Data & Policy (2026), 8: e35
doi:10.1017/dap.2026.10076

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Comparing AI-based approaches to candidate screening: recommendations for policy and regulation

Laura Vásquez-Rodríguez[1,2,3], Bertrand Audrin[4], Samuel Michel[2], Samuele Galli[5], Julneth Rogenhofer[4], Jacopo Negro Cusa[5] and Lonneke van der Plas[6]

[1]Doodle AG, Switzerland.
[2]Idiap Research Institute, Switzerland
[3]Work done while at Idiap Research Institute
[4]EHL Hospitality Business School, HES-SO, University of Applied Sciences and Arts Western Switzerland, Switzerland
[5]Arca24.com SA, Switzerland
[6]Università della Svizzera italiana, Switzerland
Corresponding author: Bertrand Audrin; Email: bertrand.audrin@ehl.ch

Received: 22 September 2025 UTC; **Revised:** 28 January 2026 UTC; **Accepted:** 01 May 2026 UTC

Keywords: artificial intelligence; candidate screening; human resource management; natural language processing; skill extraction

Abstract

Skill extraction is at the core of algorithmic hiring. It is based on identifying terms commonly found in both targets (i.e., resumes and job offers), aiming at identifying a "match" or correspondence between both. This article focuses on skill extraction from resumes, as opposed to job offers, and considers this task both from the human resource management (HRM) and artificial intelligence (AI) points of view. We discuss challenges identified by both fields and explain how collaboration is instrumental for a successful digital transformation of HRM. We argue that annotation efforts are an ideal example of where collaboration between both fields is needed and present an annotation effort on 46 resumes with 41 trained annotators, resulting in a total of 116 annotations. We analyze the skills extracted by multiple different systems and compare those to the skills selected by the annotators, and find that the skills extracted differ a lot in terms of length and semantic content. The skills extracted using conversational large language models (LLMs) tend to be very long and detailed; other systems are concise, whereas humans are in the middle. In terms of semantic similarity, conversational LLMs are closer to human outputs than other systems. Our analysis proposes a different perspective to understand the well-studied, but still unsolved, skill extraction task. Finally, we provide policy recommendations for skill extraction aligned with both HRM and computational perspectives.

Policy significance statement

As artificial intelligence (AI) becomes increasingly embedded in hiring processes, policymakers have to deal with a broad variety of systems, all of which have their strengths and limitations. Our findings highlight three key tensions that require policy attention: (1) representativeness, as training data and taxonomies often fail to capture the contextual specificity of skills; (2) accuracy versus interpretability, as more accuracy appears to come at the cost of interpretability; and (3) transparency across stakeholders, as current regulations emphasize deployers but leave job applicants with limited insight into how their own data are processed. Addressing these tensions appears to be fundamental for fair, accountable, and responsible AI in hiring.

This research article was awarded Open Data badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

1. Introduction

The role of technology in human resource management (HRM) has always been important (Marler and Fisher, 2013), but the recent developments of artificial intelligence (AI) and natural language processing (NLP) have offered new promises for HRM practices (Cheng and Hackett, 2021). Under the label of "HR Tech" (Nyathani, 2023), many traditional HRM practices have taken an AI turn throughout the whole employee life cycle, managing activities such as employee onboarding, performance management, and learning (Prikshat et al., 2023). In this article, we focus on the recruiting stage, where companies often receive hundreds of applications for a single position (Davis and de la Parra, 2024). Automation thus appears to be the only way to deal with such high volumes of applicants, overlooking the human aspect and often limiting the contribution of HR professionals.

In the field of NLP, we observe several attempts at automating candidate screening. Nowadays, skill extraction is the preferred and straightforward approach for automatically matching potential candidates with job offers (Shi et al., 2020). Skills, knowledge, certifications, and experience are identified in both texts, and the similarities between both are used to approximate the suitability of a given candidate for a specific job. The extraction of relevant terms can be performed using rule-based systems, as well as machine learning approaches using semantic similarity (Gugnani and Misra, 2020), and, recently, conversational large language models or LLMs (Li et al., 2023). Such skill extraction methods often limit the role of HR specialists and focus on the extraction from job offers because these data are more easily accessible.

The main objective of our study was to identify the challenges associated with the skill extraction task, in particular from resumes. We discuss this task both from the HR and computational perspectives and show how both perspectives are important especially when it comes to the annotation of resumes. We used existing computational methods to shed light on the task of skill extraction (Padmanabhan et al., 2022). We constructed a dataset for skill extraction by running an annotation task on 46 resumes with 41 trained annotators, resulting in a total of 116 annotations. We compared the nature of the skills identified by humans and machines from a set of resumes we collected, thereby showing (1) how the skills extracted by different systems vary in terms of length and semantic content and (2) how they differ from skills extracted by humans. These analyses not only support the understanding of the challenges of a well-studied but still unsolved task, but they also suggest that a shift to a human–AI teaming perspective would be beneficial. We finally provide a set of recommendations regarding skill extraction, which aims to bring together AI researchers, HR professionals, and policymakers.

2. Background

In this section, we present the skill extraction task from different perspectives. First, we discuss the HR perspective and analyze the process of resume screening (Section 2.1). Next, we discuss the skill extraction task from a computational perspective (Section 2.2). Then, we explain how the annotation of resumes lies at the intersection of both fields (Section 2.3). Finally, we discuss the policies related to our work within the field of NLP (Section 2.4).

2.1. Resume screening from an HR perspective

On average, recruiters spend 10 seconds on a resume (Hangartner et al., 2021). Their focus is thus very likely to be on high-level aspects of the resume rather than on details, which might go unnoticed. This short time frame also suggests that the recruiters' experience and perception play a major role and that elements such as phrasing or visual aspects can have a strong influence on resume screening (Arnulf et al., 2010).

The whole challenge of an initial resume evaluation lies in matching a position with a profile (Kristof-Brown, 2000) based on limited information. In some instances—for conventional positions such as "software developer"—the specific technical skills that are required can be easily mapped out and identified in resumes (such as "Python" or "Scrum Master"). In other instances, some position requirements in terms of skills might be harder to identify and are sometimes even open to interpretations,

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

especially if soft skills are the main inputs. Soft skills are a combination of personality traits, goals, motivations, and preferences that tend to be increasingly valued in the workplace (Heckman and Kautz, 2012). These tend to be expressed in multiple ways according to the experience and background of the candidate. Hence, the detection and standardization of soft skills are also more challenging aspects (Khaouja et al., 2019). Another challenging aspect is related to the job offers themselves, as many required skills are not explicit, and in many companies, job offers are generic and not tailored to the specific requirements of a job (Gugnani and Misra, 2020). In contrast, resumes are highly heterogeneous, as they strongly depend on the candidates and their background (industry, culture, and language), which affect how ideas, skills, and the overall resume are presented (Sajid et al., 2022).

