Discover Artificial Intelligence

RESEARCH Open Access

Using language models to extract and map emerging skills to the ESCO taxonomy

Alessia De Santo[1,2*], Navid Nobani[1,2], Lorenzo Malandri[1,2], Fabio Mercorio[1,2] and Mario Mezzanzanica[1,2]

This paper is partially supported within the research activity of a grant entitled *"PILLARS – Pathways to Inclusive Labour Markets"* – under the call H-2020 TRANSFORMATIONS 18-2020, grant agreement NUMBER 101004703. https://www.h2020-pillars.eu/.

*Correspondence:
Alessia De Santo
alessia.desanto@unimib.it
[1]Dept of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, Italy
[2]CRISP Research Centre, University of Milano-Bicocca, Milano, Italy

Abstract
As the online labour market evolves, the ability to identify not only established skill requirements but also emerging skills directly from online job ads (OJAs) has become crucial for anticipating workforce needs and designing effective training programmes. To address this challenge, we introduce the `SkiLLink` pipeline, which is articulated into two main components: (i) an identification-and-filtering component that extracts new skills as they appear in OJAs using word embeddings and language models; and (ii) a mapping component that employs embedding-based recommendations and large language models to align the extracted skills with an established labour market taxonomy. This distinction is also reflected in our evaluation strategy. The extraction-and-filtering component is assessed qualitatively through validation by labour market experts across multiple EU countries, who reviewed the relevance and correctness of the emerging skills. The mapping component is evaluated quantitatively using two benchmark datasets and a baseline derived from the European Skills taxonomy (ESCO). We further perform a robustness analysis employing cosine similarity and twelve additional distance metrics to examine the stability of embedding-based alignment.

Keywords Large language models, Entity linking, Skill extraction, Skill mapping, Taxonomy enrichment

1 Introduction

The rapid expansion of the online labour market has resulted in an unprecedented volume of OJAs being published across digital platforms. This wealth of real-time labour market data offers valuable insights into workforce trends and the evolution of occupational demands. In particular, it enables the identification of not only existing skill requirements but also emerging skills, which are crucial for shaping education and training programs in line with evolving industry needs. Recognizing this potential, Eurostat[1] has collected over 450 million OJAs from 2019 to 2024 across 27+1 countries. The goal is to monitor labour market dynamics in near real-time using the multilingual ESCO

[1] https://ec.europa.eu/eurostat

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

taxonomy[2] - a comprehensive framework of skills, competences, qualifications, and occupations - as a lingua franca to enable cross-country comparability. While ESCO includes over 13,000 documented skills, its static nature limits the ability to keep pace with the dynamic emergence of new competencies reflected in the labour market. Recent strategic priorities of the European Commission highlight the urgency of addressing skill shortages, promoting upskilling and reskilling, and preparing the European workforce for rapid technological and economic transitions[3]. Against this backdrop, the ability to detect and systematically map novel skills from unstructured OJAs is increasingly essential to inform policy and support strategic workforce development. In this paper, we present `SkiLLink`, a multilingual pipeline designed to detect and normalize emerging skills from OJAs. Developed within the PILLARS project, `SkiLLink` has been applied to a large dataset comprising over 18 million job ads across 28 European countries and 23 languages. The pipeline is organised around two main components: (i) a skill-extraction component, consisting of candidate skill identification followed by filtering, and (ii) a skill-mapping component that aligns the validated candidates with ESCO through a recommendation-based process that leverages multiple large language models.

2 Contribution

This work makes a threefold contribution to the study of novel skill extraction and mapping from OJAs:

1. *Extraction of Novel Skills* We propose a methodology to automatically extract *novel and emerging skills* directly from unstructured text in OJAs using word embeddings and LLMs. We maintain novelty by retaining only candidate skills that are lexically distinct from ESCO entries, using fuzzy string matching to filter out close matches.
2. *Skill Normalisation and Mapping* To ensure consistency and enable structured analysis, we map the newly extracted skills onto the ESCO taxonomy, aligning them with standardised classifications used in labour market intelligence.[4] Our data, OJAs collected from web labour portals across 27 EU countries and the UK in 22 languages, contain skill descriptions written in unstructured natural language, making automatic linkage to ESCO a non-trivial challenge. Addressing this challenge is a key contribution of our work, with relevance to the semantic web and labour market analytics communities.
3. *Application on Large-scale real-world scenario* We implement the developed methodology in a real-world setting, applying it to a large-scale dataset of OJAs published in the 27 EU countries and UK in 22 languages.

