such as SBERT—in contemporary hiring systems.

2.2 Algorithmic Fairness and Bias in Recruitment

Hiring decisions are widely recognised as a high-impact domain in which fairness considerations are particularly important. Employment outcomes influence income, career development, and long-term social movement, and are therefore subject to legal and ethical constraints aimed at preventing discrimination[43]. When automated systems are introduced into recruitment processes, questions arise regarding whether such systems treat candidates equitably and whether their outcomes differ systematically across demographic groups.

In the context of automated recruitment, **algorithmic bias** refers to consistent and non-random disparities in outcomes that disadvantage certain groups without a job-related justification [23]. These disparities may emerge even when protected attributes are not explicitly included as input features. A central reason is the presence of *proxy variables*: seemingly neutral features that are correlated with

protected characteristics and can therefore act as indirect signals for group membership. In recruitment, such proxies may include gaps in employment history, educational institutions, geographic information, names, or language patterns that correlate with gender, nationality, or socio-economic background. As a result, removing a sensitive attribute from the input does not guarantee fair outcomes, because discrimination can arise through these correlated variables [9].

Unlike random errors, biased behaviour follows identifiable patterns and may persist across large numbers of decisions, making it particularly problematic in large-scale hiring settings.

A key challenge in assessing fairness in hiring systems lies in the distinction between different forms of bias. Commonly discussed categories include[44, 45, 21]:

- **Direct bias**, which occurs when a system's outputs differ explicitly across demographic groups, for example when candidates associated with a specific gender or nationality are consistently ranked lower than equally qualified candidates from another group.

- **Indirect bias**, which arises when features that appear neutral are correlated with protected characteristics and influence outcomes in a discriminatory manner[9].

- **Historical bias**, which reflects inequalities embedded in past data and social structures that may be learned and perpetuated by automated systems even when sensitive attributes are removed.

Hiring systems are especially susceptible to these forms of bias because they often rely on historical information and large-scale textual data. Past recruitment practices may reflect existing imbalances in representation across roles, industries, or seniority levels[43]. When such patterns are incorporated into automated decision-support tools, they may be reinforced rather than mitigated. Importantly, this process can occur without any explicit intent to discriminate and without the direct use of sensitive attributes.

From a methodological perspective, fairness in hiring can be considered at different levels. Two perspectives are particularly relevant:

- **Group-level fairness**, which focuses on whether aggregate outcomes differ across demographic groups, such as differences in average ranking scores or selection rates [46, 47].

- **Individual-level fairness**, which concerns whether candidates with similar qualifications receive similar treatment regardless of demographic attributes[48].

In ranking-based recruitment systems, individual-level fairness is especially important, as small differences in similarity scores can translate into substantial differences in candidate visibility and selection probability. Because automated tools are often used in the early stages of recruitment to reduce large applicant pools, biased behaviour at this stage may have amplified downstream effects that are difficult to correct later in the hiring process.

Assessing fairness in automated hiring systems is further complicated by issues of transparency and interpretability. Many modern AI models operate as complex statistical systems whose internal representations are not easily accessible to human users. This opacity makes it challenging to identify the sources of observed disparities and to determine whether they stem from legitimate job-related criteria or from unintended correlations with demographic attributes.

These challenges have motivated the development of **audit-based evaluation approaches** [23], which assess fairness empirically by analysing system outputs under controlled input conditions. Audit-based approaches evaluate fairness empirically by treating the system as a *black box*: the auditor does not require access to the model architecture, parameters, or training data, but instead tests the system by observing its outputs under controlled input variations. In practice, this can be done by keeping job-relevant content fixed and modifying only demographic cues in the input to check whether the resulting scores, rankings, or selection outcomes change systematically. This form of black-box testing is particularly suitable for proprietary hiring systems, where internal details are often not accessible for external review [1, 27].

The focus of this thesis aligns with this audit-oriented perspective. By analysing ranking outcomes produced by automated similarity-based hiring systems under controlled conditions, it becomes possible to evaluate whether such systems treat candidates consistently across demographic groups. The following section builds on these fairness considerations by examining how bias can arise within Natural Language Processing models, with particular attention to text embeddings used in ranking tasks.

2.3 Bias in NLP Models and Text Embeddings

Natural Language Processing (NLP) models are increasingly employed to analyse, compare, and rank textual information across a wide range of applications. In embedding-based systems, text is transformed into numerical vector representations that encode semantic and contextual information. While these representations enable powerful downstream tasks, they also reflect statistical patterns present in the data used during training. As a result, biases embedded in language data may be transferred into the representations learned by NLP models[3, 10].

Understanding how such biases arise and how they manifest is essential for evaluating fairness in applications that rely on NLP, including automated hiring.

2.3.1 Sources of Bias in Language Representations

Most modern NLP models are trained on large-scale text corpora collected from diverse sources such as books, news articles, and online content[2]. These corpora reflect social, cultural, and historical contexts in which language is produced. Consequently, patterns of representation within the data may encode stereotypes, imbalances, or associations related to demographic attributes such as gender, nationality, or ethnicity.

Embedding models learn representations based on the **distributional hypothesis**[3], which assumes that linguistic units appearing in similar contexts tend to have similar meanings. While this principle enables models to capture semantic relationships, it also means that socially constructed associations present in the data are reflected in the embedding space. For instance, if certain occupations frequently co-occur with masculine terms in training data, the resulting representations may associate those occupations more strongly with male-related contexts.

Importantly, these biases are not introduced intentionally but emerge as a consequence of statistical learning from real-world language. Because they are embedded implicitly in high-dimensional vector spaces, they may be difficult to identify without targeted analysis.

2.3.2 Bias in Word-Level and Sentence-Level Embeddings

Early research on bias in NLP focused primarily on **static word embeddings**, such as Word2Vec and GloVe[2, 3]. These studies demonstrated that word-level representations encode measurable associations between demographic groups and attributes, including gender stereotypes and occupational roles. Such findings established that embedding spaces can reflect and amplify social biases present in their training data.

More recent systems rely on **contextualised embeddings** produced by transformer based models, which generate representations that depend on surrounding linguistic context. While these models offer richer semantic representations, subsequent research has shown that bias persists at both the word and sentence levels[10]. In sentence-level embeddings, bias may not appear as explicit associations with individual terms, but as subtle shifts in the representation of entire texts that influence similarity measures.

Sentence embeddings are particularly relevant for applications involving document comparison and ranking. In these settings, even minor representational

differences may affect similarity scores, potentially leading to systematic differences in ranking outcomes when documents differ only in limited demographic cues.

2.3.3 Implications for Similarity-Based Ranking Systems

Several evaluation methods have been proposed to quantify bias in embedding models[3, 10], including association-based tests that measure representational differences across demographic groups. While these approaches provide insight into abstract associations, they are often insufficient to characterise how bias influences system behaviour in applied settings.

In similarity-based ranking systems, embeddings serve as the primary decision signal. Because rankings are relative rather than absolute, small differences in representation can result in consistent changes in ordering, particularly near decision thresholds[23]. When such systems are used to prioritise or filter candidates, these effects may be amplified across large applicant pools.

As a result, analysing bias in text embeddings is a necessary step toward understanding fairness in embedding-driven decision-support systems. This motivates application- oriented evaluations that examine how embedding bias translates into observable differences in ranking outcomes.

2.4 Sentence-BERT: Architecture and Use in Hiring

Sentence-BERT (SBERT) is a sentence embedding model specifically designed to generate semantically meaningful vector representations of text[15] that can be efficiently compared using similarity measures.Due to its efficiency and flexibility, SBERT is well suited for applications that require large-scale semantic matching. This section introduces the SBERT architecture, explains its advantages over standard BERT models, and discusses its adoption in hiring-related applications.

2.4.1 From BERT to Sentence-Level Embeddings

Bidirectional Encoder Representations from Transformers (BERT) introduced a major advancement in NLP by enabling deep contextual representations of text. BERT models are trained using self-supervised objectives that allow them to capture syntactic and semantic relationships across entire sequences. However, standard BERT architectures are not optimised for semantic similarity tasks. In particular, comparing two texts requires jointly encoding them using a cross-encoder architecture[15], which is computationally expensive and unsuitable for large-scale ranking scenarios.

As a result, directly applying BERT to tasks such as resume–job matching becomes impractical when thousands of candidate profiles must be compared against one or more job descriptions. This limitation motivated the development of sentence embedding models that preserve BERT's representational power while enabling efficient similarity computation.

2.4.2 SBERT Architecture and Bi-Encoder Design

SBERT addresses these limitations by adopting a **bi-encoder** architecture[15]. Instead of jointly encoding pairs of sentences, SBERT processes each input text independently through a shared transformer network. The output token embeddings are then aggregated using a pooling strategy—such as mean pooling—to produce a fixed-size sentence embedding.

This architectural choice enables each resume and job description to be encoded once and stored, after which similarity scores can be computed rapidly using vector operations. Compared to cross-encoder models, SBERT significantly reduces computational cost and scales effectively to large document collections. This makes it particularly well suited to ranking-based applications where efficiency is critical.

SBERT models are commonly fine-tuned on sentence-pair datasets, such as natural language inference or semantic textual similarity tasks. This training strategy encourages embeddings of semantically similar sentences to be closer in the vector space, while dissimilar sentences are positioned further apart. As a result, SBERT embeddings are well aligned with cosine similarity as a comparison metric.

This ranking-based use of SBERT distinguishes it from classification-oriented NLP applications. Rather than producing discrete predictions, SBERT-based systems generate continuous scores that determine relative visibility within a candidate pool. As a result, fairness concerns arise not only from absolute score differences but also from systematic shifts in ranking position across demographic groups.

2.4.3 SBERT in Recruitment and Resume Screening

Due to its efficiency and semantic expressiveness, SBERT has been adopted in research prototypes for resume screening and candidate–job matching[15, 18, 20]. In these settings, SBERT enables semantic matching by embedding candidate profiles and job descriptions into a shared representation space. The embeddings generated by SBERT can be used to obtain a continuous similarity score between texts, which in return is used directly for ranking [18, 20].

Studies applying SBERT to recruitment tasks have reported improved matching accuracy and flexibility compared to traditional keyword-based methods[20]. SBERT-based approaches are particularly effective at handling diverse linguistic

expressions of skills and experience, making them attractive for multilingual or international hiring contexts[18]. Overall, SBERT acts as the backbone of these pipelines[18, 20], providing the shared representations used for evaluating similarity. Since screening decisions are then driven by similarity scores and the induced ranking, any systematic differences in how candidate groups are represented in the embedding space can translate into disparities in selection outcomes. This motivates auditing SBERT-based screening pipelines using outcome-focused fairness analyses in hiring contexts [1].

2.4.4 Bias Considerations in SBERT-Based Systems

While SBERT offers clear advantages in terms of efficiency and semantic matching, it inherits properties from the data used during pretraining and fine-tuning. Because SBERT models are trained on large corpora of natural language[10, 23], their embeddings may encode social and cultural patterns present in those sources. When SBERT is used in ranking-based decision systems, these representational properties can influence similarity scores in systematic ways.

Since SBERT operates as a black-box model in many deployed systems, identifying and quantifying such effects requires targeted evaluation strategies rather than inspection of model parameters.

For these reasons, SBERT constitutes a realistic and representative model for studying fairness in embedding-based hiring systems. Its adoption in research, reliance on semantic similarity, and role in candidate ranking make it an appropriate focus for audit-based analyses aimed at understanding potential sources of bias in automated recruitment pipelines.

2.5 Synthetic Data in Fairness Auditing

Synthetic data are increasingly used in fairness auditing when the objective is to evaluate model behaviour under controlled conditions without relying on real personal records. This is particularly useful in recruitment, where CVs may contain sensitive information and where access to real applicant data is often limited. By constructing synthetic profiles, researchers can vary selected demographic cues while keeping the professional content as comparable as possible, making it easier to isolate the effect of the attribute under study [27].

A further advantage of synthetic data is that they reduce privacy and re-identification risks compared with real CV collections, especially in settings where demographic information is part of the analysis [28]. They also support reproducibility, since the experimental material can be documented, shared, and reused more easily than proprietary or privacy-restricted datasets. At the same time, synthetic data do not fully capture the variability and complexity of real applications, so they should be understood as a tool for controlled auditing rather than as a perfect substitute for real-world evidence [29].

2.6 Ethical and Regulatory Context

The integration of AI-based systems in recruitment has raised significant ethical and regulatory concerns, particularly in relation to fairness, transparency, and non-discrimination. Because hiring decisions directly affect individuals' employment opportunities and long-term career trajectories, automated recruitment tools are considered high-impact systems that require careful oversight.

Within the European Union, existing regulatory frameworks provide important guidance for the deployment of AI in employment contexts. The General Data Protection Regulation (GDPR) establishes principles related to lawful data processing[26], transparency, and individual rights in automated decision-making. In particular, GDPR emphasises the need for meaningful human oversight and safeguards when algorithmic systems are used to support decisions with significant effects on individuals, such as recruitment.

More recently, the European Union AI Act has explicitly classified AI systems used in employment and worker management as high-risk applications[6]. Under this framework, organisations deploying such systems are required to implement risk management procedures, ensure appropriate data governance, and monitor systems for discriminatory outcomes. While the AI Act does not prescribe specific technical solutions, it highlights the importance of bias assessment and ongoing evaluation throughout the system lifecycle.

In addition to AI-specific regulations, automated hiring tools must comply with

established EU anti-discrimination and equal-treatment law. At the constitutional level, Article 21 of the Charter of Fundamental Rights of the European Union prohibits discrimination on a wide range of grounds [5]. In the employment context, this principle is implemented through the EU equality directives, notably Directive 2000/78/EC (equal treatment in employment and occupation), Directive 2000/43/EC (racial equality), and Directive 2006/54/EC (gender equality in employment), which prohibit unequal treatment in access to employment and working conditions [49, 50, 51]. These legal obligations apply regardless of whether decisions are made by humans or supported by automated tools.

In this regulatory context, audit-based evaluations play a crucial role. By empirically examining how AI systems behave across demographic groups, audits provide evidence that can support compliance with ethical principles and regulatory requirements. The methodological approach adopted in this thesis aligns with this perspective by focusing on the systematic evaluation of fairness in an embedding-based hiring scenario, without relying on proprietary datasets or access to opaque system internals.

Chapter 3

Methodology

This chapter presents a structured framework for the algorithmic auditing of Sentence-BERT (SBERT) within the domain of automated recruitment. As dense vector representations increasingly function as the primary mechanism for matching professional profiles to job opportunities, the need for rigorous and transparent evaluation of their latent biases has emerged as a pressing socio-technical issue [19, 10]. Rather than presuming that SBERT is a neutral tool, this audit conceptualises the model's latent space as a potential source of cultural and linguistic preferences [52].

Through a systematic examination of the model's responses to identical professional experiences presented across diverse demographic backgrounds, this study aims to uncover digital gatekeeping mechanisms that may shape access to professional opportunities [22]. The chapter details the upstream data-generation pipeline, the similarity-computation phase, and the downstream statistical methods used to validate the results.

3.1 Framework Overview and Objectives

3.1.1 Research questions

This thesis examines whether an SBERT-based CV–job matching system can treat candidates differently because of demographic cues, even when their qualifications are kept comparable. In many screening settings, similarity scores are used to sort candidates and decide who gets shortlisted, so small systematic differences can matter.

The study is organized around two research questions:

- **RQ1:** "If two candidates have the same qualifications, does SBERT give different similarity scores or ranking positions just because of gender or

nationality in the CV? And do these differences affect who gets shortlisted?"

- **RQ2:** "Are these differences similar at all hierarchy levels, or do they become stronger in some levels (for example junior vs senior/managerial)?"

To answer these questions, the analysis compares score patterns, ranking outcomes, and shortlisting rates across demographic groups under multiple job contexts, and then checks whether the same patterns hold across hierarchy levels.

3.1.2 Technical Selection: The all-MiniLM-L6-v2 Model

For this research, we chose the all-MiniLM-L6-v2 model. This choice was guided by this main considerations:

- **Consistent output:** In this study, the model is used in inference mode under fixed experimental conditions, so identical input texts produce identical embeddings. This property is important for reproducibility, since the audit relies on repeated and controlled comparisons across candidate profiles.
- **High-Speed Semantic Accuracy:** According to the official Sentence Transformers documentation all-MiniLM-L6-v2 offers a fast embedding generation with around 22 million parameters, ensuring a good balance between computational efficiency and semantic quality. Its lightweight architecture makes it suitable for large-scale matching experiments, while still providing embeddings that are appropriate for similarity-based analysis.[53].
- **Transparency and suitability for audit:** The model was also selected because it is open-source and well documented. This makes the experimental pipeline easier to describe, reproduce, and assess, which is particularly important in an audit-based study.

This model is a distilled version of the Transformer architecture designed for semantic similarity tasks. In the present study, it is used to encode each CV and job description into a 384-dimensional vector representation.

The final embedding is obtained by combining the information processed by the 12 attention heads. Each head contributes a 32-dimensional component to the final vector:

$$\vec{v}_{final} = [h_1 \oplus h_2 \oplus \cdots \oplus h_{12}] \in \mathbb{R}^{384} \tag{3.1}$$

where $\oplus$ denotes the concatenation operator.This combined representation forms the final vector representation of the candidate profile. The audit focuses on this output, since it is the representation used to compute cosine similarity with the corresponding job description.

Identity Compression and the 384-Dimensional Bottleneck

By converting the input text into a fixed-sized embedding means the model compresses the entire professional identity of a candidate into a single point in a 384-dimensional latent space.

This "384-dimensional bottleneck" is where the primary ethical risk resides. When a complex document like a CV is reduced to a single vector $\vec{v} \in \mathbb{R}^{384}$, the model must decide which features are the most "salient." Our methodology tests if the model clusters candidates primarily by professional skills or if the "signal" from protected attributes dominates the compression process. In this high-dimensional space, if the "Nationality" feature carries a high mathematical magnitude, it can effectively "pull" the candidate's vector into a cluster of marginalized identities. By utilizing SBERT, we can mathematically measure the **Cosine Distance** between these compressed identities, providing a quantitative audit of whether the model perceives a "British" candidate as fundamentally more similar to a role than an "Albanian" or "Moroccan" counterpart with identical credentials [10].

3.1.3 Technical Justification for Linguistic Scope

The selection of the linguistic domain is a critical control variable in an algorithmic audit. This research is conducted exclusively in **English**, a decision grounded in technical performance, bias isolation, and regulatory alignment.

Performance Superiority of Monolingual Architectures

Empirical evidence suggests that monolingual models consistently outperform multilingual "jack-of-all-trades" architectures in semantic accuracy and context understanding. A landmark study [54] demonstrated that specialized monolingual BERT models superiorly detect subtle semantic patterns compared to their multilingual counterparts. In the context of this audit, the all-MiniLM-L6-v2 model was trained on over 1 billion English sentence pairs, which makes it better suited to capturing semantic relationships in English than multilingual models trained across many languages [15]. This is important for the present study, since it reduces the risk that the results are affected by translation-related noise or weaker performance across languages.