Another perspective to keep in mind is that of candidates who face the challenge of having to convince HR professionals as well as algorithms, with potential tensions between both. On the one hand, presenting a convincing resume to HR professionals requires visual cues and formatting; on the other hand, presenting a convincing resume to an "algorithmic audience" requires focusing on appealing keywords tailored for a specific job posting. Candidates can use elaborate templates to make their resumes stand out to the human eye while also relying on conversational LLMs to craft their resume for a specific job offer. This results in a tension between customization and standardization in the content and styles of resumes, making preselection even harder.

2.2. Skill extraction from a computational perspective

Algorithmic hiring has focused on skill extraction and has become a prominent field of study (Zhang et al., 2023; Senger et al., 2024). Previous work encompasses both hard skills (Goyal et al., 2023; Zhang et al., 2024) and, to a lesser extent, also soft skills (Sayfullina et al., 2018; Zhang et al., 2022a,b). An even smaller number of works also focus on the precise match of candidates and job offers (Guo et al., 2016).

Often, the skill extraction task is tackled by using a taxonomy of terms to search (Cenikj et al., 2021) for coincidences in the text (e.g., rule-based system). A taxonomy is a collection of terms that interconnect occupations with skills for a specific domain as shown in Figure 1. In this scenario, a skill is extracted when words or phrases in the taxonomy are identified in the job description or resume. Due to its strict adherence to matches between words and phrases, these rule-based systems are conservative in that they extract only a limited number of often quite precise skills. This behavior also explains why these systems often overlook certain skills that might have been described in a way that deviates from the terms and phrases described in the taxonomy. Nevertheless, rule-based systems are effective and easy to interpret also thanks to the direct link to an overarching taxonomy (Rudin, 2019).

As an alternative, machine learning methods have been incrementally introduced, where models can learn to extract terms that have not been seen before. The advantage of these methods is that their ability to learn can help these models to generalize, that is, to find different ways of expressing the same skill. We can roughly classify previous work into models that rely on annotated data to fine-tune pretrained language models to identify skills in text (Green et al., 2022; Zhang et al., 2022a,b) and those that do not rely on annotated data but use conversational LLMs to generate skills when given a prompt containing instruction and either a job description or resume as input (Magron et al., 2024; Nguyen et al., 2024). The

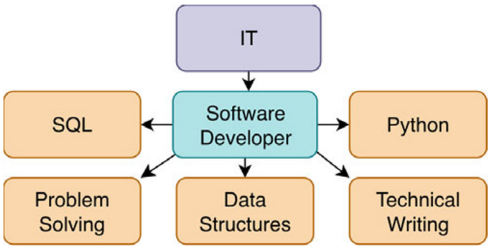


Figure 1. *Example of a skills taxonomy for the software developer occupation (IT domain).*

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

main drawback of the former is the fact that these models depend on annotated data for any given language or domain of interest. The main drawbacks of the latter are relatively high computational cost and limited control over the output. Previous work has demonstrated that the output of conversational LLMs for skill extraction results in 36% disagreement with humans (Nguyen et al., 2024). We elaborate on the annotation challenge in Section 2.3.

Multiple limitations can be found in previous work on automatic skill extraction (Zhang et al., 2022a,b; Goyal et al., 2023; Zhang et al., 2024). Firstly, the focus has been mainly on extracting skills from job descriptions (Goyal et al., 2023; Zhang et al., 2024) and largely ignored the task of extracting skills from resumes, which is an indispensable element of any applicant tracking system (ATS). One of the reasons for this discrepancy is practical. Resumes contain personal data and are therefore very hard to obtain and release as research data—nowadays often required for publishing research papers. Moreover, using commercial LLMs for extracting skills from resumes raises severe privacy concerns. However, resumes are very different from job descriptions, and a system that performs well on skill extraction from job descriptions is not guaranteed to work well on resumes. Resumes often use complicated layouts, for reasons discussed earlier, that are hard to parse automatically, among other issues.

Previous work has overlooked the fact that businesses often have clients from all parts of the world and many different expertise areas. Annotated data are not available for all languages and domains, and model transfer may also fail. The definitions and identification of what a skill is might also vary. For example, in the medical domain, certifications are important, while in business domains soft skills are essential. To alleviate issues related to data scarcity, datasets are synthetically generated using conversational LLMs (Decorte et al., 2023; Magron et al., 2024). Such methods could potentially also mitigate issues regarding privacy. However, the relevance of these methods for a skill identification task that can be considered as subjective, to an extent, is still questionable (Li et al., 2023).

The difference between manual resume screening and automatic skill extraction is quite large. First, computational systems are not trained on screening resumes based on visual elements. Second, the skill extraction task is almost always decoupled from the actual job offer in computational methods. In the Section 2.3, we will explain what consequences this has for the manual annotation of resumes.

2.3. AI and HR intersection: annotations

Supervised machine learning approaches for skill extraction require the collection of human annotations (Tamburri et al., 2020). Traditionally, researchers in the AI field work with annotators from the HR domain to perform this task (Zhang et al., 2022a,b). HR professionals then create annotations independently, according to defined guidelines, and often, annotations are rarely published (Senger et al., 2024). In this project, we created annotations and annotation guidelines in a collaboration between HR researchers, AI researchers, and HR tech professionals to include all three perspectives. We have learned several important aspects regarding the annotation process that we discuss below.

Overall, the manual annotation of job offers and especially resumes is an extremely subjective task. First, it is important to note that resumes represent a very specific form of text that is highly fragmented (resumes often feature only a few full sentences but a high number of lists and bullet points) and codified (specific information is expected in a resume, such as education and experience). However, candidates benefit from a high level of autonomy within these constraints, which leads to substantial variability between different resumes, and with it complexity when annotating their content.

The context of the job application (e.g., type of position, language) process also plays an important role in shaping resumes. In fact, the entire purpose of the resume is for a candidate to showcase that their profile matches with the position that they are applying for. A resume is thus expected to "target" a job offer. This is critical in two respects: (1) how it influences the resume itself and (2) how it influences the annotation process. The context actively shapes the content of a resume: people applying for software developer jobs are more likely to emphasize the projects they have worked on and the programming languages that they mastered, whereas people applying to HR manager jobs are more likely to focus on their tasks and responsibilities and the results. The context also shapes the annotation itself. Indeed, HR professionals

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

always compare resumes with a job description and a job specification: there is always a point of reference that can be used to contextualize the experience and the skills that can be found in a resume. An out-of-context resume is thus more difficult to analyze and its annotation requires more expertise and stricter guidelines.

In that respect, annotators require a high level of expertise in the domain of the resumes that they are annotating (e.g., computer science, finance, and HRM), as they are not able to rely on existing job offers to assist them in their analysis. This expertise also needs to be complemented with a clear set of guidelines that clarify what to annotate and how. This is particularly critical in skill annotation, as different forms of competence can be featured in a resume (e.g., theoretical knowledge, practical experience, abilities, etc.), which can lead to confusion in annotations. Having clear categories of skills with examples for different contexts thus helps reduce uncertainty and increase reliability throughout the annotations. Given these conditions, annotating skill datasets for AI represents one of the biggest challenges in developing HR tech solutions, and considerable time should be reserved for this phase.

2.4. Policies for NLP and HR

In this section, we will dive into the connection of our work to policy-making. Overall, policy development on AI matters has tended to lag behind technological innovation. While basic neural networks were invented decades ago, it was the recent, rapid advancement of powerful conversational LLMs that prompted regulations like the EU AI Act (European Parliament and Council of the European Union, 2024). In response to new regulations, the research community has begun to analyze specific implementation strategies for compliance (Billah et al., 2025). For example, the accelerated developments in the field of NLP have created significant legal and ethical challenges that are now being addressed by policy. Primary concerns regarding the applicability of NLP methods include copyright issues related to training data and its handling after models are distributed (de Castilho et al., 2018). Other cases relate to the key issues of the appropriate use of social media data and how certain applications, such as exploratory research on disinformation (Peng et al., 2025), are still acceptable.

Nevertheless, the implementation of highly structured and general regulations, such as the EU AI Act, brings to light new questions that fall outside of what is currently framed. This has prompted the community to propose more granular rules, such as the proposition of risk-based approaches and the development of formal ethical frameworks to address these emerging issues (Sillberg et al., 2025). We frame our policy discussion within the domain of HR, with a specific focus on the EU AI Act, as it is the key framework regulating our work. While the legislation offers the classification of AI systems based on the level of risk into three classes: prohibited, high-risk, and general-purpose AI, the applications in the field of HR are very likely to be designated as "High-Risk" due to the following statement:

> Employment, workers management and access to self-employment: AI systems used for recruitment or selection, particularly targeted job ads, analysing and filtering applications, and evaluating candidates. Promotion and termination of contracts, allocating tasks based on personality traits or characteristics and behaviour, and monitoring and evaluating performance (European Parliament and Council, 2024).

The replacement of human judgment with autonomous AI systems carries significant risks, including strong algorithmic biases that could drastically affect the job market and individuals' careers. We refrain from developing methods for direct candidate selection in order to avoid high-risk use of AI. Instead, we propose an approach centered on creating AI systems that assist HR professionals with enhanced information, facilitating their decision-making rather than fully automating the hiring process. In Section 3, we will focus on a specific use case to characterize different AI-based approaches to skill extraction and how they differ from human annotations. We will continue our discussion of the policy implications and recommendations in Sections 5.3 and 5.4.

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

3. Methods

To study the challenges associated with the skill extraction task in resumes, we divided our work into two parts. First, we collected resumes and performed human annotations on these documents. Second, we compared human annotations with those generated automatically (i.e., with rule-based and AI-based systems) and using the same corpus (detailed in Section 3.2). In addition, we performed a comparative analysis (Section 3.3) regarding the length and semantic content of our annotations to better understand the challenges of the annotation task from both human and computational sides.

3.1. Data collection

We collected a resume-based dataset suitable for the skill extraction task. To achieve this goal, we organized a workshop with students of a graduate-level Talent Management course in the hospitality domain. The main task was to craft their resumes to collectively gather further annotations. In total, we obtained the data consent to annotate 46 resumes out of 60. Students were provided with clear annotation instructions, and a training was performed. The annotators were then divided into groups of three or four people, and individually, they annotated each resume. Each resume was split into paragraphs (if possible), annotating in each case the corresponding label into three categories: Hard Skills/Certifications, Soft Skills, and Occupations/Positions. Forty-one students participated in the workshop, which achieved a total of 116 annotated texts, given that each CV would be annotated multiple times. Also, due to the variability in the annotators' speed, some students managed to annotate more texts, completing a total of 46 CVs, after aggregating all the annotations. We report the statistics of the dataset in Table 1 and refer to this corpus as the *EHL workshop dataset.* As a result, each resume would have a single list of labels (unique/nonunique) as shown in Table 3 (Human). We also removed inconsistencies such as different characters for separating skills. For the scope of this articles, we have selected only the analysis of hard skills, as soft skills are difficult to detect using automatic methods. In future work, we will include the detection of soft skills in our tools and taxonomies.

3.2. System selection and implementation

To compare automatic systems with human annotations, we first selected a **rule-based system** which is often used for skill extraction and is usually straightforward and low-cost. A rule-based system is a program that searches keywords based on a defined set of rules or taxonomy. For our use case, we have an *in-house* industrial taxonomy with triplets that define a domain, an occupation, and its corresponding skills as shown in Figure 1. This system was implemented using the open-source SkillNER tool as a baseline (https://github.com/AnasAito/SkillNER), where labels were identified by matching the text against the provided taxonomy using predefined rules.

For the AI-based methods, we selected two approaches: a supervised model and a conversational LLM. The **supervised model** is trained on annotated corpora, specializing the model with domain-specific knowledge, such as job offers or skill extraction concepts, and on the language of the datasets. These supervised models tend to be smaller than a conversational LLM, and they do not need large amounts of examples to perform a task (i.e., using fine-tuning methods). For our experimental setting, we selected a state-of-the-art supervised system, ESCOXLMR (Zhang et al., 2023). This model has been previously trained on the ESCO taxonomy (i.e., mainly skills/competences and occupations), and samples of

Table 1. *Statistics for the EHL workshop dataset of 46 resumes*

	Mean per resume	Standard deviation per resume	Total values
Words	465.391	181.214	21,408
Annotators	2.522	0.836	41

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

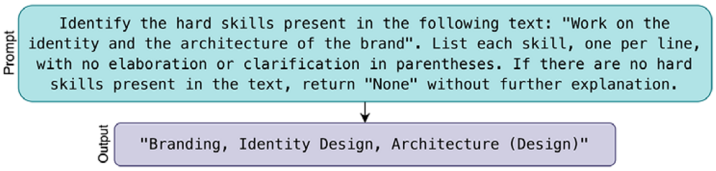


Figure 2. *Prompt example for annotating resumes with LLMs.*

Table 2. *Statistics for skills extracted*

Skills	Mean per resume	Mean per resume (unique)	Total values	Total (unique)
Human	30.522	24.478	1404	1126
Rule-based	16.478	12.848	758	591
Supervised	37.022	36.022	1703	1657
Conversational LLM	52.391	49.652	2410	2284

annotated job offers (ESCO, 2022). There are two specializations of the model: detection of skills and knowledge (Zhang et al., 2023). We used these models to annotate the text in our dataset with no further training.