Comparability and Terminological Standardization

Conducting the experiment in English also makes the results easier to compare with previous studies and established benchmarks, since much of the literature on bias in sentence embeddings and semantic similarity has been developed in English. In addition, the dataset was constructed using **ISCO-08** (International Standard Classification of Occupations) [55] vocabulary, which is standardized in English and offers a common framework for describing occupations. Using this

vocabulary helped keep job titles and professional content consistent across the generated profiles, so that differences in model output were less likely to reflect simple wording variations rather than caused by the attributes under study.

Table 3.1: Language-Bias Control Matrix

Potential Bias	**Mitigation Strategy**	**Scientific Justification**
Translation Noise	Strict use of English templates.	Avoids "Zero-Shot" transfer errors found in multilingual models.
Linguistic Variance	Standardized ISCO-08 vocabulary.	Ensures professional concepts remain stable across all 384 permutations.
Encoding Loss	Monolingual S-BERT architecture.	Monolingual models provide significantly richer semantic representations.

3.1.4 Pooling Strategies: Aggregating Professional Identity

The transition from token-level contextualized embeddings to a singular, fixed-size document vector represents the most reductive and technically significant stage in the SBERT pipeline. While the Transformer layers preserve complex, non-linear relationships between individual tokens, the *Pooling Layer* must condense this high-dimensional information into a single coordinate. For the purposes of this audit, the choice of pooling strategy—CLS, MAX, or MEAN—is a critical decision regarding how a candidate's professional identity is mathematically distilled.

Technical Comparison and Empirical Justification

In a recruitment context, where a multipage Curriculum Vitae must be compared against a concise Job Description, the model must aggregate hundreds of contextual tokens into a single 384-dimensional vector. We evaluate three primary methodologies to justify our selection of **Mean Pooling**, supported by the empirical findings of Reimers and Gurevych [15]:

- **CLS-Token Pooling:** Historically, standard BERT models utilized the output of the first special token, the $[CLS]$ (Classification) token, as the representation for the entire sequence.Originally optimized for classification tasks in BERT it has been shown to perform worse than pooling strategies such as mean pooling when used for semantic similarity tasks [15].
- **MAX-Pooling:** This method selects the largest value for each dimension across all token embeddings in the sequence. In practice, this means that the final sentence vector keeps only the strongest activation that appeared for each feature. As a result, MAX pooling highlights the most prominent words or features in the text. While this can help preserve important keywords, it may

also ignore information distributed across the rest of the document, making the representation less sensitive to the overall context of the CV[15]

- **MEAN-Pooling (The Selected Approach):** This study adopts MEAN-pooling, which calculates the arithmetic mean of all token embeddings produced by the Transformer layers.It has been shown to produce more reliable sentence embeddings for semantic similarity tasks compared to CLS or max pooling strategies[15]. Mathematically, for a sequence of n tokens with embeddings $x_1, x_2, \ldots, x_n$, the resulting document vector $\vec{y}$ is defined as:

$$\vec{y} = \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i \tag{3.2}$$

Why mean pooling matters for the audit.

From the perspective of this audit, the choice of mean pooling is important because the model does not compare individual words or sections of the CV separately. After the Transformer produces token-level embeddings, these are aggregated into a single representation that is later used to compute similarity with the job description. Mean pooling performs this step by averaging information across the sequence, so the final vector reflects the overall content of the document rather than only one special token or only the strongest token-level activations [15]. This matters for the audit because the final similarity score is based on that pooled representation. As a result, professional qualifications, work experience, skills, and any demographic cues present in the text are brought together in the same final representation. In other words, once pooling is applied, the model no longer evaluates an engineer who happens to be Italian as two fully separate pieces of information. Instead, it compares a single aggregated profile representation, in which occupational and demographic cues may both be reflected. In this sense, if demographic information influences the embedding process, that influence may also appear in the final representation used for matching [15, 10].

The Mathematical Calculus of Similarity

Once the Curriculum Vitae (A) and the Job Description (B) have been processed through the Siamese towers and synthesised via the pooling layer, they exist as fixed-length vectors in $\mathbb{R}^{384}$. To determine the semantic proximity between these two points, the model calculates the **Cosine Similarity**, which measures the cosine of the angle θ between the two vectors [15].

The formula for this operation is defined as the dot product of the vectors divided by the product of their Euclidean magnitudes:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.3}$$

In the context of this thesis, this value serves as the primary metric for the automated ranking system:

- **Unity ($\cos(\theta) = 1.0$):** Implies a perfect semantic alignment ($\theta = 0°$), where the candidate's skills are a mathematical mirror of the job requirements.
- **Orthogonality ($\cos(\theta) = 0.0$):** Implies total semantic independence ($\theta = 90°$), suggesting no relevant correlation between the two documents.
- **Diametric Opposition ($\cos(\theta) = -1.0$):** Implies the documents are semantically opposite. This is theoretically possible but rare in recruitment contexts due to the nature of the Transformer's embedding space [10].

Methodological Justification: Choice of Cosine Similarity

Cosine similarity was selected as the similarity metric for comparing CV embeddings with job descriptions because it is a standard approach used to evaluate semantic similarity between sentence embeddings in SBERT-based systems[15]. Instead of measuring the absolute distance between vectors, cosine similarity evaluates the angle between them, capturing how closely their semantic representations align in the embedding space. This makes it particularly suitable for text similarity tasks, where the goal is to compare the meaning of two pieces of text rather than their absolute vector magnitude. The use of cosine similarity for comparing embeddings is also widely established in natural language processing research. Many studies analysing semantic relationships and bias in word embeddings rely on cosine similarity to measure associations between vector representations [2, 3, 15]. This widespread adoption supports its suitability for ranking the semantic similarity between CVs and job descriptions in embedding-based matching systems.

Quantifying Bias through the Similarity Difference (Δ)

The core of our auditing methodology relies on measuring the **Similarity Difference** (Δ). Utilising the factorial design established in Section 3.1, we maintain constant professional merit (M) while toggling only the protected demographic attribute (I).

If we compare a British candidate vector (A_{Brit}) and an Albanian candidate vector (A_{Alb}) with identical professional credentials, their similarity to a Job Description (B) should, in a fair system, be identical. Any observed deviation represents a **Model similarity difference** (Δ):

$$\Delta = |\cos(\theta)_{Brit} - \cos(\theta)_{Alb}| \tag{3.4}$$

In an unbiased system, Δ should be approximately zero. Our research investigates a "change in the embedding representation"—a phenomenon where the

inclusion of an identity marker causes the candidate's vector to rotate away from the "job description" resulting in a lower similarity score [10, 56].

3.2 Experimental Design

3.2.1 Experimental Design: The Synthetic Corpus

The integrity of an algorithmic audit is entirely dependent on the quality and control of the input data [29]. While many studies rely on historical datasets, evaluating the behavior of high-dimensional language models such as SBERT benefits from carefully controlled experimental data [15]. This section outlines the development of the Synthetic Corpus[29], a dataset constructed to provide a controlled environment for analysis [9, 1, 43]. The dataset is specifically designed to separate demographic variables from indicators of professional merit, thereby enabling a focused examination of model behavior.

Rationale for Controlled Synthetic Data

A primary challenge in AI auditing is the "Black Box" nature of both the model's architecture and its original training data. If this research were to rely on real-world CVs, it would face a significant methodological challenge: the confounding variable problem [9, 57]. In real-world datasets, professional merit and demographic identity are frequently intertwined [9]. If a model assigns a lower score to a Moroccan candidate within such data, it becomes difficult to determine whether this outcome is due to bias or to other factors related to the dataset, such as differences in educational background or linguistic expression.

To achieve the level of scientific rigor necessary for a technical audit under the EU AI Act, this study therefore moves from observational data toward a controlled experimental setup based on counterfactual comparisons[56, 6].

The Counterfactual Framework and "Ceteris Paribus" To isolate the effects of nationality and gender, this methodology applies a strict *ceteris paribus* condition [58]. The *ceteris paribus* condition ensures that all other variables remain constant. Synthetic CVs make it possible to enforce this condition, making sure that each indicator of professional merit remains identical across demographic variants of a given profile.

This approach enables the evaluation of model behavior through a counterfactual logic. Let f denote the S-BERT scoring function, and let CV represent the vector representation of a candidate. We define two candidates, CV_{it} (Italian) and CV_{ma} (Moroccan), both of whom possess identical professional qualifications. In a

perfectly objective system, the model's output should satisfy the following condition of Demographic Parity:

$$f(CV_{it}) = f(CV_{ma}) \tag{3.5}$$

Therefore, any observed deviation in the similarity scores can be attributed solely to the demographic attributes under investigation:

$$\text{if } f(CV_{it}) \neq f(CV_{ma}) \rightarrow \text{Systemic Bias Detected} \tag{3.6}$$

Addressing Proxy Variables A particularly challenging form of algorithmic bias arises from proxy variables—data points that appear neutral but in fact encode demographic information [59, 9]. The use of synthetic data allows these potential proxies to be controlled. Removing geographic and cultural proxies and retaining only the identity tokens (nationality and gender) and the merit tokens (the professional content), makes possible to create a high-resolution environment for measuring the model's response to identity in its most isolated form.

3.2.2 Factorial Design Matrix

We designed a full factorial matrix. This structured approach allows for the isolation of specific "Bias Drivers" by ensuring every professional variable is crossed with every demographic variable. The total dataset comprises 384 unique CV permutations, derived from the interaction of 4 job roles, 4 seniority levels, 4 nationalities, 2 genders, and 3 lexical versions.

Dimension	Levels / Description
Job roles	Project Manager Market Research Analyst HR Specialist Content/Creative Manager
Hierarchy levels	Junior (0–2 years) Mid-Level (3–5 years) Senior (6–10 years) Lead/Manager (10+ years)
Nationality	Italian Albanian British Moroccan
Gender	Male Female
Lexical versions	**V1** – Formal and Institutional: neutral phrasing, standardized ISCO-08 style **V2** – Leadership and Action-Oriented: power verbs, more assertive narrative **V3** – Technical and Competency-Centric: tools, methodologies, and functional skills emphasized

Table 3.2: Variables included in the factorial design matrix

Rationale for Nationalities Selection

The nationalities used in this experiment were chosen to reflect different migration and employment contexts within the European labour market, based on empirical evidence from European statistical reports:

- **The EU Reference Category (Italy):** Italy represents the EU-member reference category in the nationality design and reflects a Southern European profile within the set of nationalities considered in the study. For this reason, it was selected as a comparator and used as the baseline against which the effect of other nationality cues can be assessed.

- **The Anglo-European Comparator (United Kingdom):** The United Kingdom represents a suitable nationality for inclusion because it is geographically close to continental Europe, reflects a Western developed-country profile, and carries a distinct Anglo linguistic and cultural identity. This makes it a

useful case for testing how the model responds to a nationality cue shaped by both European proximity and Anglo cultural familiarity.

- **Non-EU Mediterranean Context (Morocco):** Morocco was included to represent a non-European nationality frequently examined in studies of migrant integration and labour market discrimination in Europe. Surveys conducted by the European Union Agency for Fundamental Rights [60] report relatively high levels (around around 40–50%)of perceived employment discrimination among respondents of North African origin within EU labour markets. Furthermore, 2024 Eurostat [61] data shows that foreign-born women experience a gender employment gap of **15 to 20 percentage points**, considerably larger than that of native-born women.

- **The Balkan Integration (Albania):** Albania represents a unique European but non-EU nationality. Prague Process [62] reports indicate that Albanian nationals held more than **839,000 valid EU residence permits**,making them one of the largest migrant groups from the Western Balkan region. Research on migrant labour market integration also reports cases of skills mismatch and overqualification, where migrants' professional qualifications are not always fully recognised or are utilised in lower-skilled occupations within host labour markets.

Gender Attribute Representation

Gender in the synthetic CVs is represented using binary markers (male and female). This choice was made to maintain control within the experimental design. The use of a binary representation reflects a methodological simplification necessary for controlled experimentation and does not aim to capture the full diversity of gender identities in real-world contexts.

Seniority Levels

The synthetic CV dataset is structured across four seniority levels (Junior, Mid-level, Senior, and Managerial). This hierarchy allows the analysis to examine whether the model's behaviour changes across different stages of the professional career path.

Including multiple levels makes it possible to explore the hypothesis that demographic cues may influence the model differently depending on job seniority. One possible expectation is that disparities may decrease as professional qualifications become more explicit at higher levels, while another possibility is that biases may emerge more strongly in senior roles where attributes such as leadership potential or organisational fit may play a greater role in candidate evaluation. Labour market studies have documented cases of overqualification and skills mismatch among

migrant populations in Europe, where professional experience is not always fully recognised in host labour markets [63, 64]. These patterns provide additional motivation for examining how automated systems evaluate candidates across different experience levels

Institutional Standardization: The ISCO-08 and Hierarchical Design

To ensure consistency in the professional hierarchy, the job levels used in the synthetic CVs were aligned with the International Standard Classification of Occupations (ISCO-08) [55]. This standardized framework distinguishes between different occupational skill levels and provides a common reference for comparing professional roles across contexts. In the present study, the profiles correspond broadly to higher professional categories, ranging from technical and professional roles ,typically associated with ISCO Skill Level 3, to strategic and managerial positions aligned with Skill Level 4.

Within this framework, the synthetic CV dataset was designed to represent four stages of professional experience: Junior, Mid-Level, Senior, and Managerial. The distinctions between these levels are expressed primarily through differences in years of professional experience and in the scope of responsibilities described in the work history.

- **Junior (0–2 years):** Represents early-career candidates with limited professional experience. Work descriptions focus primarily on supporting tasks, assisting senior staff, and contributing to operational activities.
- **Mid-Level (3–5 years):** Represents professionals with increasing autonomy and responsibility. These profiles typically include project coordination, independent task management, and greater involvement in operational decision processes.
- **Senior (6–10 years):** represents experienced professionals responsible for strategic planning, supervision of teams, and evaluation of project outcomes.
- **Lead/Manager (10+ years):** Represents leadership roles involving organizational coordination, strategic decision-making, and oversight of departmental activities.

3.3.5 Linguistic Robustness: The Three-Version Protocol

To reduce the possibility that the results depend on specific wording patterns, the synthetic corpus implements a three-version protocol (V1, V2, V3) for every candidate profile. Each version describes the same professional history and qualifications but varies in vocabulary and sentence structure. This approach ensures that the

audit evaluates the model's response to the underlying professional content rather than to the tone of language used.

Each version adopts a slightly different linguistic tone:

- **V1: Formal and Institutional (Standardized):** This version utilizes neutral, passive-voice descriptors and standardized ISCO-08 terminology [55](e.g., "Responsible for the coordination of...") to serve as a baseline reflecting a conventional CV style aligned with formal organizational language.

- **V2: Leadership and Action-Oriented (Deterministic):** This version employs high-impact "power verbs" (e.g., "Spearheaded global initiatives") to reflect a narrative emphasizing initiative and leadership.

- **V3: Technical and Competency-Centric (Functional):** This version focuses on specific tools and methodologies (e.g., "Utilized Agile methodologies") to highlight technical competence and domain-specific expertise.

The use of multiple textual versions also helps reduce the influence of specific wording choices on the similarity scores produced by the model. Transformer-based sentence encoders can produce slightly different embeddings when the same information is expressed using different vocabulary or sentence structures [10]. By evaluating the similarity scores across the three versions of each profile, the analysis reduces the risk that the results depend on writing style of the CV. In addition, this variation mirrors real-world conditions, where CVs describing similar professional experiences may differ in tone, emphasis, and narrative style depending on how candidates present their skills and achievements.

This approach relies on controlled counterfactual variation, changing only the attribute under examination [10]. Similar strategies have also been used in research on bias in sentence encoders, where demographic attributes are modified while keeping the rest of the text constant [10]. If similar patterns appear across all three versions, the observed differences are less likely to be caused by stylistic phrasing and more likely to reflect the demographic attributes under investigation.

Occupational Selection

The occupations included in the study—Project Manager (ISCO 2421), Market Research Analyst (ISCO 2431), HR Specialist (ISCO 2423), and Content/Creative Manager (ISCO 1222)—were selected to provide a coherent professional setting for the audit experiment. These roles are suitable for semantic matching because candidate evaluation in these fields depends strongly on written descriptions of experience, responsibilities, coordination skills, analytical abilities, and soft skills. For example, a Project Manager may be evaluated through descriptions such as

leading cross-functional teams, coordinating project timelines, or managing stakeholder communication. Similarly, Market Research Analysts or Content/Creative Managers are often assessed through descriptions such as conducting market analysis, interpreting consumer trends, or developing communication strategies. In these roles, professional experience is largely expressed through narrative descriptions in both CVs and job postings.

This contrasts with occupations where hiring decisions depend more strongly on formal qualifications or regulated credentials. For example, professions such as nursing typically require professional licenses and certified clinical training, while many engineering roles emphasize specific degrees or highly specialized technical competencies. Although experience in these professions may also be described in text, the initial evaluation of candidates often relies more heavily on structured credentials than on the interpretation of narrative descriptions. For this reason, professional service roles that rely more extensively on textual descriptions of responsibilities and achievements provide a suitable context for analysing a model that compares documents using semantic similarity.

A second reason for selecting these occupations was to avoid professions whose titles may themselves function as strong gender proxies. Terms such as nurse, or some highly male-coded technical titles, can carry strong gender associations in both labour-market patterns and language representations [9, 2, 3, 10]. In such cases, differences in model output would be harder to interpret, because the profession label itself may already introduce a strong gender signal.

A further advantage of these occupations is that they can be clearly mapped to the *International Standard Classification of Occupations* (ISCO-08) framework provided by the International Labour Organisation [55]. This makes it possible to define the professional content of both CVs and job descriptions in a standardized way and to maintain comparability across the dataset.

Taken together, these characteristics make the selected occupations appropriate for the purposes of the study: they are text-rich professional roles, they can be standardized through an established occupational framework, and they provide a more interpretable setting for auditing whether demographic attributes influence the behaviour of a semantic matching system.

Job Description (JD) Construction

To ensure that the job descriptions used in the audit are based on a consistent professional reference, they were constructed using the International Standard Classification of Occupations (ISCO-08) framework provided by the International Labour Organisation (ILO) [55].

Instead of collecting job postings from online job boards, which often vary widely in wording and may include company-specific language or formatting, the descriptions were derived directly from the official ISCO unit group definitions.

For each occupation, the core tasks and duties listed in the corresponding ISCO category were used as the basis for constructing the job description.

This approach provides a standardized description of the professional requirements associated with each role and ensures that the job descriptions reflect internationally recognised occupational definitions rather than individual company-specific phrasing. Using the same occupational framework for both the job descriptions and the synthetic CVs also helps maintain consistency across the dataset, allowing demographic variables to be analysed while the professional context remains constant.