Finally, **conversational LLMs** are trained on large-scale corpora and specifically tuned to generate responses in a dialogue-oriented format. They are typically used off the shelf, as retraining or fine-tuning them requires substantial computational resources. For our skill extraction setting, we selected the open-source model *Mixtral 8x7B* Instruct v0.1 model that can be run locally given that resumes handle personal information, in contrast to online commercial models that upload the data provided in each query (Nguyen et al., 2024). To extract the target skills, we designed a prompt that included a set of instructions with the model task and explicitly formatted the input text as shown in Figure 2. Taken together, we use these models to automatically extract skills from the *EHL workshop dataset.* We present the collected annotations (Vásquez-Rodríguez et al., 2025) and their statistics in Tables 2 and 3.

3.3. A comparative analysis

We conducted a comparative analysis of the annotations based on their surface properties (i.e., skill length) and based on their meaning (i.e., semantic similarity of skills). For the **length analysis**, we calculate the average length of the extractions within each resume, independently between systems and humans. We normalized all the observations using the *percent* statistic via Seaborn's visualization framework, so labels from each source will represent a portion within totality (i.e., 100%) of the observations. Second, we assess the differences in the average length per resume. For the length analysis, the implementation is straightforward, as we report the total of characters in a skill and then average for all the labels in the resume.

For the **semantic analysis**, compare the meaning of extracted skills between the following pairs: human versus LLM, human versus rule-based, and human vs supervised. This experiment aims to reflect similarities between skills collected for a given resume by humans and the different systems. We calculated the semantic similarity between resumes using the cosine similarity metric. This metric represents documents (i.e., groups of skills) as abstract representations where the distance between them reflects how close they are in meaning. We calculated the similarity between text representations using the SBERT (Reimers and Gurevych, 2019) model, specifically using the all-mpnet-base-v2 version, which reports the highest average performance for semantic similarity. Similar skills would have a value near 1, while dissimilar labels would show values close to 0. Similar to the length analysis, we analyze the distribution (by percentage) of the similarity values from the systems' outputs. We also include an analysis

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

Table 3. *Manual and automatic annotations from our dataset*

Source	Annotations
Human	negotiation, market research, project planning, language skill-English, Microsoft 365, training skills, tlb management, marketing and strategy, culinary certificate, interview experience
Rule-based	hotel led, contractor, management graduated, mathematics, sports, prepared materials, digital strategy, business schools, sales consultant, operational processes provided support
Supervised	maintained consistent communication with customers, nexus, overseeing, led the role of a project manager, my vision, degree, and maintained real estate project tracking and portfolio, systems, with market research, the restaurants daily operations, sustainability
Conversational LLM	Forbes standards, contract negotiation, coaching, multitasking, brand image and communication strategy, menu planning, budget & financial management, carrying out feasibility studies of new projects/refurbishment projects, design, managing large groups

Note: We selected 10 random annotations from each model.

per resume to understand the individual contribution to the overall scores. For both analyses, it is important to note that it is very unlikely to have a 1:1 correspondence between all skills.

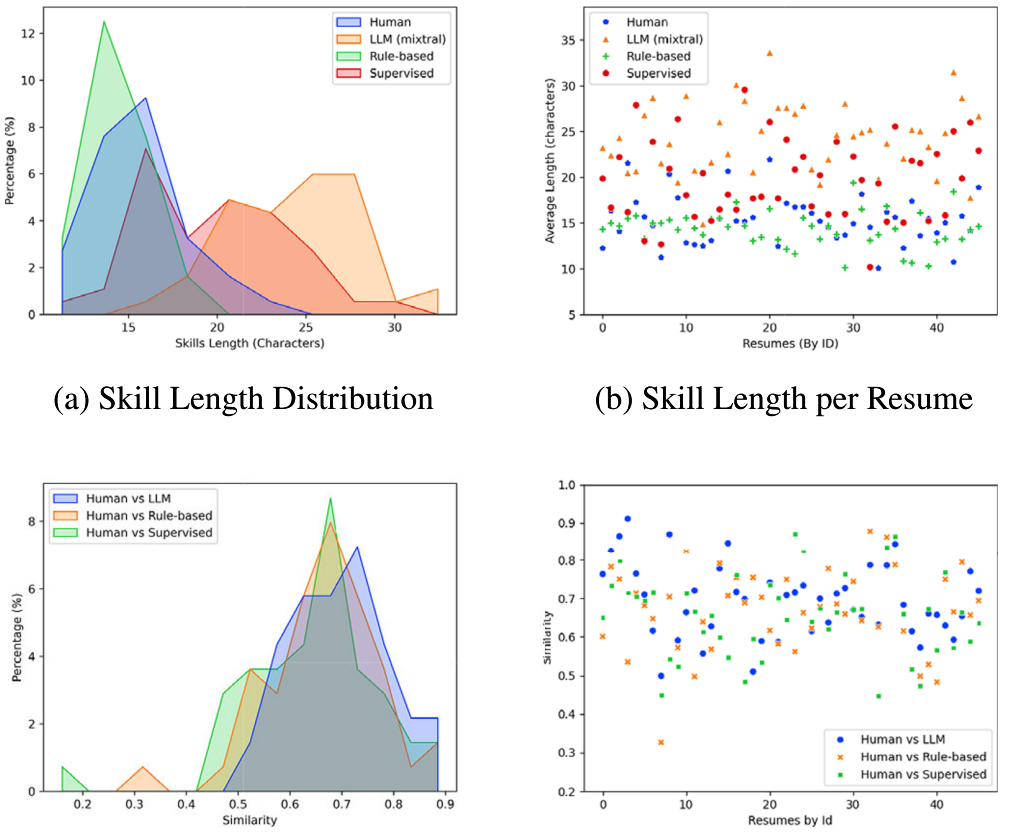
4. Results

In this section, we present our results from our experiments in Section 3.3 Figure 3a highlights the distribution of the skill lengths, aggregating all resumes for each data collection (e.g., human and supervised). We can observe that most human annotations have an approximate length of 15 characters, while skills generated by conversational LLMs are generally more verbose, with varying lengths between 15 and 35. In contrast, the rule-based system is shown to be more concise but still closer to the length of human annotations. The supervised method results in extractions that are more wordy than human annotations, but still more concise than the conversational LLM. From the analysis per resume (Figure 3b), it is clear that conversational LLMs produce the longest extractions, with some exceptions from the supervised method.

Figure 3 presents the results of the semantic analysis. In Figure 3c, we plotted the semantic similarity of outputs of the three systems in comparison with human extractions aggregated over all resumes. We see that most of the documents were somewhat related, with values classified between 0.5 and 0.8. However, there were no examples with values between 0.9 and 1, showing that there were no exact matches. Also, we present the semantic similarity per resume in Figure 3d. We compared the similarity of the human annotations against all the systems. Once again, no system produced outputs that were highly similar to human output, but there was considerable semantic overlap because the similarity between them still ranges between 0.5 and 0.8, with the conversational LLM showing the highest similarity with human content.

Finally, to better understand the quality of our quantitative analysis, we provide a list of extractions, both from humans and systems, in Table 3. We show a list of 10 randomly selected skill extractions for each model. We argue that the random selection of samples helps to understand the variability of the labels (in size and semantics) of each model. Furthermore, in Table 4, we manually mapped common concepts from five random resumes to point out some distinctions and similarities between the different approaches.

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press



(a) Skill Length Distribution

(b) Skill Length per Resume

(c) Similarity Frequency

(d) Similarity per Resume

Figure 3. *Comparative analysis of skill length and semantic similarity across human annotations and automatic systems. (a) Distribution of extracted skills by character length (x-axis), reported as percentages (y-axis). (b) Average skill length (characters, y-axis) per resume (x-axis). (c) Distribution of semantic similarity scores between human annotations and automatic systems (x-axis), reported as percentages (y-axis). (d) Semantic similarity scores (y-axis) between systems and across individual resumes (x-axis).*

5. Discussion

We focus our discussion on four avenues: a comparison of the system's capabilities and its limitations (Section 5.1), the importance of interdisciplinary collaborations and trust in AI systems (Section 5.2), the implications of our experiments for policy-making (Section 5.3) and HR practice (Section 5.4).

5.1. Model comparison and variability

First, we comment on our comparative analysis. Rule-based systems are shown to be more standardized, as they adhere to a predefined taxonomy where skills have a maximum length. In contrast, conversational LLMs are generative models (i.e., they output text based on previously seen texts) that output not only explicitly mentioned skills but may also infer content from the text. Supervised models tend to stay closer to the actual task they have been explicitly trained to do but will be able to abstract away from literal mentions of a particular skill as given in a taxonomy. The aim of training such models is for them to be able to generalize well to previously unseen skills. It is difficult to conclude the superiority of one system over the other from these analyses, and this is also not the aim of this work. These analyses bring to light some

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

Table 4. *Common annotations between humans and our proposed models*

Human	Rule-based	Supervised	Conversational LLM
Data collection and analysis	Data analysis	Data analysis	Determined the origin of lots/items through desk research, data, and sources analysis
Budget management	Budget	Estimated tender budget	Estimated tender budget
Interview experience		Conducted face-to-face interviews with airport users	Conducted face-to-face interviews
Customer service, office	Front desk	Provide customers with thoughtful service, reply to, serve every guest warmly and thoughtfully, on the phone, with customers	Receiving phone calls, customer service, familiarity with customer check-in and check-out procedures

characteristics of these system outputs. In the context of this study, conversational LLMs show the highest overlap in semantic content, but they are also extremely wordy; rule-based keyword-matching systems produce very short outputs that are linked to the items found in a taxonomy that seem to cover the semantic content a bit less well; and supervised are somewhere in between.

Concerning our analyses, we highlight that these results should be considered with caution, as they result from a small dataset of 46 observations. We acknowledge that their generalization can be limited and further investigations with larger datasets are needed to formulate more decisive statements. Nevertheless, the results from our experiments can help create awareness of the variability in output across different model architectures and humans beyond pure numbers. Concerning the conversational LLM-based skill extraction, we present the results for a specific LLM—*Mixtral-8x7B model*—with our customized, single prompt. These models are highly variable and nondeterministic, so results may vary according to the architecture and prompts. Prompts could include a length limitation, for example. Although the definition of a taxonomy in the context of the LLM could mitigate the variable nature of the skill extraction, these could be extremely large, comprising over 13,000 distinct skills in the case of ESCO, representing nontrivial challenges for its implementation.

As mentioned before, the human identification of skills is a subjective task due to differences in background, domain knowledge, and appreciation. Despite this, the automation of skill extraction has been widely implemented without systematically accounting for such variability. In this article, we show that systems extracting skills from the same resume also vary substantially in their outputs. While these differences may appear intuitive, they are often overlooked or misunderstood in practice. Candidates might be aware that their CVs are bound to be screened by an ATS, but they often do not realize that different systems might lead to different recommendations, which represents a high risk without human oversight. This has important implications for the fairness and transparency of how candidates are evaluated.

From our findings, we recommend developing clear and well-defined guidelines to ensure as much consistency in the data as possible. While annotation guidelines have been released in the past (Senger et al., 2024), following an elaboration process that is specific to the hiring and recruitment context and involves HR professionals is instrumental to creating clear documentation and ensuring a more structured and valid process.

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

5.2. Interdisciplinary collaborations and system trust

Finally, we discuss the benefits of our collaboration efforts in our HR tech project and the relevance of trust in skill extraction systems.

First, we noted the importance of NLP researchers to work closely with HR researchers so that both sides benefit from each other's expertise. We also experienced that it took some time for the interdisciplinary team to find a common language and develop a mutual understanding of the task at hand and the contributions from both sides. This has been particularly instrumental in the annotation task to ensure that the output meets the requirements from both NLP and HR perspectives. In the same vein, partnering with industry is also fundamental, not only to understand the practical application of a given solution (e.g., skill extraction) but also for the implications and benefits to the business, which could include the adaptation to existing taxonomies, limited computational resources, and compatibility with existing solutions. In the context of this project, this has led us to focus on how to meet objectives from an industry perspective and also in terms of academic research: developing experimentation and novelty to thrive, while at the same time keeping an eye on cost, time-effectiveness, and long-term maintenance of systems.

Another important aspect is the trust in selection algorithms (Groß, 2021). While there is a strong debate among practitioners on whether algorithms can be trusted or not for selection decisions, our results highlight a certain degree of variability between systems and human annotations in terms of length and meaning. Also, we explained that the output from certain methods (rule-based) is easier to explain than from others. These observations are linked to well-known challenges of explainability (i.e., why did the system output a certain result?), precision, and overall trustworthiness of AI-based recruitment systems. In that respect, building robust, interpretable, and explainable systems seems to be a priority to increase the trust of users and the adoption of the technologies, also in view of AI regulations.

5.3. Implications and recommendations for policy

Our results will help formulate policy implications, specifically with respect to potential bias, performance, and interpretability of AI methods. We highlight that both human and automatic methods come with their own biases when it comes to hiring. These aspects are also directly correlated with the established policy frameworks, which guide end users to their right applicability. We discuss these aspects in Sections 5.3.1–5.3.4.

5.3.1. Data representativeness and systematic bias

Rule-based systems are strongly dependent on their taxonomy and how well it covers the corpus to be analyzed (e.g., a resume). Some terms related to skills or qualifications may be missing from the taxonomy, which can result in an inaccurate identification of applicants' skills and the discarding of potential and relevant candidates. The fact that some terms are missing can stem from the evolution of work (e.g., the development of new skills and new qualifications) or from choices from the institution developing the taxonomy (e.g., not including soft skills). For supervised systems, biases can originate from the dataset they were trained on, including its size, as well as from other factors, such as imbalanced demographic data and uneven class distribution. This is nontrivial in hiring, as the variety of skills and qualifications usually extensively exceeds what the initial corpus consists of.