3.3 Research Pipeline: A Step-by-Step Workflow

3.3.1 System Architecture Overview

The research pipeline figure 3.1 is organised as a modular, multi-stage system for analysing potential shifts in the 384-dimensional embedding space. Its structure separates the data generation phase from the computational analysis phase, ensuring that the synthetic profiles are constructed independently of the model outputs. This design helps preserve the controlled nature of the audit and reduces the risk of unintended interactions between experimental stages. The pipeline is therefore divided into four main functional stages, each serving to maintain the consistency and interpretability of the counterfactual analysis

1. **The Generation Tier:** Synthetic Corpus Expansion via `mass_generator.py`.
2. **The Encoding Tier:** Semantic Vectorization and Embedding via `run_audit.py`
3. **The Comparison Tier:** Geometric Similarity Engine using Cosine Similarity.
4. **The Analytical Tier:** Statistical Post-Processing and Result Persistence.

.

3.3.2 Stage 1: The Generation Tier (`mass_generator.py`)

This stage converts a small set of manually written templates into the full set of unique synthetic candidate profiles used in the experiment.

Deterministic Placeholder Logic

The predefined template-based approach for the pipeline is chosen to maintain strict experimental control.

- **Syntactic Invariance:** The system loads base `.txt` templates from the `cv_dataset` directory. These templates serve as the fixed semantic anchors of the study.The complete set of templates is available in the corresponding project directory of the repository.
- **Identity Token Substitution:** The script performs a literal string replacement of the `[NATIONALITY]` and `[GENDER]` markers. By using deterministic replacement, the system ensures that for every iteration, the only fluctuating variables are the demographic identity tokens. The full dataset of CVs is save into `cv_dataset_expanded` folder of the repository.

- **Meritocratic Parity:** Because only these specific tokens change, the underlying professional structure of the CV—the fixed skeleton—remains the same across all versions. This controlled isolation allows any variation in similarity scores to be attributed to the nationality and gender markers being tested.

3.3.3 Stage 2: The Encoding Tier (`run_audit.py`)

In this tier, the CVs and JD synthetic text is transformed into mathematical coordinates using the Sentence-BERT (S-BERT) backbone.

Transformer Architecture: all-MiniLM-L6-v2

The pipeline utilizes the all-MiniLM-L6-v2 model, a distilled version of the BERT architecture [40]optimized for clustering and semantic search.

- **Dimensionality:** Each CV adn JD is mapped to a 384-dimensional dense vector space

- **Tokenization and Truncation:** The all-MiniLM-L6-v2 model has a limit of truncation of 256 word pieces. By keeping documents under this limit, the pipeline ensures that the professional profile—from the the header of the file to the footer—remains within the model's active attention window without being cut off. Transformer models utilize a self-attention mechanism that allows every token to "attend" to every other token[37]. By staying well below the 256 ceiling, the model can simultaneously process the relationship between the candidate's identity markers and their professional achievements, resulting in a more reliable embedding of the full profile.

- **Document Length Standardisation:** Literature suggests that similarity scores can be skewed if document lengths vary significantly [23, 3]. By standardising all files to a uniform word count, the pipeline ensures that the Cosine Similarity ($\vec{S_c}$) results are a reflection of the content and identity rather than the volume of text processed. All templates were designed to be around 180-200 words. This ensures they fit within the model's optimal processing window, preventing "Length Bias"[3, 23] where longer documents might receive higher scores due to sheer keyword density.

3.3.4 Stage 3: The Comparison Tier (Geometric Engine)

The core of the audit is the calculation of the semantic relationship between the candidates and the Job Descriptions (JDs).

Reference Vector Mapping

Four distinct JDs act as "Reference Vectors"—fixed coordinates in the latent space representing the "Ideal Candidate" for each seniority level. By maintaining the JD as a geometric constant, we can quantify the changes in similarity produced by nationality or gender markers.

Cosine Similarity Operationalization

The engine calculates the angular distance between the JD vector ($\vec{A}$) and the CV vector ($\vec{B}$) using the formula:

$$S_c = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|} \tag{3.7}$$

- **Magnitude Independence:** Since cosine similarity evaluates the relative orientation of the two vectors, it indicates how closely the CV vector aligns with the JD vector and gives a continuous score to that semantic similarity.
- **Precision Management:** The script utilizes `float32` precision to preserve small shifts in similarity scores, since even subtle numerical differences may be relevant when examining potential demographic effects.

Stage 4: The Analytical Tier

The final stage consisted of exporting and organising the raw similarity scores into a structured dataset, which was then used for the statistical evaluation described in Section 3.4.

Data traceability The similarity scores were saved as `final_bias_audit_results.csv` in the `results` folder. Each record retained the corresponding JD and CV filenames and the structered list of attributes related to it, allowing every similarity score to be traced back.

Score organisation The scores were grouped by nationality, gender, hierarchy level, and linguistic version. This structure made it possible to compare the results across demographic groups while preserving the distinctions between profile versions.

Preparation for statistical analysis Once exported, the dataset was reviewed and organised in Excel so that the matched comparisons required for the statistical tests could be constructed. In particular, the data structure allowed the formation of

gender-based matched pairs and nationality-based matched profile blocks, consistent with the evaluation framework described in Section 3.4.

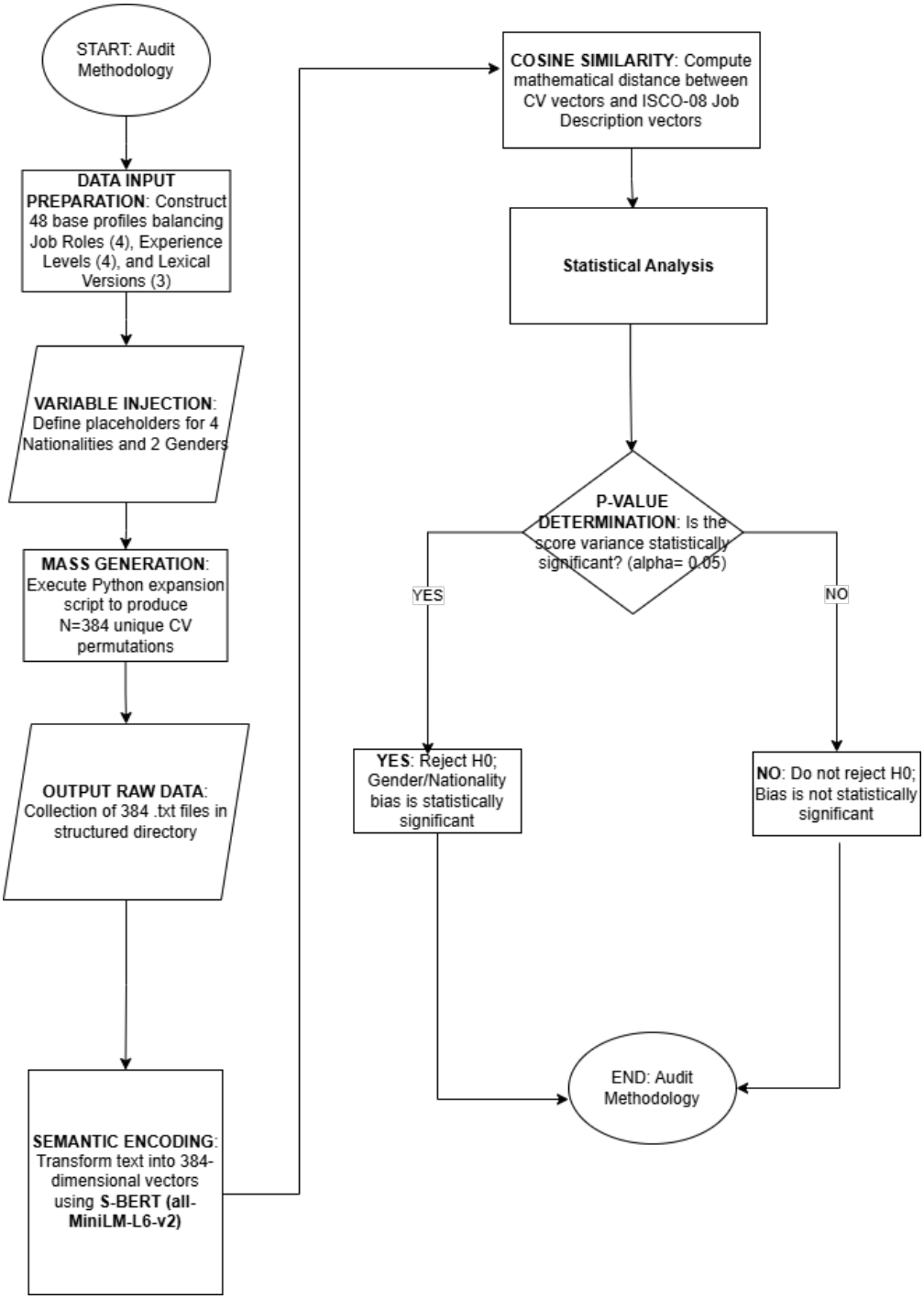

Figure 3.1: Audit Methodology Flowchart: From data input preparation to statistical bias determination.

3.4 Statistical Evaluation Framework

3.4.1 General Analytical Approach

The statistical analysis builds directly on the controlled structure of the synthetic dataset. The experiment combines four nationalities, two genders, four hierarchy levels, and three template versions, resulting in a balanced set of synthetic CVs. Within this design, profiles are organised so that professional content is held as constant as possible, while only one demographic attribute is varied at a time. This makes it possible to construct pairs or matched profile sets. The profiles belong to the same experimental condition and are used to isolate the effect of the demographic attribute under examination. Under these conditions, any systematic difference in the model output can be interpreted more directly as being associated with the modified demographic cue, rather than with differences in qualifications, experience, or role alignment described in the CV[43], consistent with a counterfactual audit logic[56].

This matching logic is central to the way the tests are applied. For gender, the analysis is based on matched male–female pairs that share the same job description, nationality, hierarchy level, and template version. For nationality, the analysis is based on matched four-profile blocks in which nationality varies while the other characteristics remain fixed. To create the pairwise comparison, in this case we use one nationality as baseline. This matching structure defines the basis on which the statistical tests are performed. The evaluation is performed in three complementary forms. First, the score-level analysis examines the raw similarity scores. Second, the threshold-based analysis converts those scores into binary screening decisions. Third, the ranking analysis examines the relative ordering of candidates once they are sorted by score.

The same analytical logic is applied both at aggregate level and by hierarchy level. The aggregate analysis provides an overall view across the full dataset, while the hierarchy-level analysis makes it possible to verify whether the observed patterns remain stable across junior, mid-level, senior, and managerial profiles.

3.4.2 Score-Level Analysis

The first part of the analysis focuses on the raw similarity scores produced by the model. The statistical comparison is based on score differences computed within pairs.

For gender, each male CV score is compared with its female counterpart. For nationality, since we have more than 2 possible values for it, each profile is compared with the it corresponding Italian counterpart profile. The null hypothesis for both

gender analysis and nationality analysis with italian as a baseline is:

$$H_0 : \mu_d = 0$$

where μ_d is the mean of the paired score differences.

For each matched pair, the score difference was calculated as:

$$d_i = s_{1,i} - s_{2,i}$$

where $s_{1,i}$ and $s_{2,i}$ are the similarity scores of the two matched profiles.

A paired t -test was applied to the differences.The test shows whether that average difference is large enough, relative to the variation across pairs, to suggest that the model is responding systematically to that attribute rather than producing random fluctuations. The t-test evaluates whether the observed difference is statistically associated with the attribute under consideration, but it does not measure the magnitude of the difference itself. The paired t-test statistic was computed as:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where $\bar{d}$ is the mean of the paired differences, s_d is their standard deviation, and n is the number of matched pairs.

The p-value was obtained from the t distribution with $n - 1$ degrees of freedom. In Excel, the p-value can be calculated with:

```
T.DIST.2T(ABS(t), n-1).
```

For nationality, the pairwise comparisons relative to Italy were complemented by the Friedman test, which evaluates whether nationality has an overall effect across the four matched nationality conditions. The Friedman test is appropriate when comparing more than two related samples measured within the same experimental blocks.

The null hypothesis is:

$$H_0 : \text{Nationality has no effect on the similarity scores.}$$

In other words, the four nationality-specific profiles within each matched block are expected to receive similar scores on average.

To compute the test statistic, the four scores within each block are first converted into ranks from 1 to 4. The ranks are then summed across all blocks for each nationality. Let R_j denote the total rank sum for nationality j. The Friedman statistic is calculated as:

$$\chi^2 = \frac{12}{n\,k(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

where n is the number of matched blocks and k is the number of nationality groups. Larger values of χ^2 indicate stronger differences in scores across the nationality conditions.

The p-value is obtained from the chi-squared distribution with $k-1$ degrees of freedom. In Excel, it can be calculated as:

$$p = \texttt{CHISQ.DIST.RT}(\chi^2, k-1)$$

3.4.3 Threshold-Based Screening Analysis

The second part of the analysis examines the model output after similarity scores are converted into binary screening decisions. This step simulates a pass/fail screening process. In this sense, the threshold serves as an operational proxy for the minimum score required by a candidate to pass the screening stage.

To represent different levels of screening strictness, three operating thresholds were defined: Top 10%, Top 20%, and Top 30%. Candidates above the relevant threshold were coded as selected (1), while the remaining candidates were coded as not selected (0).

In the aggregate analysis, the threshold was applied to the full set of CVs. Thus, under the Top 10% condition, only the highest-scoring 10% of all CVs were treated as having passed the screening stage. In the hierarchy-level analysis, the same rule was applied separately within each hierarchy group. For example, under the Top 10% condition, only the highest-scoring 10% of CVs within a given hierarchy level were coded as selected.

For gender, the resulting binary outcomes were compared within matched male–female pairs. The null hypothesis is:

$$H_0 : P(M = 1, F = 0) = P(M = 0, F = 1)$$

This means that male and female profiles are assumed to have the same probability of being selected. To test this hypothesis, McNemar's test was applied. McNemar's test is appropriate for paired binary outcomes and focuses only on the discordant pairs, that is, cases in which the two matched profiles receive different selection outcomes.

Let:

- b denote the number of cases in which the male profile was selected and the female profile was not;
- c denote the number of cases in which the female profile was selected and the male profile was not.

The McNemar test statistic was computed as:

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

The p-value was obtained from the chi-squared distribution with 1 degree of freedom. In Excel, it can be calculated as:

$$\texttt{CHISQ.DIST.RT}(\chi^2, 1)$$

For nationality, the comparison involved four related binary outcomes within each matched block. In this case, Cochran's Q test was used, as it extends McNemar's test to more than two related conditions. The null hypothesis is:

$$H_0 : P(\text{selection} \mid IT) = P(\text{selection} \mid AL) = P(\text{selection} \mid MO) = P(\text{selection} \mid BR)$$

This means that the four nationality conditions are assumed to have the same probability of selection within the matched blocks.

To compute the test, each profile was coded as 1 if selected and 0 otherwise. The total number of positive outcomes was then summed both by nationality condition and by matched block. Let C_j denote the total number of selections for nationality j, and let R_i denote the total number of selections within block i. The Cochran's Q statistic was computed as:

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} C_j^2 - \left(\sum_{j=1}^{k} C_j\right)^2\right]}{k\sum_{i=1}^{n} R_i - \sum_{i=1}^{n} R_i^2}$$

where n is the number of matched blocks and k is the number of nationality groups. Larger values of Q indicate stronger differences in selection probabilities across nationality conditions. Under the null hypothesis, the Cochran's Q statistic Q follows approximately a chi-squared distribution with $k-1$ degrees of freedom. In Excel, the corresponding p-value can be calculated as:

$$\texttt{CHISQ.DIST.RT}(Q, k-1)$$

Alongside the inferential tests, descriptive disparity measures were also reported. These include the selection rate, the risk difference, and the risk ratio.

- **Selection Rate (SR):**

$$SR_g = \frac{\text{Selected in group } g}{\text{Total in group } g}$$

Question answered: What proportion of candidates in group g is selected?

- **Risk Difference (RD):**

$$RD_{g-r} = SR_g - SR_r$$

Question answered: By how much does the selection probability of group g differ from that of the reference group r in absolute terms?

- **Risk Ratio (RR):**

$$RR_{g/r} = \frac{SR_g}{SR_r}$$

Question answered: How many times more or less likely is selection in group g compared with the reference group r?

These measures are useful because they show not only whether a difference is statistically significant, but also how large it is in practical terms.

3.4.4 Ranking-Based Analysis

The third part of the analysis examines ranking outcomes. Candidates were ordered according to their similarity scores and evaluated in terms of Top-K inclusion, with K = 1,3,5. This choice reflects a practical recruitment setting, where the central question is often which profile emerges as the best candidate among the applications. For values of K greater than 1, the analysis instead considers whether a profile appears among a fixed number of the best-performing candidates.

For each value of K, every candidate was coded as included (1) or not included (0) in the Top-K set:

$$z^{(K)} = \begin{cases} 1 & \text{if candidate is ranked in the Top-}K\text{ positions,} \\ 0 & \text{otherwise,} \end{cases}$$

For gender, the resulting binary outcomes were compared within matched male–female pairs. The null hypothesis is:

$$H_0 : P(\text{Top-}K \mid M) = P(\text{Top-}K \mid F)$$

This means that the probability of appearing in the Top-K positions is assumed to be the same for the two matched profiles. McNemar's test was applied again in this case, as the comparison involves paired binary outcomes.

The p-value was obtained from the chi-squared distribution with 1 degree of freedom. In Excel, it can be calculated as:

$$\texttt{CHISQ.DIST.RT}(\chi^2, 1)$$

For nationality, the comparison involved four related binary Top-K outcomes within each matched block. The null hypothesis is:

$$H_0 : P(\text{Top-}K \mid IT) = P(\text{Top-}K \mid AL) = P(\text{Top-}K \mid BR) = P(\text{Top-}K \mid MO)$$

This means that the probability of appearing in the Top-K positions is assumed to be the same across all four nationality conditions. Cochran's Q test was used to evaluate this hypothesis, as it extends McNemar's test to the case of more than two related binary conditions.

The p-value was obtained from the chi-squared distribution with $k - 1$ degrees of freedom. In Excel, it can be calculated as:

$$\texttt{CHISQ.DIST.RT}(Q, k - 1)$$

Alongside the inferential tests, descriptive disparity measures were also reported. These include the Top-K inclusion rate, the risk difference, and the risk ratio.

- **Top-K Inclusion Rate (TKR).** For each group g, the Top-K inclusion rate is

$$TKR_g = \frac{\#\text{ of Top-K inclusions for group } g}{\#\text{ of opportunities for group } g}.$$

  Question answered: How frequently does group g appear among the Top-K most visible candidates?

- **Risk Difference (RD).** The absolute representation gap between group g and a reference group r is

$$RD_{g-r} = TKR_g - TKR_r.$$

  Question answered: By how many percentage points does the Top-K inclusion rate of group g differ from that of the reference group r?

  Here, the number of opportunities corresponds to the total number of ranking situations where that group could appear in Top-K.

- **Risk Ratio (RR).** The relative representation gap between group g and a reference group r is

$$RR_{g/r} = \frac{TKR_g}{TKR_r}.$$

  Question answered: How many times more or less frequently does group g appear in the Top-K positions compared to the reference group r?

These measures give a more clear description of the results we see.

3.4.5 Significance Level and Interpretation

For all statistical tests, the significance level was set at $\alpha = 0.05$. Therefore, when the p-value was lower than 0.05, the null hypothesis was rejected and the result was interpreted as statistically significant. When the p-value was above this threshold, the null hypothesis was not rejected.