Dataset-related limitations can prevent the ability of such systems to accurately identify applicant skills if they come from a different line of work or background than the one the system has been trained on. In hiring, this is associated with challenges in terms of data type and contextualization. First, many academic supervised systems are trained on job offers and not resumes (Senger et al., 2024). While job offers present some similarities with resumes, the vocabulary that is used in resumes differs greatly and has more variability. The same skill can be expressed differently by each candidate, for example, "Software Developer" and "Software Engineer." Second, supervised systems, even if based on resumes, have to select a pool of resumes that are associated with an industry and location. These selections can also be subject to privacy concerns if not crafted carefully, as anonymization tools tend to perform poorly (Magron et al., 2024). Overall, codes and norms on what to include or not in resumes strongly differ

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

between industries and locations, thus leading to challenges in terms of accurate skill identification. The same reasoning applies to conversational LLMs, but on a greater scale and with even less oversight on the characteristics of the data included, and thus less room for bias mitigation.

These aspects imply that representativeness can be difficult to define, as it refers to different types of intersecting contexts (occupations, industries, locations, and careers, among others). Nonetheless, providing actors with a clear explanation of expectations toward representativeness can be key to assessing any system, whether rule-based, supervised, or a conversational LLM. For example, if a candidate screening system has been trained on job offers, would this qualify as "representative" if the content it is applied to consists of resumes, which contain a high variability of skills? The span of representativeness, thus, would require more clarification in policies.

5.3.2. Interpretability and generalizability

With regard to the aspect of interpretability, rule-based systems are the most interpretable, as we can clearly understand the internal mechanisms of the selection algorithms. Supervised systems and LLMs, on the other hand, are less interpretable, especially because of the complexity of their internal mechanisms (Vásquez-Rodríguez et al., 2024). However, the latter possess better abilities to infer skills from existing information and generalize to multiple domains. Moreover, when a significant amount of context (i.e., data) is involved, it is difficult to explain the internal mechanisms of LLMs that boost a model's performance. This can create a conflict between goals as one seems to come at the cost of the other. Policy frameworks can guide developers and organizations in setting priorities and making critical design choices. For example, they can establish whether a small trade-off in predictive accuracy is acceptable in exchange for a more interpretable system. This could involve using a simple, rule-based system or employing specific AI libraries (Hasan, 2024) to analyze a model's decisions and show which factors most influenced a particular outcome.

5.3.3. Task subjectivity

In a broader dimension, the nature of skills themselves requires a form of interpretation. Indeed, a skill is commonly defined as "the ability to retrieve knowledge and apply it to a task in a proficient manner" (Lamri and Lubart, 2023). What this definition highlights is the contextual nature of skills, a context that rule-based systems greatly ignore. For example, "reporting" means something very different whether you are an accountant or a journalist. In that respect, there are clear tradeoffs that have great implications in terms of policies, as policy-related decisions have to account for (1) the variety of the systems that are currently labeled as "AI hiring," (2) the level of risk for these AI-based systems and, (3) balancing the trade-offs for the development of responsible AI algorithms, in aspects such as fairness (minimal bias), interpretability, and privacy (Kemmerzell and Schreiner, 2024).

5.3.4. Considerations for the EU AI Act

There are current policies that regulate the use of AI in HRM activities (Autischer et al., 2025), but some key questions are only partially addressed. Some of it pertains to the fact that policies tend to adopt broad labels (i.e., "high-risk AI system" in the EU AI Act) to refer to a broad spectrum of technical systems, which all have different characteristics and outcomes.

From a user's perspective, the EU AI Act primarily focuses on regulating AI deployers (organizations and intermediaries), yet it creates a significant gap in transparency for a key stakeholder: the job applicant. This lack of clarity on how AI systems make decisions (especially when data (i.e., resumes) and systems vary greatly across companies) creates an unfair hiring environment. For example, a candidate applying for the same position at two distinct companies could be assessed drastically differently by AI systems that lack transparency and a common standard. To ensure fairness and accountability, we advocate that future policy should mandate transparent processes for all users, including a clear right for applicants to understand how AI-driven decisions are made.

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

5.4. Implications and recommendations for HR practice

Our results also make it possible to formulate recommendations for HR professionals and candidates.

For HR professionals, the differences between distinct types of ATSs and the way they extract skills from resumes are fundamental. Being aware of how various systems function and understanding their respective pros and cons (i.e., regarding interpretability and generalization) is fundamental to making a choice that best aligns with their needs. For example, in a company with very standardized positions, a rule-based system might be more efficient, as it can more effectively extract mandatory skills whereas a conversational LLM-based ATS might be more relevant for companies with more variable jobs that look for less-fixed profiles.

Moreover, as the results show a great amount of variability, it is fundamental to avoid taking the results of any ATS for granted, as each system comes with its set of limitations. In that respect, the professional responsibility of HR practitioners is to critically review resumes of applicants that have been identified as potential matches by an ATS, as well as those of some other applicants, to ensure that the ATS did not make errors.

Finally, while CV screening can be useful to exclude irrelevant applications, it does not represent a selection tool with a high predictive validity. In that respect, the first assessment of resumes made by an ATS should only be considered as an early indication that needs to be complemented by other selection methods with higher predictive validity (such as structured interviews, cognitive ability tests, situational judgment tests, etc.).

For candidates, it is crucial to understand that their CVs are likely to be processed by various automatic systems, with distinct ways of processing their documents. This means first making sure that skills can be identifiable and extracted by systems that do not operate in the same way. In that respect, focusing on a thorough alignment with terminology is very helpful.

Similar to HR professionals, candidates should keep in mind that the ATS only represents the "first filter." And while it is important to make it through this first filter, the end goal is to build a long-lasting relationship with an employer by successfully going through a more complex selection process. In that respect, the frontier between tailoring one's resume to make sure that skills can be extracted and inflating one's skill set can be quite blurred. Nevertheless, the following steps of the selection process are likely to highlight discrepancies between the documents and reality.

6. Limitations and future research directions

We have proposed the comparison of multiple models as a means to understand the variability of different sources. Nevertheless, it is important to state their limitations. For the supervised setting, we chose publicly available state-of-the-art models; however, these are only trained on job offers, as resumes are not freely available. For the rule-based model, we are limited by the predefined taxonomy of skill concepts, thus, the performance will depend on the quality and nature of the taxonomy. Concerning the conversational LLMs, we chose a zero-shot approach to generate extractions from our dataset. However, it would also be beneficial to experiment with few-shot scenarios and in-context learning so that the models can see practical examples and also, it becomes possible to provide additional feedback to improve the quality of the output.