The results were not interpreted on the basis of statistical significance alone. They were also examined together with descriptive measures such as mean score differences, selection rates, risk differences, and risk ratios.

3.5 Implementation and Reproducibility

The reproducibility of the audit depends on a clearly defined computational pipeline. This section details the software environment, project structure, and execution steps required to replicate the analysis.

3.5.1 Tools, Environment, and Frameworks

The auditing framework was developed using **Python 3.13.1**, which was the current stable version at the time the project began.

- **Sentence-Transformers (Hugging Face):**The core framework used to load and run the all-MiniLM-L6-v2 architecture. This model generates 384-dimensional dense vectors with a standardised token window to ensure semantic consistency.
- **Virtual Environment (`.venv`):** A dedicated isolated environment was implemented to manage all project dependencies. This ensures that library versions for `torch`, `transformers`, and `pandas` remain static across hardware configurations,hereby reducing the risk of version-related inconsistencies.
- **Data Serialization:** The `Pandas` library is utilized for structured metadata management, enabling the transition from raw similarity scores to a structured CSV file for analysis.

3.5.2 Data Accessibility and Project Structure

All data required for the experiment is organised within a hierarchical directory structure to ensure transparency and reproducibility. The project repository is available at `[https://github.com/XhoanaShkajoti/Thesis]`.

- **Raw Templates (`cv_dataset/`):** Contains the base professional profiles used for the synthetic expansion.
- **Job Descriptions(`job_descriptions/`):** Stores the job descriptions.
- **Generated Corpus (`cv_dataset_expanded/`):** A dynamic directory where the 384 candidate profiles are instantiated during the generation phase.
- **Audit Results (`results/`):** Stores the `final_bias_audit_results.csv`, which contains the similarity scores indexed by nationality, gender, seniority level and tonality version.

3.5.3 Reproduction and Execution Protocol

To replicate the experiment a researcher must follow this deterministic three-step execution protocol:

Step 1: Environment Initialization

The researcher must activate the virtual environment (`.venv`) and install the dependencies from the `requirements.txt` file. This establishes the exact software versions required for deterministic transformer inference.

Step 2: Deterministic Corpus Instantiation (`mass_generator.py`)

Execute the `mass_generator.py` script to perform automated string replacement. This script iterates through 4 nationalities, 2 genders, 4 seniority levels, and 3 versions per level to construct a balanced factorial matrix of 384 unique CVs.

Step 3: Algorithmic Audit (`run_audit.py`)

Execute the `run_audit.py` script to complete the audit. The script automates the following actions:

1. Loads the all-MiniLM-L6-v2 model in evaluation mode (`.eval()`) to ensure consistent output across repeated runs.
2. Encodes the 384 generated CVs and the Job Description anchors into vector representations.
3. Calculates the Cosine Similarity between each candidate vector and its corresponding job anchor.
4. Exports the results to the results directory which then can be used for statistical analysis.

Chapter 4

Results

This chapter presents and discusses the empirical results of the fairness audit conducted on the SBERT-based hiring pipeline, with the aim of quantitatively assessing whether gender- and nationality-related disparities emerge in the simulated recruitment process.

4.1 Similarity score analysis

This section focuses on the similarity score analysis and examines whether the SBERT-based hiring pipeline assigns different matching scores by gender and nationality. The results were obtained by comparing paired candidate profiles with the same job descriptions through the matching procedure described in Chapter 3. The analysis was then carried out by observing how the resulting scores vary across the demographic groups considered in the audit.

4.1.1 Aggregate-level score differences

This part presents the similarity score results by considering all hierarchy levels together. This first view makes it possible to identify the general pattern of differences by gender and nationality before examining the results separately across hierarchical levels.

By Gender

The analysis of aggregate-level gender score differences makes it possible to verify whether gender has a systematic influence on the similarity scores assigned by the model. Table 4.1 reports the main descriptive statistics of the paired score differences between male and female profiles.

Statistic	Value
Number of pairs (n)	192
Mean difference (M−F)	0.00475
Std. deviation	0.00280
Minimum difference	-0.00110
Maximum difference	0.01440
% positive differences	97.4%

Table 4.1: Descriptive statistics of aggregate-level gender score differences

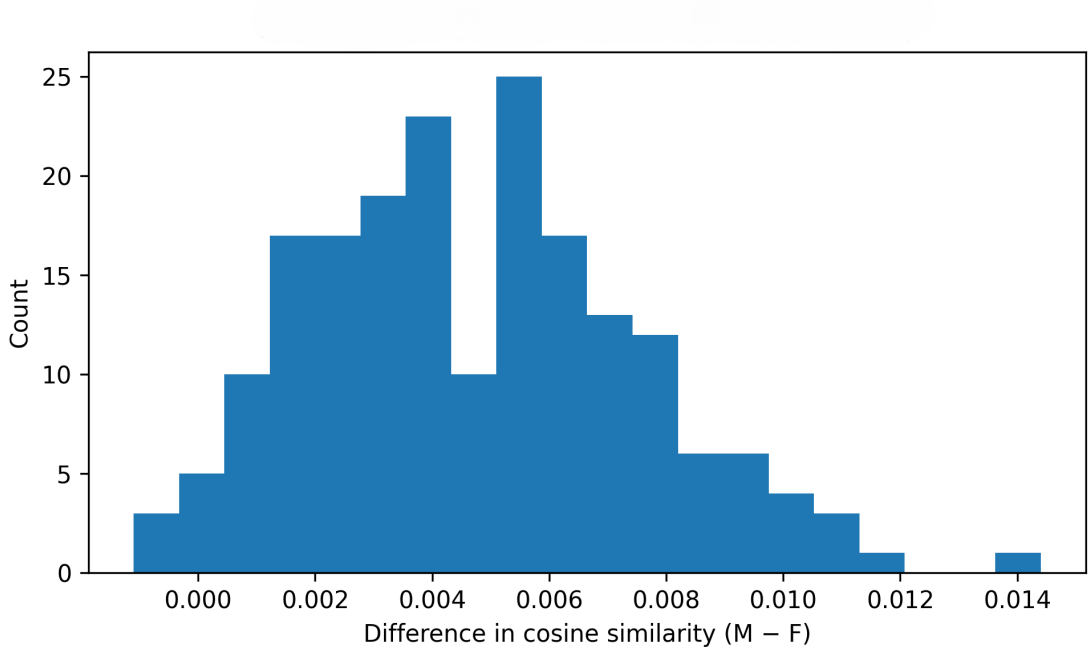

Figure 4.1: Histogram of aggregate-level gender score differences (M−F).

As illustrated in Figure 4.1, the distribution of the paired differences is concentrated mainly above zero. This shows that, in most matched configurations, male profiles receive a slightly higher similarity score than the corresponding female profiles. The same pattern emerges from Table 4.1, where 97.4% of the differences are positive.

The paired t-test confirms this result. The test returns a large positive statistic, $t(191) = 23.53$, with an extremely small p-value ($p = 2.42 \times 10^{-58}$). The null hypothesis of zero mean difference is therefore not supported.Although the statistical evidence is very strong, this does not imply a large score gap in practical terms. The mean difference is equal to 0.00475, which remains very small on a similarity

scale bounded between -1 and 1. Rather, the large t-value is influenced by the strong directional consistency of the paired differences: since they are positive in 97.4% of the matched blocks and remain relatively concentrated, the test detects a systematic effect even if its magnitude is limited.

The results indicate a clear and statistically significant direction of the effect, with male profiles systematically receiving slightly higher similarity scores than female profiles.

By Nationality

The nationality analysis examines whether similarity scores differ across the candidate groups considered in the audit. Table 4.2 shows that the Friedman test strongly rejects the null hypothesis of equal score distributions across nationalities ($\chi^2(3) = 209.83$, $p = 3.17 \times 10^{-45}$), indicating that nationality has an overall effect on the model output.

Test	$\boldsymbol{\chi^2}$	**df**	**p-value**
Friedman	209.83	3	3.17×10^{-45}

Table 4.2: Friedman Repeated-Measures Test.

To understand more clearly where this effect comes from, Table 4.3 reports the pairwise comparisons relative to the Italian baseline, while Figure 4.2 summarises the corresponding mean differences and Figure 4.3 shows the distribution of the paired differences for each comparison.

The pairwise results all point in the same direction. In each comparison, the mean difference relative to Italy is positive, showing that Albanian, British, and Moroccan profiles receive on average higher similarity scores than the corresponding Italian profile under the same contextual conditions. The paired t-tests supports these results, as the null hypothesis of zero mean difference is rejected for all three nationality pairs.

However, the size and strength of the effect are not the same across groups. Figure 4.2 shows that BR$-$IT has the largest mean difference (0.02275), followed by AL$-$IT (0.01606), while MO$-$IT (0.00641) remains much smaller. This suggests that the effect of nationality is most pronounced for the British profiles, intermediate for the Albanian profiles, and more limited for the Moroccan profiles.

Pair	Mean Diff	Std. Dev.	% $d > 0$	t-stat	p-value
AL − IT	0.01606	0.01150	91.7%	13.68	3.42×10^{-24}
BR − IT	0.02275	0.01096	100.0%	20.34	2.22×10^{-36}
MO − IT	0.00641	0.00807	75.0%	7.78	8.86×10^{-12}

Table 4.3: Paired nationality differences relative to the Italian baseline.

Figure 4.3 helps clarify this pattern. The distribution for BR−IT lies clearly on the positive side, showing a very stable advantage over the Italian baseline. This is fully consistent with the fact that BR−IT is positive in 100% of the observed configurations. AL−IT also shows a clear positive pattern, with most of the mass above zero and 91.7% positive differences. MO−IT remains positive on average as well, but its distribution is closer to zero and more dispersed, which fits its lower mean difference and its weaker directional consistency (75.0% positive differences).

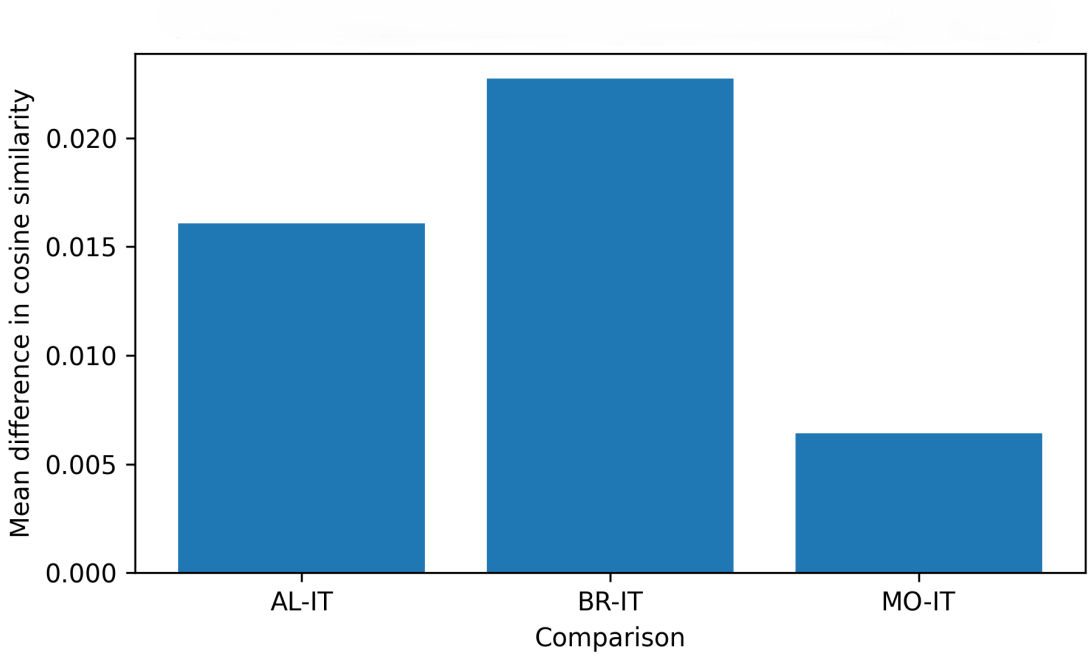

Figure 4.2: Mean within-block nationality differences relative to the Italian baseline (IT).

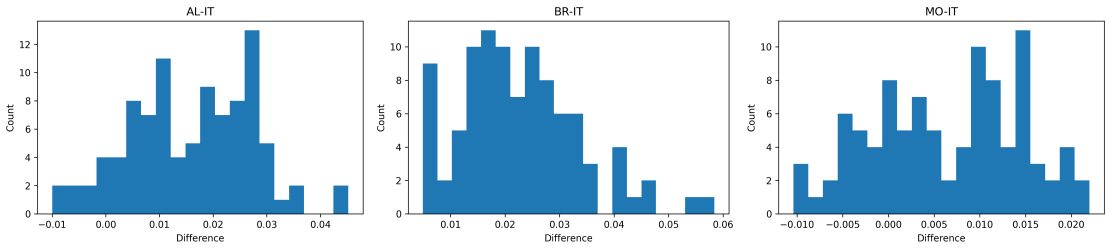

Figure 4.3: Distribution of within-block nationality differences relative to IT: AL−IT, BR−IT, and MO−IT.

As in the gender analysis, the very small p-values should not be read as evidence of large score gaps in absolute terms. The mean differences remain small in cosine-similarity units. What makes the results statistically strong is the fact that the differences tend to fall on the same side of zero across many configurations, especially in the British and Albanian comparisons.

Taken together, these results show that the nationality effect detected by the Friedman test is driven by a clear and directionally consistent advantage relative to the Italian baseline. This pattern is strongest for British profiles, still evident for Albanian profiles, and weaker but still statistically significant for Moroccan profiles.

4.1.2 Score differences by hierarchy level

This part presents the similarity score results by hierarchy level. Looking at the results separately for each level helps show whether the patterns observed in the aggregated analysis are consistent across the organisational structure or more evident in some levels than in others.

By Gender

The analysis by hierarchy level makes it possible to assess whether the gender-related score differences observed in the aggregate results remain stable across the organisational structure. Table 4.4 reports the main descriptive and inferential results for each hierarchy level, while Figure 4.4 compares the corresponding mean differences.

Level	Mean Diff (M−F)	Std. Dev.	% Positive	t-stat	p-value
JR	0.00701	0.00203	100.00%	23.88	6.31×10^{-28}
MID	0.00330	0.00215	93.75%	10.62	4.49×10^{-14}
SR	0.00627	0.00224	100.00%	19.42	4.34×10^{-24}
MGR	0.00243	0.00168	95.83%	10.03	2.88×10^{-13}

Table 4.4: Gender score differences by hierarchy level.

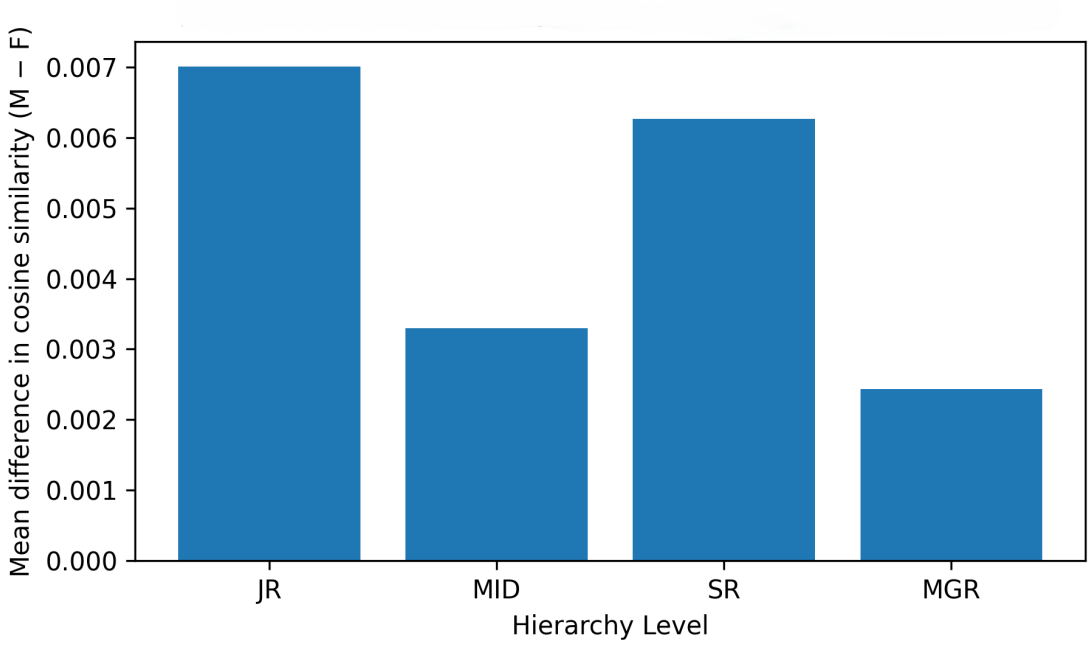

Figure 4.4: Comparison of mean gender score differences by hierarchy level.

As shown in Table 4.4, the pattern observed in the aggregate analysis remains the same across all hierarchy levels. In every group, the mean difference is positive, which means that male profiles systematically receive slightly higher similarity scores than female profiles under the same contextual conditions. This result is also reflected in the share of positive differences, which is very high in all cases and reaches 100% for the senior and junior groups.

The statistical analysis supports this trend. The paired t-tests are statistically significant at every hierarchy level, and the null hypothesis of zero mean difference is therefore rejected for all four levels. As in the aggregate case, this strong statistical evidence should not be interpreted as a large score gap in absolute terms. The mean differences remain small in cosine-similarity units.

That said, the size of the gap is not the same across the hierarchy. Figure 4.4

shows that the largest average difference is observed at the junior level (0.00701), followed by the senior level (0.00627), while smaller values emerge in the mid-level (0.00330) and managerial (0.00243) groups. In particular, the junior gap is more than twice as large as the mid-level and managerial ones. This suggests that, although the direction of the effect is stable across all levels, some of them show a more pronounced gender-related disparity than others.

The results show the null hypothesis is rejected at every hierarchy level, but they also show that the gender-related score shift is not uniform, as some levels display a more marked gap than others.

By Nationality

This part examines whether the nationality pattern changes once the results are separated by hierarchy level. Table 4.5 shows that the Friedman test rejects the null hypothesis of equal nationality score distributions in all four levels. This means that the overall nationality effect remains present throughout the hierarchy.

Level	χ^2	p**-value**
JR	54.56	8.50×10^{-12}
MID	44.60	1.13×10^{-9}
SR	58.20	1.42×10^{-12}
MGR	58.80	1.06×10^{-12}

Table 4.5: Friedman test by hierarchy level.

The pairwise results in Table 4.6 follow the same pattern already observed in the aggregate analysis. In every level, the mean differences relative to Italy are positive, and the paired t-tests reject the null hypothesis of zero mean difference for all three comparisons. British profiles show the clearest dominance, with BR−IT positive in 100% of the cases at every level and with the largest mean gap in each group. Albanian profiles remain in an intermediate position and follow a fairly stable pattern across levels.

By contrast, the differences in mean magnitude across hierarchy levels are not especially marked. The ordering stays the same from one level to another, and Figure 4.5 shows only limited variation in the size of the average gaps. In this sense, separating the results by hierarchy does not change the main picture already seen in the aggregate analysis.

Level	**Pair**	**Mean Diff**	**Std. Dev.**	**% $d > 0$**	***t*-stat**	***p*-value**
JR	AL−IT	0.01746	0.01440	91.7%	5.94	4.65×10^{-6}
JR	BR−IT	0.02680	0.01313	100.0%	10.00	7.66×10^{-10}
JR	MO−IT	0.00590	0.00813	75.0%	3.56	1.68×10^{-3}
MID	AL−IT	0.01357	0.01222	83.3%	5.44	1.58×10^{-5}
MID	BR−IT	0.02299	0.01234	100.0%	9.13	4.17×10^{-9}
MID	MO−IT	0.00630	0.00946	58.3%	3.26	3.44×10^{-3}
SR	AL−IT	0.01609	0.01003	91.7%	7.86	5.79×10^{-8}
SR	BR−IT	0.02181	0.00961	100.0%	11.12	1.00×10^{-10}
SR	MO−IT	0.00526	0.00773	75.0%	3.33	2.90×10^{-3}
MGR	AL−IT	0.01713	0.00891	100.0%	9.41	2.35×10^{-9}
MGR	BR−IT	0.01940	0.00703	100.0%	13.53	1.96×10^{-12}
MGR	MO−IT	0.00818	0.00700	91.7%	5.72	7.88×10^{-6}

Table 4.6: Nationality score differences by hierarchy level relative to the Italian baseline.

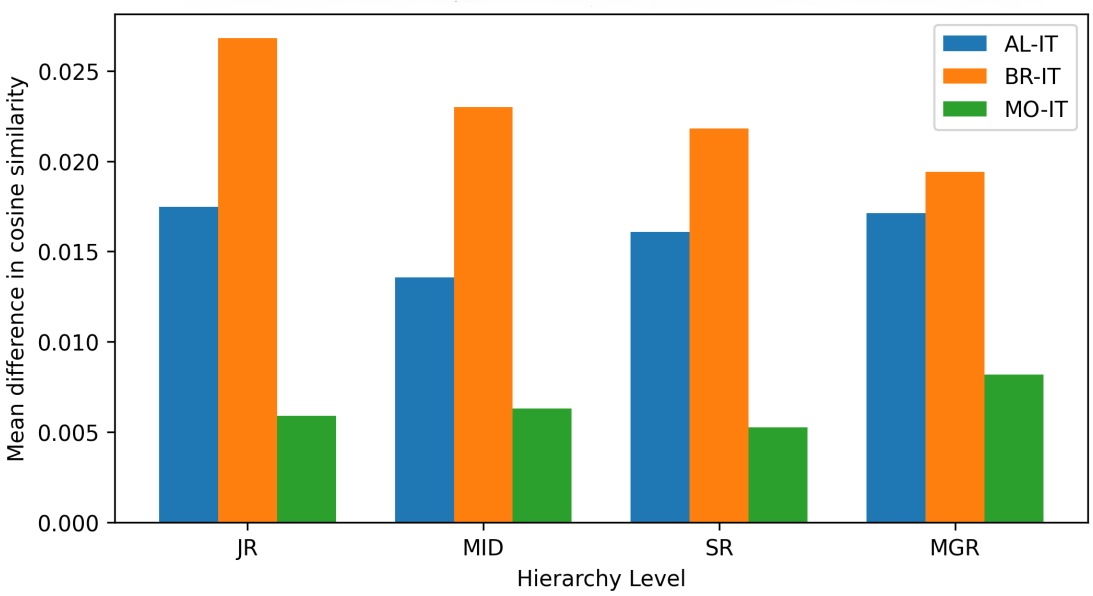

Figure 4.5: Mean nationality differences by hierarchy level relative to the Italian baseline (IT).

The most visible fluctuation appears in the Moroccan comparison. MO−IT remains positive and statistically significant in every level, but the share of positive differences changes more noticeably than in the other cases. In particular, it

falls to 58.3% at the mid level and rises to 91.7% at the managerial level. This suggests a less stable pattern for Moroccan profiles across levels, whereas the British comparison remains consistently strong in all groups.

Taken together, these results suggest that the hierarchy-level breakdown does not alter the general nationality pattern, but it does highlight a greater fluctuation in the Moroccan case.

4.1.3 Robustness to tonality variation

This section examines whether the score-level patterns observed for gender and nationality remain stable across the three tonal versions of the synthetic CVs. The aim is to verify whether the detected disparities are limited to a specific writing style or persist across different formulations of the same profile.

By Gender

Table 4.7 and Figure 4.6 show that the gender-related score difference remains positive across all three tonal versions. The size of the gap varies slightly, with V2 showing the largest average difference, while V1 and V3 remain smaller but still consistent in direction. Overall, this suggests that the gender pattern is not driven by a single writing style, but remains visible across tonal formulations.

Version	**mean(M−F)**	**median(M−F)**	**% positive**
V1	0.0040375	0.00420	92.19%
V2	0.0060313	0.00570	100%
V3	0.0041859	0.00375	100%

Table 4.7: Paired gender gap (M−F) by tonality version.

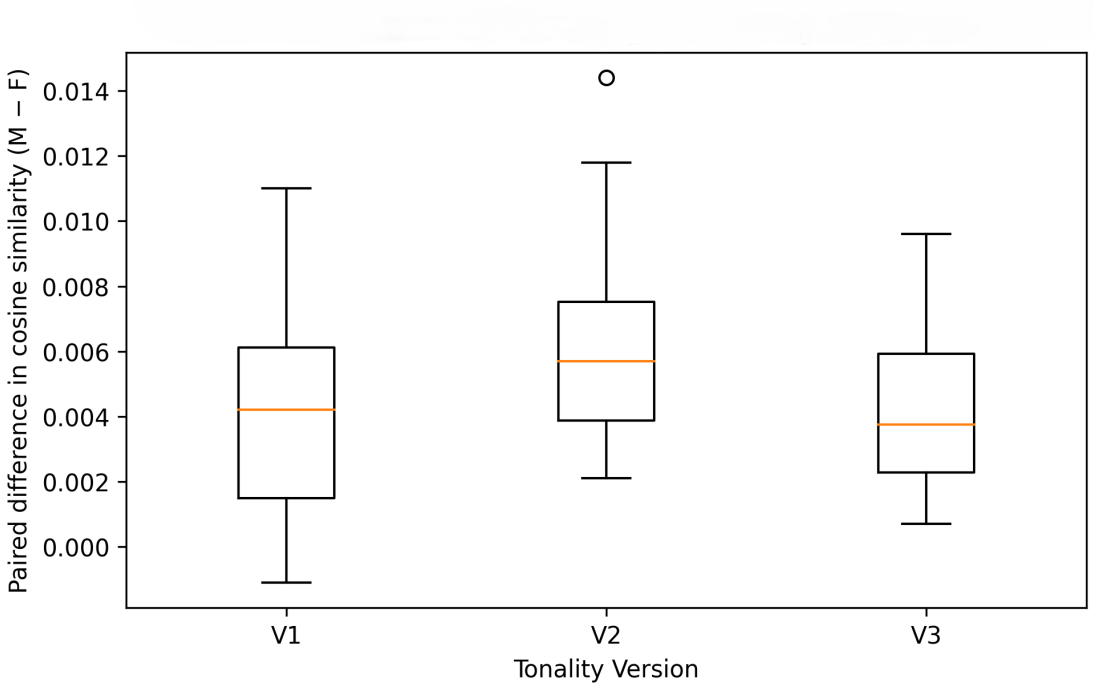

Figure 4.6: Distribution of paired gender gaps (M−F) by tonality version.

By Nationality

Table 4.8 and Figure 4.7 show that the strongest nationality patterns also remain visible across tonal versions. In particular, AL−IT and BR−IT stay positive in all three versions and continue to show clear and statistically significant differences relative to the Italian baseline. MO−IT is weaker and less stable: it remains positive in V1 and V2, but becomes much smaller and no longer statistically significant in V3. A possible explanation is that, since V3 also exhibits lower similarity scores in general than the other versions, the corresponding Moroccan gap is less clearly detectable in this version. The robustness analysis supports the main score-level results, since the direction of the observed gender and nationality patterns remains broadly stable across the different tonality versions.

Version	Pair	Mean Diff	Std. Dev.	% $d > 0$	t-stat	p-value
V1	AL–IT	0.02627	0.00761	100.00%	19.52	5.63×10^{-19}
V1	BR–IT	0.03183	0.01057	100.00%	17.03	2.72×10^{-17}
V1	MO–IT	0.00803	0.00759	81.25%	5.99	1.25×10^{-6}
V2	AL–IT	0.01804	0.00633	100.00%	16.14	1.24×10^{-16}
V2	BR–IT	0.02193	0.00738	100.00%	16.82	3.86×10^{-17}
V2	MO–IT	0.01018	0.00695	90.63%	8.30	2.27×10^{-9}
V3	AL–IT	0.00388	0.00653	75.00%	3.36	2.07×10^{-3}
V3	BR–IT	0.01449	0.00676	100.00%	12.12	2.71×10^{-13}
V3	MO–IT	0.00101	0.00681	53.13%	0.84	4.10×10^{-1}

Table 4.8: Nationality differences vs. IT by tonality version.

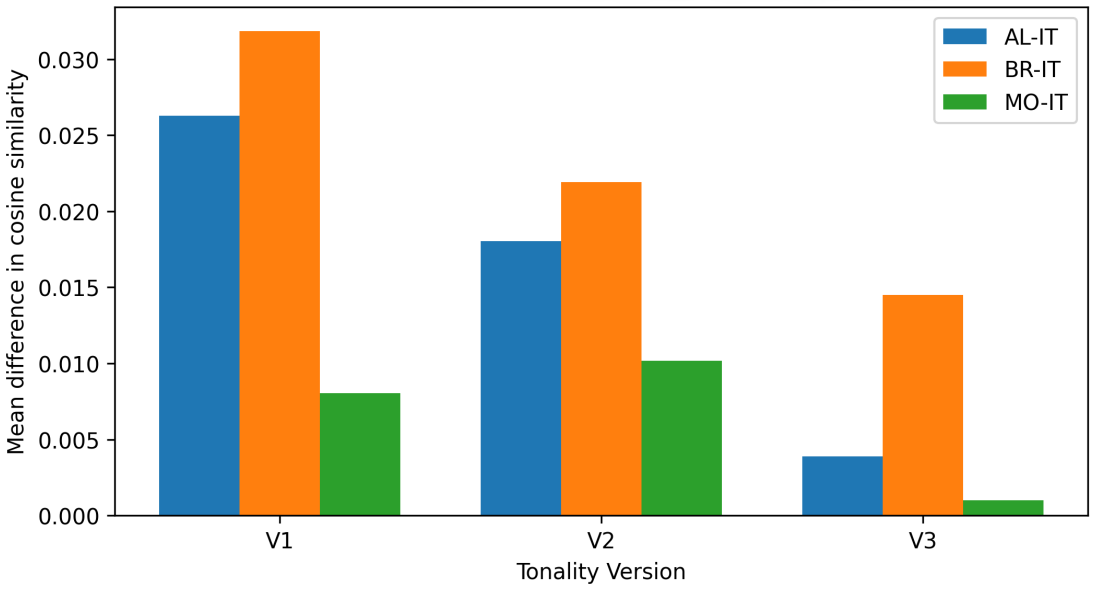

Figure 4.7: Mean nationality difference vs IT by tonality version.

4.2 Threshold-based screening analysis

This section looks at what happens when similarity scores are turned into screening decisions through fixed selection thresholds. The aim is to see whether the small score differences observed in the previous section remain limited at the score level or become more visible once candidates are either selected or excluded. The results are shown for different operating points, so that it is possible to observe how the pattern changes as the screening becomes more or less selective.

4.2.1 Aggregate-level threshold-based disparities

This part presents the threshold-based results by considering all hierarchy levels together.

By Gender

The gender threshold analysis examines whether the score differences observed earlier turn into different selection outcomes once a global cutoff is applied. Table 4.9 reports the main results, while Figure 4.8 shows the corresponding selection rates for male and female candidates across the three operating points.

As shown in Table 4.9 and Figure 4.8, male candidates have a slightly higher selection rate than female candidates at all three thresholds. However, a statistically significant difference emerges only at the 30% cutoff, where the McNemar test rejects the null hypothesis of equal selection probability ($\chi^2 = 7.00$, $p = 0.0156$). At the 10% and 20% thresholds, by contrast, the null hypothesis is not rejected. These results suggest that the gender-related score shift does not translate into a clearly detectable screening imbalance under the more restrictive cutoffs. A likely explanation is that the absolute score difference remains small, so its effect on the final selection outcome becomes more visible only when the threshold is relaxed.

Threshold	SR_M	SR_F	#M	#F	RD	RR	χ^2	p-value
Top 10%	0.1042	0.0938	20	18	0.0104	1.1111	2.00	0.5000
Top 20%	0.2031	0.1979	39	38	0.0052	1.0263	1.00	1.0000
Top 30%	0.3177	0.2812	61	54	0.0365	1.1296	7.00	0.0156

Table 4.9: Gender selection rates under global thresholds.

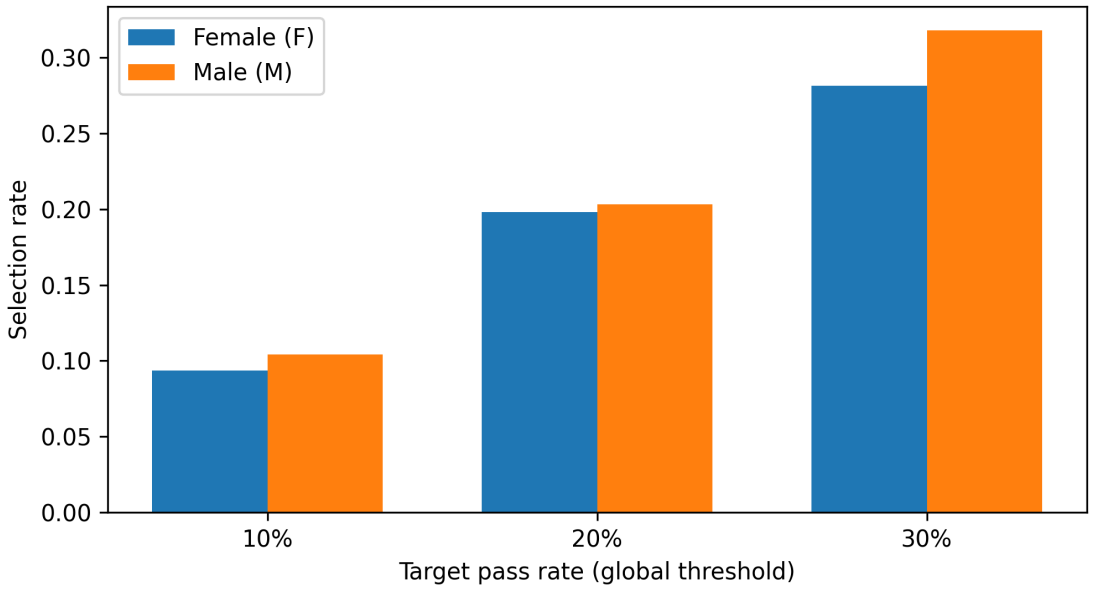

Figure 4.8: Gender selection rate by threshold.

By Nationality

The nationality threshold analysis examines whether the score differences observed earlier also appear once screening decisions are based on fixed cutoffs. Table 4.11 reports the selection rates across the four nationality groups, while Table 4.12 and Figure 4.9 show the disparities relative to the Italian baseline.

As shown in Table 4.10, Cochran's Q is statistically significant at all three operating points. The null hypothesis of equal selection probability across nationalities is therefore rejected at the Top 10%, Top 20%, and Top 30% thresholds. This means that nationality differences remain visible even after the scores are translated into binary screening decisions.

Threshold	#AL	#BR	#IT	#MO	Q-stat	p-value
Top 10%	12	14	6	6	20.40	1.40×10^{-4}
Top 20%	23	24	14	16	24.24	2.22×10^{-5}
Top 30%	29	33	24	29	15.77	1.26×10^{-3}

Table 4.10: Nationality frequencies and Cochran's Q test results by threshold.

Threshold	SR_{AL}	SR_{BR}	SR_{IT}	SR_{MO}
Top 10%	0.1250	0.1458	0.0625	0.0625
Top 20%	0.2396	0.2500	0.1458	0.1667
Top 30%	0.3021	0.3438	0.2500	0.3021

Table 4.11: Nationality selection rates by threshold.

Threshold	RD_{AL}	RR_{AL}	RD_{BR}	RR_{BR}	RD_{MO}	RR_{MO}
Top 10%	0.0625	2.0000	0.0833	2.3333	0.0000	1.0000
Top 20%	0.0938	1.6429	0.1042	1.7143	0.0208	1.1429
Top 30%	0.0521	1.2083	0.0938	1.3750	0.0521	1.2083

Table 4.12: Nationality disparities using Italy as baseline.

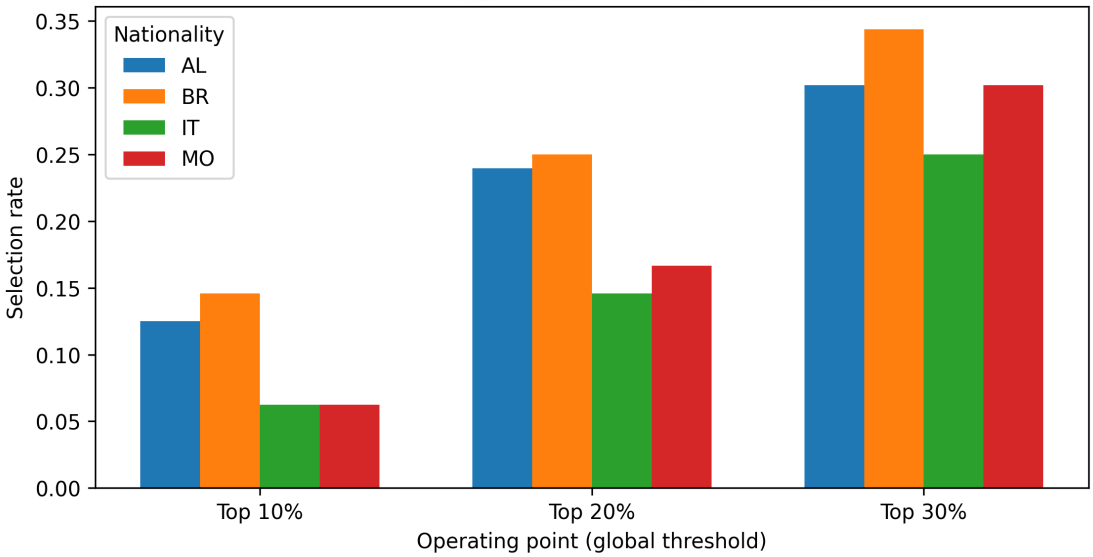

Figure 4.9: Nationality selection rates by threshold.