In future work, we would like to extend the number of annotated datasets to include resumes from various domains and languages. This will improve the generalization abilities of the models to more scenarios. With respect to the conversational LLMs, these are sensitive to minimal changes in the selected prompts. By using multiple prompting techniques and models, we could arrive at a deeper understanding of their performance. This includes the limitation on the length of candidates' skills and the introduction of external knowledge from taxonomies, or other domain knowledge. In particular, we would like to include soft skills, which are fundamental in the workplace and their identification is one of the main goals of our project. However, in order to have a fair comparison between the different approaches, which is our objective in this study, we would need a taxonomy that specifically focuses on soft skills as well. Such a

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

taxonomy, tailored to our needs, is currently not available, so we decided to leave soft skills out of this specific work while hoping to include them in future research.

7. Conclusion

In this article, we discussed the different perspectives of HR and AI on the skill extraction task. We focused on resumes, in contrast to previous work, and collected 46 annotated texts. These allowed us to perform comparative analyses of the output generated by three different computational methods on the task of automatic skill extraction. We showed the difference in the nature of what these computational systems produce. We also discussed the pros and cons of these approaches from several angles focusing on the human aspects fed by interactions with HR professionals and HR tech professionals, such as privacy, trustworthiness, and efficiency. We expect that our research has shed light on the variations between systems and how they differ from human annotations, emphasizing the role of this often-neglected actor in a field that tends to rely more and more on algorithmic hiring. Moreover, our results can inform policymakers, especially regarding how various systems deal with bias, interpretability, and performance.

Data availability statement. The data (Vásquez-Rodríguez et al., 2025) supporting the results of this research are permanently archived on Zenodo (https://doi.org/10.34777/peay-xe10). The ongoing development of the project can also be found on GitHub (https://github.com/idiap/human_skill_extraction).

Acknowledgments. We gratefully acknowledge the support from Innosuisse (grant 104.069 IP-ICT). In addition, generative AI tools were used in a limited capacity to support language editing, including grammar correction and paragraph structure improvement during manuscript preparation. No AI tools were used to generate original scientific arguments, interpret results, draw conclusions, or make policy recommendations. All substantive content, analysis, and interpretations remain the responsibility of the authors, who have reviewed and verified the manuscript in full.

Author contribution. Conceptualization:L.V.R., B.A., S.M., S.G., J.R., J.N.C., L.vd P.; Data curation: L.V.R., B.A., S.M., S.G., J.R.; Data visualization: L.V.R., S.M., S.G.; Methodology: L.V.R., B.A., S.M., S.G., J.R., J.NC., L.vd P.; Software: L.V.R., S.M., S.G.; Writing—original draft: L.V.R., B.A.; Writing—review and editing: L.V.R., B.A., S.M., S.G., J.R., J.N.C., L.vd P.; All authors have approved the final submitted draft.

Funding statement. This research was supported by grants from Innosuisse (grant 104.069 IP-ICT).

Competing interests. None.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Autischer G, **Waxnegger K and Kowald D** (2025) *AI certification and assessment catalogues: Practical use and challenges in the context of the European AI Act.* In Proceedings of Fourth European Workshop on Algorithmic Fairness (Weerts H, Pechenizkiy M, Allhutter D, Corrêa AM, Grote T and Liem C (eds.)), Proceedings of Machine Learning Research, *294*, PMLR, 30 June–02 July, pp. 492–498, Eindhoven University of Technology, Eindhoven, The Netherlands.

Billah MM, **Hamjaya HS**, **Shiralizade H**, **Singh V and Inam R** (2025) Large language models' trustworthiness in the light of the EU AI Act—A systematic mapping study. *Applied Sciences 15*(14). https://doi.org/10.3390/app15147640.

de Castilho RE, **Dore G**, **Margoni T**, **Labropoulou P and Gurevych I** (May 2018) *A legal perspective on training models for natural language processing.* In Calzolari N, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, and Tokunaga T (eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA).

Cenikj G, **Vitanova B and Eftimov T** (2021) *Skills named-entity recognition for creating a skill inventory of today's workplace.* 2021 IEEE International Conference on Big Data (Big Data), pp. 4561–4565 https://bigdataieee.org/BigData2021/index.html.

Cheng MM and Hackett RD (2021) A critical review of algorithms in hrm: Definition, theory, and practice. *Human Resource Management Review 31*(1), 100698. https://doi.org/10.1016/j.hrmr.2019.100698.

Davis SJ and de la Parra BS (2024) *Application flows. Working Paper 32320, National Bureau of Economic Research.*

Decorte J-J, **Verlinden S**, **Van Hautte J**, **Deleu J**, **Develder C and Demeester T** (2023) *Extreme multi-label skill extraction training using large language models. Computation and Language, arXiv:2307.10778.*

ESCO. 2022. Machine Learning Assisted Mapping of Multilingual Occupational Data to ESCO (Part 1). https://esco.ec.europa.eu/en/about-esco/data-science-and-esco/machine-learning-assisted-mapping-multilingual-occupational-data-esco-part-1.

European Parliament and Council of the European Union (2024) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Annex III. https://artificialintelligenceact.eu/annex/3/.

Goyal N, **Kalra J**, **Sharma C**, **Mutharaju R**, **Sachdeva N and Kumaraguru P** (2023) JobXMLC: EXtreme multi-label classification of job skills with graph neural networks. In *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2181–2191.

Green T, **Maynard D and Lin C** (2022) *Development of a benchmark corpus to support entity recognition in job descriptions. Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association*, pp. 1201–1208.

Groß M (2021) Yes, AI can: *The artificial intelligence gold rush between optimistic HR software providers, skeptical HR managers, and corporate ethical virtues, Management for Professionals*. In Vieweg SH (ed), *AI for the Good*. Cham: Springer, pp. 191–225. https://doi.org/10.1007/978-3-030-66913-3_10.

Gugnani A and Misra H (2020) Implicit skills extraction using document embedding and its use in job recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence 34*(8), 13286–13293.

Guo S, **Alamudun F and Hammond T** (2016) Résumatcher: A personalized résumé-job matching system. *Expert Systems with Applications 60*, 169–182. https://doi.org/10.1016/j.eswa.2016.04.013.

Hangartner D, **Kopp D and Siegenthaler M** (2021) Monitoring hiring discrimination through online recruitment platforms. *Nature 589*, 572–576.

Hasan MM (2024) Understanding model predictions: A comparative analysis of shap and lime on various ml algorithms. *Journal of Scientific and Technological Research 5*(1), 17–26.

Heckman JJ and Kautz T (2012) Hard evidence on soft skills. *Labour Economics 19*(4), 451–464. https://doi.org/10.1016/j.labeco.2012.05.014.

Kemmerzell N and Schreiner A (2024) Quantifying the trade-offs between dimensions of trustworthy ai - an empirical study on fairness, explainability, privacy, and robustness. In Hotho A and Rudolph S (eds), *KI 2024: Advances in Artificial Intelligence (Cham)*. Cham: Springer Nature Switzerland, pp. 128–146. https://doi.org/10.1007/978-3-031-70893-0_10.

Khaouja I, **Mezzour G**, **Carley KM and Kassou I** (2019) Building a soft skill taxonomy from job openings. *Social Network Analysis and Mining 9*(1), 43. https://doi.org/10.1007/s13278-019-0583-9.

Kristof-Brown AL (2000) Perceived applicant fit: Distinguishing between recruiters' perceptions of person–job and person–organization fit. *Personnel Psychology 53*(3), 643–671. https://doi.org/10.1111/j.1744-6570.2000.tb00217.x.

Lamri J and Lubart T (2023) Reconciling hard skills and soft skills in a common framework: The generic skills component approach. *Journal of Intelligence 11*(6). https://doi.org/10.3390/jintelligence11060107.

Li N, **Kang B and De Bie T** (2023) *Skillgpt: A restful API service for skill extraction and standardization using a large language model.* arXiv:2304.11060.

Li Z, **Zhu H**, **Lu Z and Yin M** (2023) *Synthetic data generation with large language models for text classification: Potential and limitations. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics*, pp. 10443–10461.

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

Marler JH and Fisher SL (2013) An evidence-based review of e-hrm and strategic human resource management. *Human Resource Management Review 23*(1), 18–36. https://doi.org/10.1016/j.hrmr.2012.06.002.

Nguyen K, **Zhang M**, **Montariol S and Bosselut A** (2024) *Rethinking skill extraction in the job market domain using large language models. Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024). St. Julian's, Malta: Association for Computational Linguistics*, pp. 27–42.

Nyathani R (2023) Preparing for the future of work: How hr tech is shaping remote work. *Journal of Technology and Systems 5*(1), 60–73. https://doi.org/10.47941/jts.1498.

Padmanabhan B, **Fang X**, **Sahoo N and Burton-Jones A** (2022) Editor's comments: Machine learning in information systems research. *MIS Quarterly 46*(1), iii–xix. https://doi.org/10.25300/MISQ/2022/461E1.

Peng Q, **Moro R**, **Gregor M**, **Srba I**, **Ostermann S**, **Simko M**, **Podrouzek J**, **Mesarčík M**, **Kopčan J and Søgaard A** (2025) *SemEval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval*. In Rosenthal S, Rosá A, Ghosh D and Zampieri M (eds.), *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Vienna, Austria: Association for Computational Linguistics, pp. 2498–2511.

Prikshat V, **Islam M**, **Patel P**, **Malik A**, **Budhwar P and Gupta S** (2023) Ai-augmented hrm: Literature review and a proposed multilevel framework for future research. *Technological Forecasting and Social Change 193*, 122645. https://doi.org/10.1016/j.techfore.2023.122645.

Reimers N and Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410.

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Sajid H, **Kanwal J**, **Bhatti SUR**, **Qureshi SA**, **Basharat A**, **Hussain S and Khan KU** (2022) Resume parsing framework for e-recruitment. In *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (pp. 1–8). IEEE. https://doi.org/10.1109/IMCOM53663.2022.9721762.

Sayfullina L, **Malmi E and Kannala J** (2018) Learning representations for soft skill matching. In Panchenko, A *et al*. (eds), *Analysis of Images, Social Networks and Texts. AIST 2018. Lecture Notes in Computer Science*, vol. *11179*. Cham: Springer International Publishing, pp. 141–152. https://doi.org/10.1007/978-3-030-11027-7_15.

Senger E, **Zhang M**, **van der Goot R and Plank B** (2024) *Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings. Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024). St. Julian's, Malta: Association for Computational Linguistics*, pp. 1–15.

Shi B, **Yang J**, **Guo F and He Q** (2020) Salience and market-aware skill extraction for job targeting. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, pp. 2871–2881. https://doi.org/10.1145/3394486.3403338.

Sillberg CV, **De Cerqueira JS**, **Sillberg P**, **Kemell K-K and Abrahamsson P** (2025) *The EU AI Act is a good start but falls short*. In PapatheocharousE, Farshidi S, Jansen S and Hyrynsalmi S (eds), *International Conference on Software Business (Cham)*. Cham: Springer Nature Switzerland, pp. 114–130. https://doi.org/10.1007/978-3-031-85849-9_10.

Tamburri DA, **Van Den Heuvel W-J and Garriga M** (2020) *Dataops for societal intelligence: A data pipeline for labor market skills extraction and matching. 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI). arXiv:2104.01966*, pp. 391–394.

Arnulf JK, **Tegner L and Larssen Ø** (2010) Impression making by résumé layout: Its impact on the probability of being shortlisted. *European Journal of Work and Organizational Psychology 19*(2), 221–230. https://doi.org/10.1080/13594320902903613.

Vásquez-Rodríguez L, **Audrin B**, **Michel S**, **Galli S**, **Rogenhofer J**, **Cusa JN and Van Der Plas L** (2024) *Hardware-effective approaches for skill extraction in job offers and resumes. RecSys in HR'24: The 4th Workshop on Recommender Systems for Human Resources, in conjunction with the 18th ACM Conference on Recommender Systems, October 14–18, 2024, Bari, Italy: CEUR Workshop Proceedings.*

Vásquez-Rodríguez L, **Audrin B**, **Michel S**, **Galli S**, **Rogenhofer J**, **Cusa JN and van der Plas L** (2025) Skill Extraction [dataset]. Idiap Research Institute. Zenodo. https://doi.org/10.34777/peay-xe10.

Zhang M, **Jensen K**, **Sonniks S and Plank B** (2022a) *SkillSpan: Hard and soft skill extraction from English job postings. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, US: Association for Computational Linguistics*, pp. 4962–4984.

Zhang M, **Jensen KN and Plank B** (2022b) *Kompetencer: Fine-grained skill classification in Danish job postings via distant supervision and transfer learning.* Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, pp. 436–447.

Zhang M, **van der Goot R and Plank B** (2023) ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain. In *Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers)*. Toronto, Canada: ACL, pp. 11871–11890.

Zhang M, **van der Goot R**, **Kan M-Y and Plank B** (2024) *NNOSE: Nearest neighbor occupational skill extraction. Proceedings of the 18th Conference of the European Chapter of the ACL (*Volume 1: *Long Papers). St. Julian's, Malta: ACL*, pp. 589–608.

Cite this article: Vásquez-Rodríguez L, Audrin B, Michel S, Galli S, Rogenhofer J, Negro Cusa J and van der Plas L (2026). Comparing AI-based approaches to candidate screening: recommendations for policy and regulation. *Data & Policy*, 8: e35. doi:10.1017/dap.2026.10076

https://doi.org/10.1017/dap.2026.10076 Published online by Cambridge University Press