As shown in Table 4.11 and Figure 4.9, the threshold analysis reflects the same general pattern already seen at score level, with British candidates showing the clearest advantage across all three operating points.

Table 4.12 makes this pattern more explicit. At the Top 10% threshold, British candidates are selected more than twice as often as Italian candidates ($RR_{BR} = 2.3333$), also Albanian candidates are selected about twice as often ($RR_{AL} = 2.0000$).

As the threshold is relaxed, the risk ratios tend to decrease, but the differences in selection rates remain visible across all three operating points.

These results show that nationality does affect the screening scenario. The same ordering seen at score level reappears after thresholding, with British candidates consistently more likely to be selected and Italian candidates remaining the least selected group.

4.2.2 Threshold-based disparities by hierarchy level

By Gender

This part examines whether the threshold-based gender pattern changes once the analysis is separated by hierarchy level. Table 4.13 reports the results under level-specific thresholds, while Figure 4.10 and Figure 4.11 show the corresponding selection rates and risk differences. As shown in Table 4.13, male candidates generally retain slightly higher selection rates than female candidates across most hierarchy-threshold combinations. However, these differences remain small, and none of the McNemar tests rejects the null hypothesis of equal selection probability within level.

Level	Threshold	SR_M	SR_F	#M	#F	RD	RR	χ^2	p-value
JR	Top 10%	0.1250	0.0833	6	4	0.0417	1.5000	2.00	0.5000
JR	Top 20%	0.2292	0.1667	11	8	0.0625	1.3750	3.00	0.2500
JR	Top 30%	0.3333	0.2708	16	13	0.0625	1.2308	3.00	0.2500
MID	Top 10%	0.1250	0.0833	6	4	0.0417	1.5000	2.00	0.5000
MID	Top 20%	0.2083	0.2083	10	10	0.0000	1.0000	0.00	1.0000
MID	Top 30%	0.3333	0.2708	16	13	0.0625	1.2308	3.00	0.2500
SR	Top 10%	0.1250	0.0833	6	4	0.0417	1.5000	2.00	0.5000
SR	Top 20%	0.2083	0.1875	10	9	0.0208	1.1111	1.00	1.0000
SR	Top 30%	0.3125	0.2917	15	14	0.0208	1.0714	1.00	1.0000
MGR	Top 10%	0.1042	0.1042	5	5	0.0000	1.0000	0.00	1.0000
MGR	Top 20%	0.2292	0.1875	11	9	0.0417	1.2222	2.00	0.5000
MGR	Top 30%	0.3333	0.2708	16	13	0.0625	1.2308	3.00	0.2500

Table 4.13: Gender selection rates under level-specific thresholds.

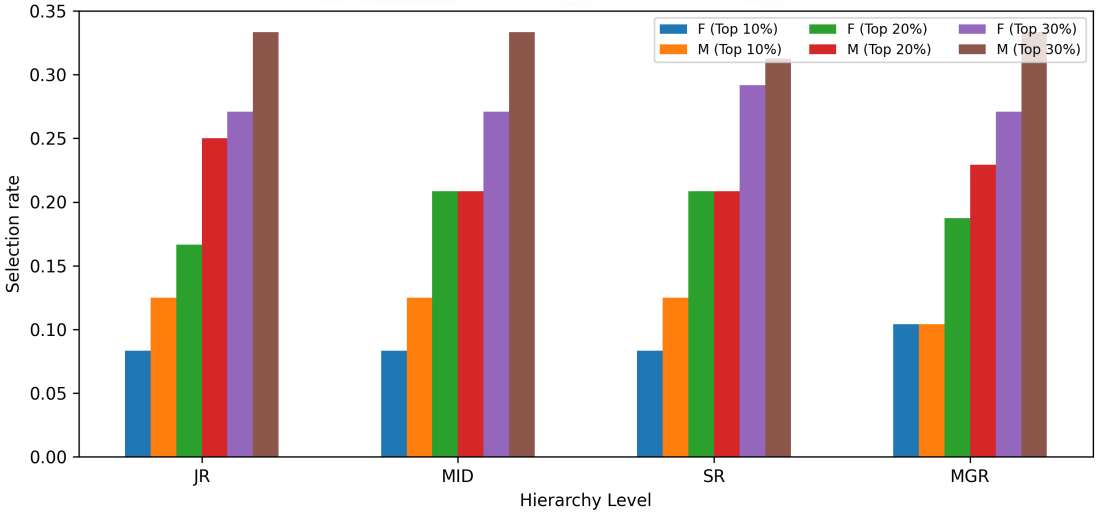

Figure 4.10: Gender selection rates by hierarchy level and threshold.

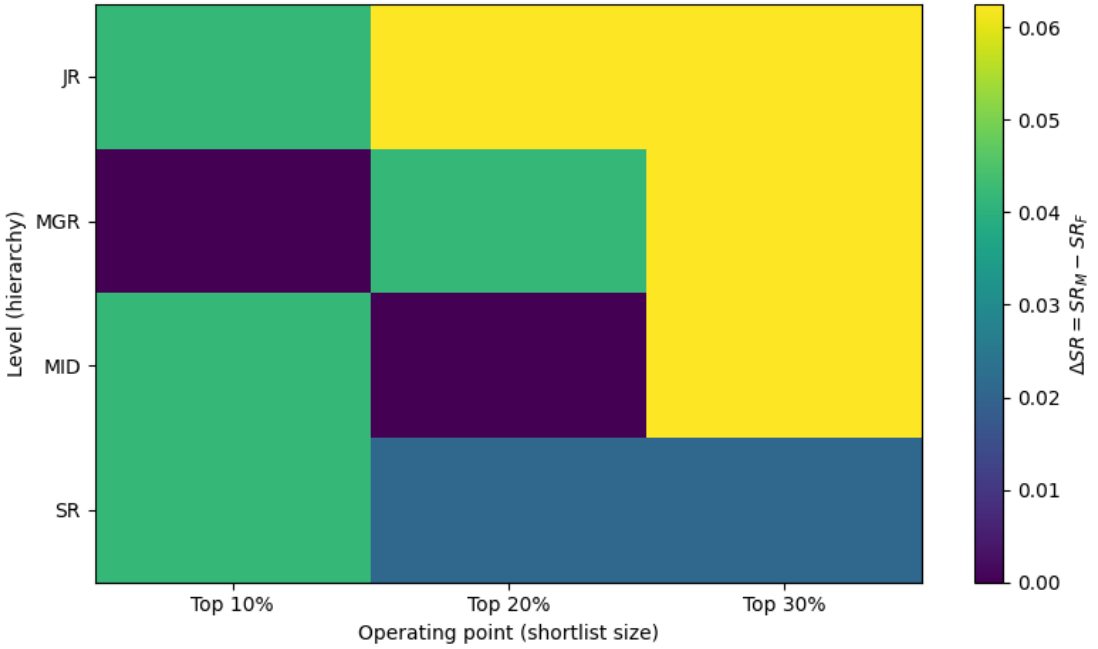

Figure 4.11: Heatmap of gender selection gap under level-specific thresholds.

The heatmap 4.11 suggests that the junior group shows, on average, the highest risk differences across thresholds. However, this pattern is not supported by statistically significant results, so it should be interpreted only as a descriptive

tendency rather than as clear evidence of a stronger hierarchy-specific disparity.

After calibrating thresholds within each hierarchy level, the differences in screening outcomes between genders remain limited and are not statistically supported by the paired analysis.

By Nationality

This part examines whether nationality-based screening disparities remain stable once the analysis is separated by hierarchy level. Table 4.14 4.15 and 4.16 report the level-specific results, while Figure 4.12, 4.13 and Figure 4.14 summarise the corresponding selection rates and gaps relative to the Italian baseline.

The results show that statistical significance is concentrated mainly in the junior group. Cochran's Q is significant at all three thresholds in JR, while among the other levels significance appears only in the managerial group at the Top 30% cutoff ($Q = 13.96$, $p = 0.0030$). This means that the evidence for a nationality effect at the screening stage is strongest and most consistent in the junior case.

Level	Threshold	#AL	#BR	#IT	#MO	Q-stat	p-value
JR	Top 10%	5	4	0	1	11.33	0.0101
JR	Top 20%	6	7	2	4	9.32	0.0254
JR	Top 30%	9	10	4	6	13.00	0.0046
MID	Top 10%	2	4	2	2	6.00	0.1116
MID	Top 20%	6	6	4	4	6.00	0.1116
MID	Top 30%	7	9	6	7	5.18	0.1590
SR	Top 10%	3	3	2	2	3.00	0.3916
SR	Top 20%	6	6	4	4	6.00	0.1116
SR	Top 30%	8	9	6	6	7.36	0.0612
MGR	Top 10%	4	2	2	2	6.00	0.1116
MGR	Top 20%	5	6	4	5	4.00	0.2615
MGR	Top 30%	8	11	4	6	13.96	0.0030

Table 4.14: Observed frequencies and Cochran's Q test results under level-specific thresholds.

Level	Threshold	SR_{AL}	SR_{BR}	SR_{IT}	SR_{MO}
JR	Top 10%	0.2083	0.1667	0.0000	0.0417
JR	Top 20%	0.2500	0.2917	0.0833	0.1667
JR	Top 30%	0.3750	0.4167	0.1667	0.2500
MID	Top 10%	0.0833	0.1667	0.0833	0.0833
MID	Top 20%	0.2500	0.2500	0.1667	0.1667
MID	Top 30%	0.2917	0.3750	0.2500	0.2917
SR	Top 10%	0.1250	0.1250	0.0833	0.0833
SR	Top 20%	0.2500	0.2500	0.1667	0.1667
SR	Top 30%	0.3333	0.3750	0.2500	0.2500
MGR	Top 10%	0.1667	0.0833	0.0833	0.0833
MGR	Top 20%	0.2083	0.2500	0.1667	0.2083
MGR	Top 30%	0.3333	0.4583	0.1667	0.2500

Table 4.15: Nationality selection rates under level-specific thresholds.

Level	Threshold	RD_{AL}	RR_{AL}	RD_{BR}	RR_{BR}	RD_{MO}	RR_{MO}
JR	Top 10%	0.2083	–	0.1667	–	0.0417	–
JR	Top 20%	0.1667	3.0000	0.2083	3.5000	0.0833	2.0000
JR	Top 30%	0.2083	2.2500	0.2500	2.5000	0.0833	1.5000
MID	Top 10%	0.0000	1.0000	0.0833	2.0000	0.0000	1.0000
MID	Top 20%	0.0833	1.5000	0.0833	1.5000	0.0000	1.0000
MID	Top 30%	0.0417	1.1667	0.1250	1.5000	0.0417	1.1667
SR	Top 10%	0.0417	1.5000	0.0417	1.5000	0.0000	1.0000
SR	Top 20%	0.0833	1.5000	0.0833	1.5000	0.0000	1.0000
SR	Top 30%	0.0833	1.3333	0.1250	1.5000	0.0000	1.0000
MGR	Top 10%	0.0833	2.0000	0.0000	1.0000	0.0000	1.0000
MGR	Top 20%	0.0417	1.2500	0.0833	1.5000	0.0417	1.2500
MGR	Top 30%	0.1667	2.0000	0.2917	2.7500	0.0833	1.5000

Table 4.16: Risk differences and risk ratios under level-specific thresholds, with Italy as baseline.

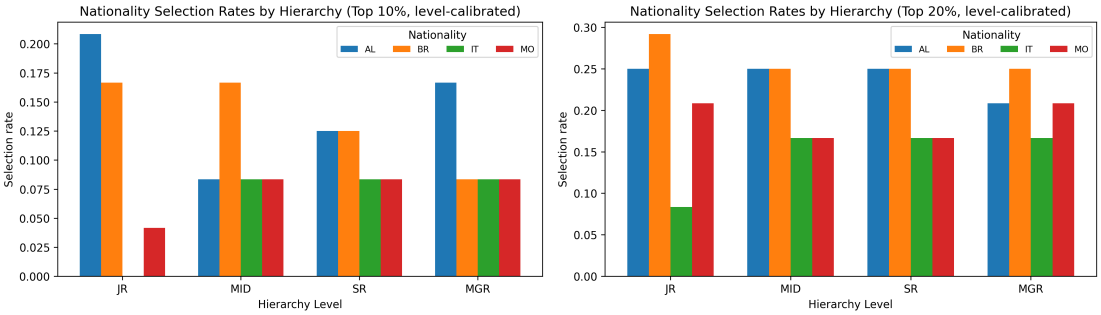

Figure 4.12: Nationality selection rates by hierarchy level and threshold(Top 10% and 20%).

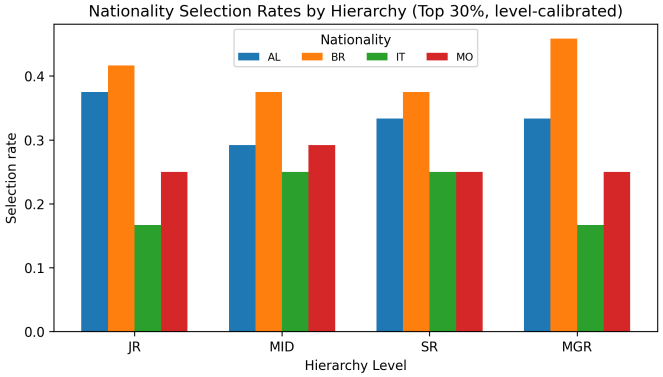

Figure 4.13: Nationality selection rates by hierarchy level and threshold (Top 30%).

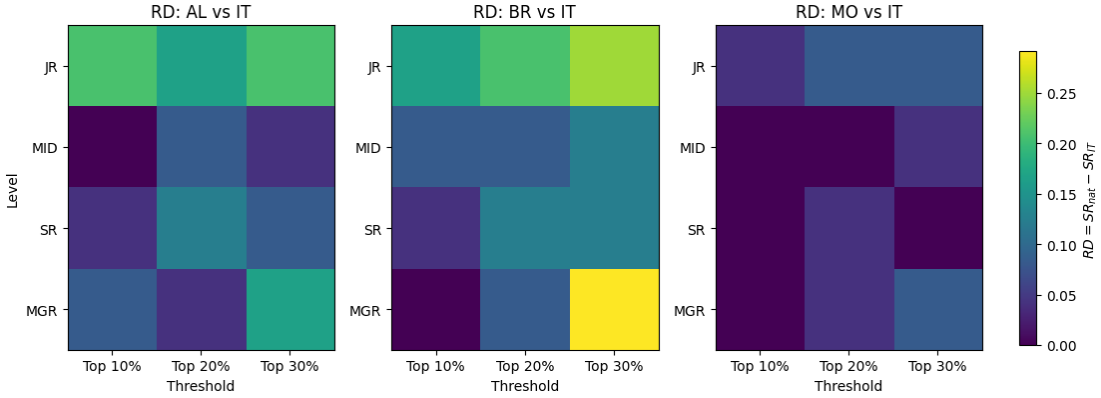

Figure 4.14: Heatmaps of nationality selection gaps relative to Italy .

The general disparity pattern remains similar to the one observed in the aggregate analysis, with British candidates usually showing the clearest advantage over the Italian baseline. The junior group is especially marked at the Top 10% threshold, where the Italian selection rate drops to zero while the other nationality profiles still record positive selection rates. This creates the sharpest separation at the strictest cutoff and helps explain why the overall test is already significant in JR.

The heatmap also points in the same direction. It shows that the junior group experiences, on average, the largest selection gaps relative to Italy across thresholds. British profiles most often display the widest positive differences, while Albanian profiles remain intermediate and Moroccan profiles stay closer to the Italian baseline.

The managerial result at the Top 30% threshold appears to be driven mainly by the particularly strong British advantage at that operating point. Here BR reaches a selection rate of 45.83% against 16.67% for Italy, producing the largest gap in the table. This suggests that the significant managerial result is linked above all to that stronger separation between the British and Italian profiles.

The hierarchy-level threshold analysis suggests that nationality does not affect screening outcomes in a uniform way across the hierarchy. The clearest statistical evidence appears in the junior level, where significance is observed at all three thresholds, while the other levels show either weaker patterns or only isolated significant results.

4.3 Ranking analysis

This section examines whether the score differences observed earlier also shape the final ranking of candidates. In particular, it looks at whether some groups are more likely to appear in the top positions.

4.3.1 Aggregate-level ranking disparities

By Gender

This part examines whether the gender pattern observed earlier also appears in the final ranking. In particular, it looks at whether male and female candidates differ in their probability of reaching the top positions. Table 4.17 and Figure 4.15 show that the null hypothesis of equal Top-K inclusion by gender is rejected for all three values of K, with very small p-values throughout. The clearest difference appears at Top-1, where no female candidate is ever ranked first and all first positions are occupied by male candidates. Although the gap narrows when the shortlist is expanded to Top-3 and Top-5, male candidates still remain much more likely to appear in the top positions. The ranking results show a strong and persistent gender imbalance in favour of male profiles.

K	#M	#F	TKR_M	TKR_F	RD	RR	χ^2	p-value
1	48	0	0.2500	0.0000	0.2500	∞	48.00	7.11×10^{-15}
3	100	44	0.5208	0.2292	0.2917	2.2727	56.00	2.78×10^{-17}
5	145	95	0.7552	0.4948	0.2604	1.5263	46.30	1.65×10^{-13}

Table 4.17: Gender Top-K inclusion across all ranking contexts.

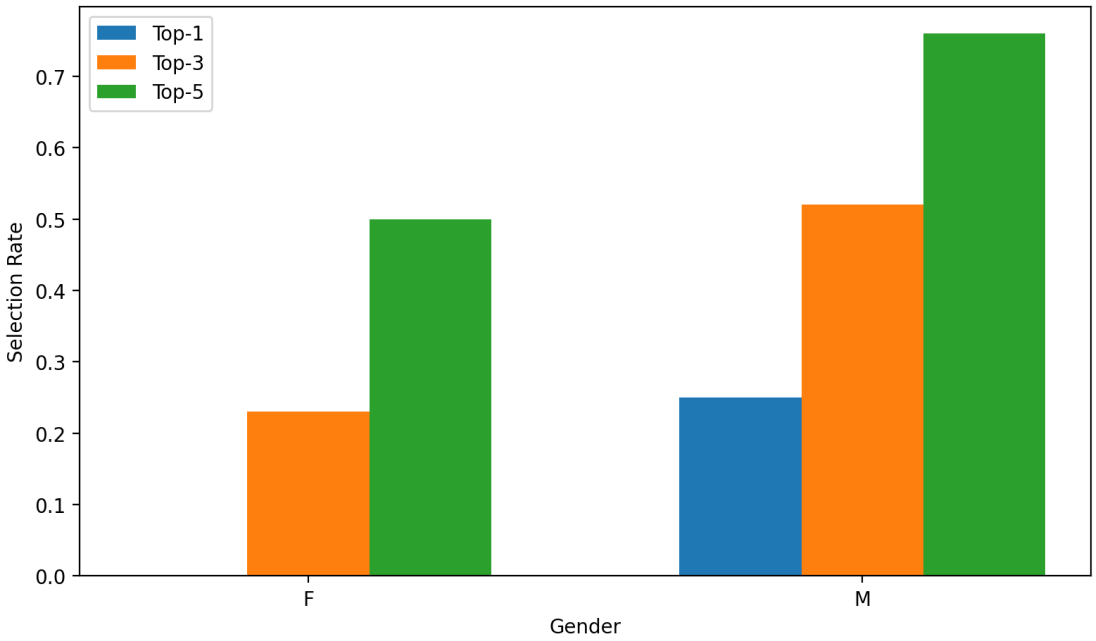

Figure 4.15: Gender Top-1, Top-3, and Top-5 inclusion rates (TKR).

By Nationality

Table 4.18, Table 4.20, and Figure 4.16 show that the null hypothesis of equal Top-K inclusion across nationalities is rejected for all three values of K, with extremely small p-values throughout.

The most striking disparity appears at Top-1. British candidates occupy the first position 42 times, while the second most represented group, Albanian candidates, appears only 5 times. Moroccan candidates appear once, and Italian candidates never appear in first position at all. This shows a very strong concentration of top-rank visibility in favour of British profiles.

K	#AL	#BR	#IT	#MO	Cochran Q	p-value
1	5	42	0	1	101.17	$< 10^{-16}$
3	47	84	4	9	94.72	$< 10^{-16}$
5	82	95	20	43	58.91	1.00×10^{-12}

Table 4.18: Nationality Top-K inclusion counts and Cochran's Q test results.

K	TKR_{AL}	TKR_{BR}	TKR_{IT}	TKR_{MO}
1	0.0521	0.4375	0.0000	0.0104
3	0.4896	0.8750	0.0417	0.0938
5	0.8542	0.9896	0.2083	0.4479

Table 4.19: Nationality Top-K inclusion rates.

K	RD_{AL-IT}	$\mathrm{RR}_{AL/IT}$	RD_{BR-IT}	$\mathrm{RR}_{BR/IT}$	RD_{MO-IT}	$\mathrm{RR}_{MO/IT}$
1	0.0521	∞	0.4375	∞	0.0104	∞
3	0.4479	11.7504	0.8333	21.0000	0.0521	2.2512
5	0.6459	4.1000	0.7813	4.7500	0.2396	2.1500

Table 4.20: Nationality Top-K disparities relative to the IT baseline

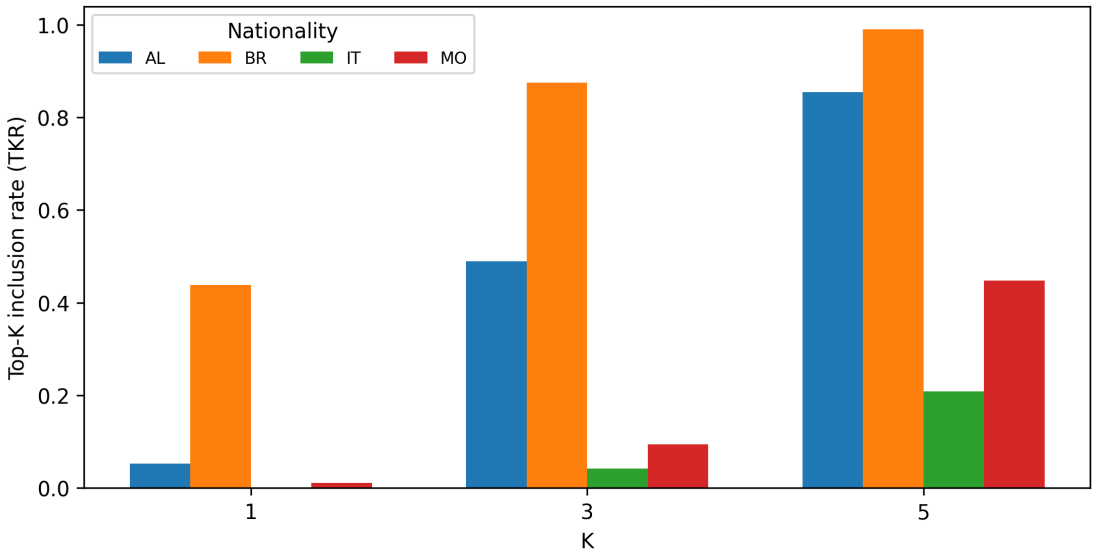

Figure 4.16: Nationality Top-K inclusion rate (TKR) as a function of K.

When the shortlist expands to Top-3 and Top-5, the risk ratios become smaller, so the relative advantage over the Italian baseline is less pronounced than at Top-1. Even so, the selection gaps remain large, and the differences continue to be statistically significant for all three values of K.

The results point to a strong and persistent nationality effect on Top-K visibility.

4.3.2 Ranking disparities by hierarchy level

By Gender

Table 4.21 and Figure 4.17 show that the null hypothesis of equal Top-K inclusion by gender is rejected at every hierarchy level and for all three values of K, with consistently small p-values.

The most striking result appears at Top-1. In every hierarchy level, no female candidate is ever ranked first, so the first position is occupied only by male candidates. For Top-3 and Top-5, female inclusion increases, but male candidates still remain more likely to appear in the top positions across all levels. Figure 4.17 shows that the ranking gap is positive across all hierarchy levels and all Top-K values. On average, the junior group displaysa larger risk differences than the other levels, although the size of the gap varies across ranking cutoffs. Taken together with the results in Table 4.21, this indicates that gender has a statistically significant effect on Top-K inclusion at every hierarchy level, even if its magnitude is not identical across rankings.

Level	K	#M	#F	TKR_M	TKR_F	RD	RR	χ^2	p-value
JR	1	12	0	0.2500	0.0000	0.2500	∞	12.00	0.000488
JR	3	26	10	0.5417	0.2083	0.3333	2.6000	16.00	0.000031
JR	5	37	23	0.7708	0.4792	0.2917	1.6087	14.00	0.000122
MID	1	12	0	0.2500	0.0000	0.2500	∞	12.00	0.000488
MID	3	25	11	0.5208	0.2292	0.2917	2.2727	14.00	0.000122
MID	5	35	25	0.7292	0.5208	0.2083	1.4000	8.33	0.006348
SR	1	12	0	0.2500	0.0000	0.2500	∞	12.00	0.000488
SR	3	24	12	0.5000	0.2500	0.2500	2.0000	12.00	0.000488
SR	5	38	22	0.7917	0.4583	0.3333	1.7273	16.00	0.000031
MGR	1	12	0	0.2500	0.0000	0.2500	∞	12.00	0.000488
MGR	3	25	11	0.5208	0.2292	0.2917	2.2727	14.00	0.000122
MGR	5	35	25	0.7292	0.5208	0.2083	1.4000	8.33	0.006348

Table 4.21: Gender Top-K inclusion by hierarchy level.

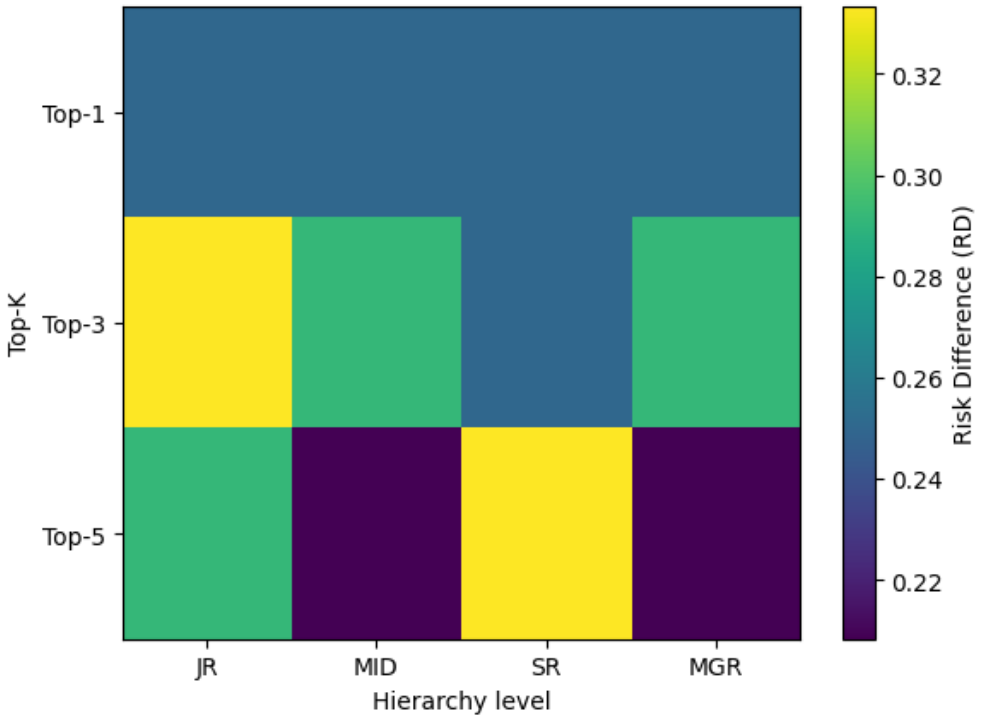

Figure 4.17: Heatmap of Risk Difference (RD) across hierarchy levels and Top-K values.

By Nationality

Table 4.22, 4.23, and 4.24, and Figure 4.18 show that the hierarchy-level results closely mirror the aggregated analysis. The null hypothesis of equal Top-K inclusion across nationalities is rejected in every hierarchy level and for all three values of K, which indicates that the nationality effect on ranking visibility remains statistically significant throughout the hierarchy.

The clearest imbalance appears at Top-1. In all levels, the first position is strongly dominated by British candidates, while Italian candidates never appear in first place. This confirms that the concentration of top-rank visibility seen in the aggregated results is not driven by one specific hierarchy group, but remains present across the full organisational structure. When the shortlist expands to Top-3 and Top-5, the same ordering remains visible in every level. British candidates continue to show the highest inclusion rates, Albanian candidates generally follow, Moroccan candidates remain in an intermediate position, and Italian candidates stay the least represented. The differences in magnitude also remain fairly similar across levels, which suggests a largely uniform pattern rather than strong variation from one hierarchy group to another.

Level	K	#AL	#BR	#IT	#MO	Cochran Q	p-value
JR	1	1	11	0	0	28.67	0.000003
JR	3	11	21	1	3	22.17	0.000060
JR	5	20	24	6	10	12.64	0.005493
MID	1	0	11	0	1	28.67	0.000003
MID	3	8	22	2	4	16.53	0.000882
MID	5	18	24	7	11	10.00	0.018566
SR	1	1	11	0	0	28.67	0.000003
SR	3	14	21	1	0	30.50	0.000001
SR	5	22	23	6	9	13.20	0.004223
MGR	1	3	9	0	0	18.00	0.000440
MGR	3	14	20	0	2	26.00	0.000002
MGR	5	22	24	1	13	28.67	0.000003

Table 4.22: Nationality Top-K inclusion counts and Cochran's Q test results by hierarchy level.

Level	K	TKR_{AL}	TKR_{BR}	TKR_{IT}	TKR_{MO}
JR	1	0.0417	0.4583	0.0000	0.0000
JR	3	0.4583	0.8750	0.0417	0.1250
JR	5	0.8333	1.0000	0.2500	0.4167
MID	1	0.0000	0.4583	0.0000	0.0417
MID	3	0.3333	0.9167	0.0833	0.1667
MID	5	0.7500	1.0000	0.2917	0.4583
SR	1	0.0417	0.4583	0.0000	0.0000
SR	3	0.5833	0.8750	0.0417	0.0000
SR	5	0.9167	0.9583	0.2500	0.3750
MGR	1	0.1250	0.3750	0.0000	0.0000
MGR	3	0.5833	0.8333	0.0000	0.0833
MGR	5	0.9167	1.0000	0.0417	0.5417

Table 4.23: Nationality Top-K inclusion rates by hierarchy level.

Level	K	RD_{AL-IT}	$\mathrm{RR}_{AL/IT}$	RD_{BR-IT}	$\mathrm{RR}_{BR/IT}$	RD_{MO-IT}	$\mathrm{RR}_{MO/IT}$
JR	1	0.0417	∞	0.4583	∞	0.0000	–
JR	3	0.4166	11.0000	0.8333	21.0000	0.0833	3.0000
JR	5	0.5833	3.3332	0.7500	4.0000	0.1667	1.6668
MID	1	0.0000	–	0.4583	∞	0.0417	∞
MID	3	0.2500	4.0000	0.8334	11.0000	0.0834	2.0000
MID	5	0.4583	2.5714	0.7083	3.4286	0.1666	1.5714
SR	1	0.0417	∞	0.4583	∞	0.0000	–
SR	3	0.5416	14.0000	0.8333	21.0000	-0.0417	0.0000
SR	5	0.6667	3.6668	0.7083	3.8332	0.1250	1.5000
MGR	1	0.1250	∞	0.3750	∞	0.0000	–
MGR	3	0.5833	∞	0.8333	∞	0.0833	∞
MGR	5	0.8750	22.0000	0.9583	24.0000	0.5000	13.0000

Table 4.24: Nationality Top-K disparities relative to the IT baseline by hierarchy level. *Risk ratios are reported as ∞ when $TKR_{IT} = 0$ and the numerator is > 0;"–" indicates ratios that are not informative when both numerator and baseline are 0.*

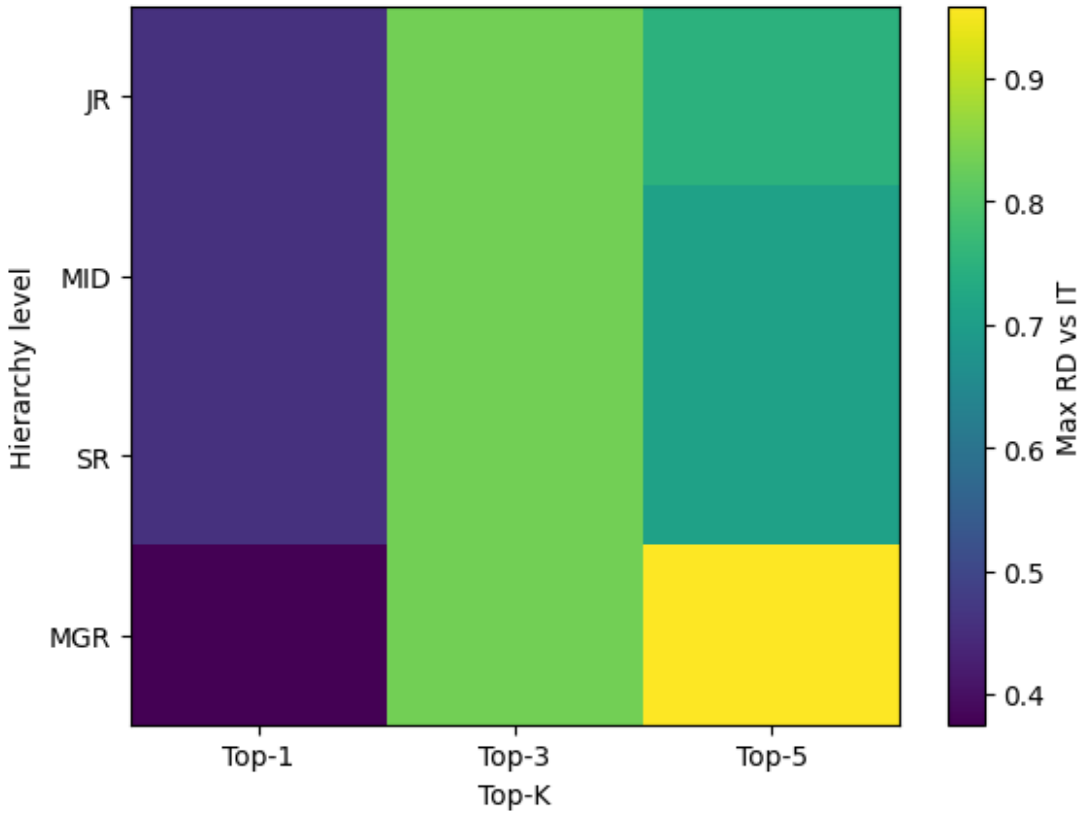

Figure 4.18: Heatmap of risk difference (RD) across hierarchy levels and Top-K values.

Figure 4.18 makes this easier to see. The heatmap shows that the maximum gap relative to the Italian baseline tends to increase from Top-1 to Top-3 and remains high at Top-5. The largest disparity is observed at the managerial level for Top-5, but the overall pattern remains fairly uniform across all levels.

The hierarchy-level Top-K analysis confirms a strong and persistent nationality effect on ranking visibility, with a pattern that remains very close to the aggregated one.

Chapter 5

Conclusion

5.1 Summary of key findings

This thesis examined whether an SBERT-based CV–job description matching system changes its output when demographic cues vary, even though the job-related content of the CV remains the same. To study this, the analysis was carried out on synthetic CVs and looked at the model from three connected angles: score, screening, and ranking. Taken together, the results show that the pipeline is not fully neutral with respect to the attributes considered.

At the score level, the gender effect is small but consistent. Male profiles tend to receive slightly higher similarity scores than their matched female counterparts, and this pattern appears across all hierarchy levels. The size of the gap is not large, but it remains stable throughout the analysis. For nationality, the result is more marked. Compared with the Italian baseline, British profiles show the strongest advantage, Albanian profiles generally follow, and Moroccan profiles remain closer to the baseline, although still not fully neutral. This ordering appears repeatedly in the score analysis, which suggests that the nationality effect is not occasional.

At the screening level, the differences become more concrete, especially for nationality. For gender, the threshold-based results are less regular and less pronounced than the ranking results. For nationality, instead, the disparity is much clearer. Under the strictest global threshold, British profiles can be selected up to 2.33 times as often as Italian profiles, while Albanian profiles can be selected up to 2 times as often. This shows that even moderate differences in score can become more important once the model is used to decide who passes an initial screening stage.

The ranking analysis demonstrates the effects of these disparities even more clearly. For gender, the imbalance is clear in the most selective positions. In the Top-K results, male profiles are more frequently placed in the highest ranks, and at

the most restrictive level, Top-1, female representation drops to zero. This means that, even when the average score difference is small, the final ordering can still produce a much less balanced distribution of visibility. For nationality, the same logic appears even more strongly. British profiles are the most represented in the top positions, while Italian profiles are consistently the least represented. The ranking results therefore confirm that the gap is not only numerical, but also affects who is actually seen first by the system.

The hierarchy analysis does not support a single common trend. However, comparing the hierarchy levels with each other still shows an important pattern. Junior profiles are usually the ones where both gender and nationality produce the strongest effects, especially when the model is used for screening and ranking. In the other levels, the disparities are still present, but they are generally less strong or less consistent. This is relevant in practical terms, because junior roles often attract larger numbers of applicants, so distortions at that stage could affect a wider group of candidates.

The main finding of the study is that demographic differences do not emerge in the same way at every stage of the process. Looking only at similarity scores would show only part of the picture. Once the model is used to shortlist and rank candidates, the effect becomes much clearer. In this sense, the thesis shows that the practical impact of bias depends not only on whether a score difference exists, but also on how that difference is translated into visibility and selection.

5.2 Critical interpretation

The results of this study suggest that the audited matching system is not reacting only to professional content. Even when qualifications, experience, and role fit are kept constant, changing demographic cues still affects how candidates are positioned by the model. This point is important because it shows that semantic matching is not neutral by definition. A system may appear technically sophisticated and still produce outputs that are not fully independent from attributes that should not influence the evaluation.

A second important point concerns the way these differences become more serious once the model is used in practice. At score level, some gaps may seem limited, especially in the gender analysis. However, recruitment systems are not used only to assign similarity values. They are used to sort applicants, reduce visibility, and decide who moves forward. In this study, the strongest concern does not come only from the existence of score differences, but from the fact that these differences become more meaningful when translated into ranking and screening. In other words, the issue is not simply that some groups receive slightly different scores, but that these differences can shape access to shortlist positions and early

selection opportunities.

This is particularly relevant in the nationality analysis. The results show a stable ordering across the tested conditions, with the Italian baseline often performing worse than the other profiles. This is a useful finding because it shows that the reference category chosen in the experimental design is not automatically the advantaged one. The model produces its own ordering, and that ordering has to be measured rather than assumed. More broadly, this suggests that bias in these systems may not always follow intuitive expectations. For this reason, fairness cannot be judged on the basis of assumptions about which groups are likely to benefit or suffer. It has to be tested directly.

The hierarchy analysis also helps to clarify the practical meaning of the findings.The results do not indicate a clear monotonic pattern across hierarchy levels, as disparity does not change consistently in either direction. At the same time, the comparison across hierarchy levels is still informative. Junior roles are usually the ones where both gender and nationality show the strongest effects. This matters because junior positions are often the stage at which applicant pools are larger and initial filtering is more severe. In a real hiring context, even moderate distortions at that point could affect a substantial number of candidates.

From a theoretical point of view, this thesis contributes to the study of fairness in AI-assisted hiring by showing that score-level analysis alone is not enough. Looking only at average similarity values would have captured only part of the problem. By combining score analysis with screening, ranking, and hierarchy-based comparisons, the study shows more clearly where disparities become operationally relevant. In this sense, the thesis proposes a more complete way to evaluate fairness in embedding-based recruitment systems, especially when the model is used as a black-box matching tool.

Findings suggest that although semantic embeddings are effective at capturing textual similarity in ways that go beyond explicit rule-based approaches, they should not be assumed to be neutral. Their use in recruitment requires careful auditing, especially in the early stages of candidate selection, where ranking and visibility already shape opportunities. The main implication, then, is not that these tools must be rejected, but that they should not be adopted without systematic testing, cautious interpretation, and safeguards that take into account how small score differences can become unequal outcomes in practice.

5.3 Limitations of the study and future perspectives

The findings of this thesis should be read in light of some important limitations.

First, the analysis is based on synthetic CVs. This was a deliberate choice,

because it made it possible to control the content of the profiles and vary only the demographic cues under study. In this sense, the synthetic design was useful for isolating the effect of gender and nationality more clearly. At the same time, synthetic CVs cannot capture the full complexity of real applications. Real resumes are more varied in style, structure, wording, and background, and this means that the results should be interpreted as evidence from a controlled audit rather than as a direct picture of real labour-market outcomes.

Second, the study focuses on one specific embedding model, all-MiniLM-L6-v2, used in a single matching setup. This provides a clear and transparent case for analysis, but it also limits the scope of the conclusions. The results cannot be automatically extended to other sentence-embedding models, multilingual systems, fine-tuned architectures, or commercial recruitment tools, which may behave differently.

Third, the analysis considers only a limited set of demographic attributes. The study focuses on gender and nationality, but fairness issues in hiring may also involve other characteristics and, above all, combinations of characteristics. In real contexts, different forms of disadvantage often overlap. Since intersectional effects were not analysed systematically here, the thesis captures only part of the broader fairness problem.

A further limitation concerns the stage of the hiring process that was examined. The thesis focuses on the matching, ranking, and screening stage of recruitment. This is an important part of AI-assisted hiring, but it is not the whole process. Real hiring decisions are also shaped by recruiter judgement, interviews, organisational practices, and institutional rules. For this reason, the results should not be read as a complete account of discrimination in recruitment, but rather as evidence about one specific and influential stage of the pipeline.

These limitations also point to possible directions for future research. A first extension would be to compare the audited model with other embedding systems, including multilingual and domain-adapted models, in order to understand whether the same patterns remain stable across architectures. A second step would be to test the framework on more realistic and heterogeneous CV collections, closer to real hiring materials. Future work could also include intersectional analyses, for example by studying how gender and nationality interact rather than treating them separately. Finally, it would be useful to examine mitigation strategies and human-in-the-loop settings, in order to understand whether these disparities are reduced, amplified, or corrected once the model is placed in a more realistic decision environment.

5.4 Final conclusion

This thesis shows that an SBERT-based CV–job description matching system should not be treated as inherently neutral. Even when candidates have comparable qualifications and experience, the model does not always treat demographic cues in an invariant way. The analysis shows that these differences are not equally visible at every stage: some appear already at score level, while others become more evident when the system is used to rank candidates or to decide who passes an initial screening.

The results also show that the practical impact of bias depends on how the model is used. A small difference in score may seem limited on its own, but it can become more meaningful once it affects visibility, shortlist positions, and access to the next stage of selection. This is especially important in a hiring context, where early filtering already shapes who is seen and who is excluded.

For this reason, the main conclusion of the thesis is that fairness in AI-assisted recruitment should not be evaluated only in terms of technical performance. A model may work well as a matching tool and still produce unequal outcomes across comparable candidates. What matters is not only whether the system is efficient, but also whether that efficiency comes at the cost of unequal treatment.

The broader implication is that systems of this kind should not be adopted as neutral decision-support tools by default. They need to be tested carefully, interpreted with caution, and evaluated at the points where scores become real decisions. In this sense, the thesis argues that fairness auditing should be treated as a necessary part of the design and use of AI in recruitment, not as an optional check added at the end.

Bibliography

[1] Manish Raghavan et al. «Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices». In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (2020) (cit. on pp. 1, 3–5, 11, 15, 24).

[2] Tolga Bolukbasi et al. «Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings». In: *Advances in Neural Information Processing Systems* (2016) (cit. on pp. 1, 4, 12, 23, 30).

[3] Aylin Caliskan et al. «Semantics derived automatically from language corpora contain human-like biases». In: *Science* (2017) (cit. on pp. 1, 4, 11–13, 23, 30, 33).

[4] Nikhil Garg et al. «Word embeddings quantify 100 years of gender and ethnic stereotypes». In: *Proceedings of the National Academy of Sciences* (2018) (cit. on p. 1).

[5] European Union Agency for Fundamental Rights. *Charter of Fundamental Rights of the European Union: Article 21 – Non-discrimination*. `https://fra.europa.eu/en/eu-charter/article/21-non-discrimination` (cit. on pp. 1, 17).

[6] European Parliament and Council. *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. 2024. URL: `https://eur-lex.europa.eu/eli/reg/2024/1689/oj` (cit. on pp. 1, 16, 24).

[7] *EU Artificial Intelligence Act: Article 6 – Classification rules for high-risk AI systems*. `https://artificialintelligenceact.eu/article/6/` (cit. on p. 1).

[8] *EU Artificial Intelligence Act: Annex III – High-risk AI systems referred to in Article 6(2)*. `https://artificialintelligenceact.eu/annex/3/` (cit. on p. 1).

[9] Solon Barocas and Andrew Selbst. «Big Data's Disparate Impact». In: *California Law Review* (2016) (cit. on pp. 2, 10, 24, 25, 30).

[10] Chandler May et al. «On Measuring Social Biases in Sentence Encoders». In: *NAACL*. 2019 (cit. on pp. 2, 4, 11–13, 15, 18, 20, 22–24, 29, 30).

[11] Jeffrey Dastin. *Insight: Amazon scraps secret AI recruiting tool that showed bias against women*. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/. 2018 (cit. on p. 2).

[12] Workday. *What is an Applicant Tracking System?* https://www.workday.com/en-us/topics/hr/applicant-tracking-system.html (cit. on p. 3).

[13] Patrick van Esch et al. «Recruiting and selecting talent with artificial intelligence: A systematic review». In: *Journal of Business Research* (2021) (cit. on p. 3).

[14] Miranda Bogen and Aaron Rieke. *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*. https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20--%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms%2C%20Equity%20and%20Bias.pdf. 2018 (cit. on p. 3).

[15] Nils Reimers and Iryna Gurevych. «Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks». In: *EMNLP* (2019) (cit. on pp. 3, 4, 8, 13, 14, 20–24).

[16] Greenhouse. *Talent Matching – Data Processing FAQ*. https://support.greenhouse.io/hc/en-us/articles/41131616864283-Talent-Matching-Data-Processing-FAQ (cit. on pp. 3, 8).

[17] Textkernel. *Semantic Search & Advanced Matching for HR Software*. https://www.textkernel.com/learn-support/blog/semantic-search-advanced-matching/ (cit. on pp. 3, 8).

[18] Dor Lavi, Volodymyr Medentsiy, and David Graus. *conSultantBERT: Fine-tuned Siamese Sentence-BERT for Matching Jobs and Job Seekers*. https://arxiv.org/abs/2109.06501. 2021 (cit. on pp. 4, 8, 14, 15).

[19] Mohammed-Hassan Ajjam and Hamed Al-Raweshidy. «AI-driven semantic similarity-based job matching framework for recruitment systems». In: *Information Sciences* 724 (2025). Available online 2025; published in volume 2026, pp. 1–17 (cit. on pp. 4, 8, 18).

[20] S. Deshmukh and A. Raut. «Enhanced Resume Screening for Smart Hiring Using Sentence-BERT». In: *International Journal of Advanced Computer Science and Applications* (2024) (cit. on pp. 4, 7–9, 14, 15).

[21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. «A Survey on Bias and Fairness in Machine Learning». In: *ACM Computing Surveys* 54.6 (2021). DOI: 10.1145/3457607 (cit. on pp. 4, 10).

[22] Alessandro Fabris, Stefano Messlas, Gianmaria Silvello, and Gian Antonio Susto. «Fairness and Bias in Algorithmic Hiring: a Multidisciplinary Survey». In: *arXiv preprint arXiv:2309.13933* (2023) (cit. on pp. 4, 5, 18).

[23] Shomir Wilson and Aylin Caliskan. «Large Language Models for Resume Retrieval: An Empirical Study of Bias». In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (2024) (cit. on pp. 4, 9, 11, 13, 15, 33).

[24] Kyra Wilson and Aylin Caliskan. «Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval». In: *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24)*. 2024. DOI: 10.1145/3630106.3658932 (cit. on pp. 4, 5).

[25] L. An et al. «Measuring Gender and Racial Biases in Large Language Models: Intersectional Evidence from Automated Resume Evaluation». In: *PNAS Nexus* 4.3 (2025). URL: https://academic.oup.com/pnasnexus/article/4/3/pgaf089/8071848 (cit. on p. 4).

[26] European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation)*. Official Journal of the European Union. 2016 (cit. on pp. 4, 16).

[27] Inioluwa Deborah Raji, Andrew Smart, Rebecca White, et al. «Closing the AI accountability gap: Defining an end-to-end auditing framework». In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 2020, pp. 33–44. DOI: 10.1145/3351095.3372873 (cit. on pp. 4, 5, 11, 16).

[28] Steven M. Bellovin, Preetam K. Dutta, and Nathan Reitinger. «Privacy and Synthetic Data». In: *Stanford Technology Law Review* 22 (2019), pp. 1–52 (cit. on pp. 4, 5, 16).

[29] Shreeharsha Bharadhwaj. «Fake It Till You Make It: Synthetic Data and Algorithmic Bias». In: *International Journal of Communication* (2021) (cit. on pp. 5, 16, 24).

[30] Indeed Hiring Lab. *Mehr Bewerbungen bei weniger Stellen: Einseitige Dynamik am Stellenmarkt zum Jahresauftakt*. https://www.hiringlab.org/de/blog/2026/02/16/mehr-bewerbungen-bei-weniger-stellen-einseitige-dynamik-am-stellenmarkt-zum-jahresauftakt/. 2026 (cit. on p. 7).

[31] Indeed Hiring Lab. *Le marché de l'emploi en 2026 : en pleine mutation face aux nouveaux équilibres économiques*. https://www.hiringlab.org/fr/blog/2025/12/10/le-marche-de-lemploi-en-2026-en-pleine-mutation-face-aux-nouveaux-equilibres-economiques/. 2025 (cit. on p. 7).

[32] LinkedIn Economic Graph. *Labor Market Report: Building a Future of Work That Works*. Tech. rep. LinkedIn, 2026. URL: https://economicgraph.linkedin.com/content/dam/me/economicgraph/en-us/PDF/linkedIn-labor-market-report-building-a-future-of-work-that-works-jan-2026.pdf (cit. on p. 7).

[33] Oracle. *Taleo Enterprise Edition: Candidate Advanced Search with Keywords*. https://docs.oracle.com/en/cloud/saas/taleo-enterprise/otrcg/c-advancedsearchkeywords.html (cit. on p. 8).

[34] SAP. *Quick Facts About Candidate Search (SAP SuccessFactors Recruiting)*. https://help.sap.com/docs/successfactors-recruiting/setting-up-and-maintaining-sap-successfactors-recruiting/quick-facts-about-candidate-search (cit. on p. 8).

[35] iCIMS. *Using Simplified Candidate Search (Boolean logic support)*. https://community.icims.com/articles/Knowledge/Using-Simplified-Candidate-Search (cit. on p. 8).

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/ (cit. on p. 8).

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. «Attention is all you need». In: *Advances in Neural Information Processing Systems*. 2017, pp. 5998–6008 (cit. on pp. 8, 33).

[38] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://arxiv.org/abs/1907.11692. 2019 (cit. on p. 8).

[39] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. https://arxiv.org/abs/1910.01108. 2019 (cit. on p. 8).

[40] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. «MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers». In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 5776–5788 (cit. on pp. 8, 33).

[41] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. *MPNet: Masked and Permuted Pre-training for Language Understanding*. https://arxiv.org/abs/2004.09297. 2020 (cit. on p. 8).

[42] Eightfold AI. *AI-powered talent matching: The tech behind smarter and fairer hiring*. https://eightfold.ai/engineering-blog/ai-powered-talent-matching-the-tech-behind-smarter-and-fairer-hiring/. 2025 (cit. on p. 8).

[43] Marianne Bertrand and Sendhil Mullainathan. «Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination». In: *American Economic Review* (2004) (cit. on pp. 9, 10, 24, 37).

[44] Batya Friedman and Helen Nissenbaum. «Bias in Computer Systems». In: *ACM Transactions on Information Systems* 14.3 (1996), pp. 330–347. DOI: 10.1145/230538.230561 (cit. on p. 10).

[45] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, et al. «Bias in data-driven artificial intelligence systems—An introductory survey». In: *WIREs Data Mining and Knowledge Discovery* 10.3 (2020), e1356. DOI: 10.1002/widm.1356 (cit. on p. 10).

[46] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. «Building Classifiers with Independency Constraints». In: *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops (ICDMW '09)*. IEEE Computer Society, 2009, pp. 13–18. DOI: 10.1109/ICDMW.2009.83 (cit. on p. 10).

[47] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. «Data mining for discrimination discovery». In: *ACM Transactions on Knowledge Discovery from Data* 4.2 (2010). DOI: 10.1145/1754428.1754432 (cit. on p. 10).

[48] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. «Fairness through Awareness». In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. Association for Computing Machinery, 2012, pp. 214–226. DOI: 10.1145/2090236.2090255 (cit. on p. 10).

[49] European Union. *Directive 2000/78/EC establishing a general framework for equal treatment in employment and occupation*. https://eur-lex.europa.eu/eli/dir/2000/78/oj/eng. 2000 (cit. on p. 17).

[50] European Union. *Directive 2000/43/EC implementing the principle of equal treatment between persons irrespective of racial or ethnic origin*. https://eur-lex.europa.eu/eli/dir/2000/43/oj/eng. 2000 (cit. on p. 17).

[51] European Union. *Directive 2006/54/EC on equal opportunities and equal treatment of men and women in employment and occupation (recast)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32006L0054. 2006 (cit. on p. 17).

[52] L. Beattie, I. Corpus, L. H. Lin, and P. Ravichandran. «Evaluation Framework for Understanding Sensitive Attribute Association Bias in Latent Factor Recommendation Algorithms». In: *arXiv preprint arXiv:2310.20061* (2023). URL: https://doi.org/10.48550/arXiv.2310.20061 (cit. on p. 18).

[53] Sentence Transformers. *Pretrained Models*. https://www.sbert.net/docs/sentence_transformer/pretrained_models.html. 2026 (cit. on p. 19).

[54] Debora Nozza, Federico Bianchi, and Dirk Hovy. «What the [MASK]? Making Sense of Language-Specific BERT Models». In: *arXiv preprint arXiv:2003.02912* (2020). URL: https://arxiv.org/abs/2003.02912 (cit. on p. 20).

[55] International Labour Organization. *International Standard Classification of Occupations (ISCO-08)*. 2008. URL: https://www.ilo.org/public/english/bureau/stat/isco/isco08/ (cit. on pp. 20, 28–30).

[56] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. «Counterfactual Fairness». In: *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. 2017 (cit. on pp. 24, 37).

[57] Stephen Casper et al. «Black-Box Access is Insufficient for Rigorous AI Audits». In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. Association for Computing Machinery, 2024, pp. 2254–2272. DOI: 10.1145/3630106.3659037. URL: https://doi.org/10.1145/3630106.3659037 (cit. on p. 24).

[58] Shiqi Fang, Zexun Chen, and Jake Ansell. «Peer-induced Fairness: A Causal Approach for Algorithmic Fairness Auditing». In: *arXiv preprint arXiv:2408.02558* (2024) (cit. on p. 24).

[59] S. Mitros et al. *Discrimination by Proxy in AI Blindspot: A Discovery Process for Preventing, Detecting, and Mitigating Bias*. 2024. URL: https://aiblindspot.media.mit.edu/ (cit. on p. 25).

[60] European Union Agency for Fundamental Rights. *Being Muslim in the EU: Experiences of Muslims*. Tech. rep. 2024. URL: https://fra.europa.eu/en/publication/2024/being-muslim-eu (cit. on p. 27).

[61] Eurostat. *Employment gaps for women and people with disabilities*. Tech. rep. 2025. URL: https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20250527-1 (cit. on p. 27).

[62] Prague Process. *Albania - Migration Profile 2025*. Tech. rep. ICMPD, 2025. URL: https://www.pragueprocess.eu/en/countries/826-albania (cit. on p. 27).

[63] International Labour Office. *Skills mismatch of natives and immigrants in Europe*. Tech. rep. International Labour Organization, 2016. URL: https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@ed_protect/@protrav/@migrant/documents/publication/wcms_548911.pdf (cit. on p. 28).

[64] Eurostat. *Migrant integration statistics - over-qualification*. Tech. rep. European Commission, 2025. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php/Migrant_integration_statistics_-_over-qualification (cit. on p. 28).