Our contribution lies in the design of a scalable and structured methodology for emerging skill discovery in a multilingual, real-world setting. Rather than proposing a standalone algorithm, we develop an integrated framework that combines embedding-based retrieval, lexical filtering, LLM-based validation, and expert assessment to detect, normalise, and map novel skills from OJAs. This approach enables the systematic tracking

[2] ESCO (European Skills, Competences, Qualifications and Occupations) is the European multilingual classification of Skills, Competences and Occupations.

[3] https://commission.europa.eu/topics/eu-competitiveness/union-skills_en

[4] The ESCO taxonomy is structured into three pillars–skills, occupations, and qualifications–with the skill pillar encompassing knowledge, abilities, and attitudes, grouped by reuse level (specific, sector-specific, or transversal). Each skill is linked to its related occupations and qualifications.

of evolving skill requirements and supports downstream applications in labour market intelligence, including policy-making, employment services, and education planning.

3 Background and related work

This section addresses the theoretical underpinnings of the proposed framework, with a detailed examination of the concepts of skill extraction from online job advertisements and the normalization of these skills into a standardized taxonomy.

3.1 Skills extraction

Skill Extraction (SE) can be viewed as a specialised application of Information Extraction (IE) in the labour market domain. IE is the process of identifying and extracting structured information of predefined types from unstructured natural language texts [1].

Formally, let $\mathcal{T} = \{t_1, \ldots, t_n\}$ be the set of texts, $\mathcal{P} = \{p_1, \ldots, p_m\}$ the set of information types, and $\mathcal{A}$ the universe of atomic items. Formally, let $\mathcal{T} = \{t_1, \ldots, t_n\}$ be the set of texts, $\mathcal{P} = \{p_1, \ldots, p_m\}$ the set of information types, and $\mathcal{A}$ the universe of atomic items.

Let $\tau : \mathcal{A} \to \mathcal{P}$ assign each item its type.

$$\mathrm{IE} : \mathcal{T} \times \mathcal{P} \to 2^{\mathcal{A}}, \qquad \mathrm{IE}(t, p) = \{\, a \in \mathcal{A} \mid \mathrm{occ}(a, t) \ \wedge \ \tau(a) = p \,\}. \tag{1}$$

$$\mathrm{occ}(a, t) \overset{\text{def}}{\iff} \exists\, 1 \le i \le j \le |t| \text{ such that } t_i \ldots t_j = a.$$

For a fixed pair (t_i, p_j) the output is a finite set

$$\mathrm{IE}(t_i, p_j) = \{a_1, \ldots, a_{k_{ij}}\}. \tag{2}$$

Drawing from this formalisation, we define *Skill Extraction (SE)* as the specific IE task of identifying and extracting skill entities from unstructured textual data commonly found in the labour market–such as job advertisements, resumes, or job descriptions. Let $\mathcal{T}$ be the set of unstructured text documents (e.g., job ads, résumés), $\mathcal{P}$ the set of information types, and $\mathcal{S}$ the universe of canonical skill entities. For a document $t_i \in \mathcal{T}$ and a target type fixed to skills, $p_s = \textsc{skill} \in \mathcal{P}$, the goal is to return all skill entities present in t_i. We define the skill–extraction function

$$\mathrm{SE} : \ \mathcal{T} \times \{\textsc{skill}\} \to 2^{\mathcal{S}}, \tag{3}$$

which maps a document and the target information type SKILL to the set of recognised skill entities it contains.

For a document $t_i \in \mathcal{T}$, the extraction result is

$$\mathrm{SE}(t_i, p_s) \ = \ \left\{\, s \in \mathcal{S} \;\middle|\; s \preceq t_i \,\right\} \ \subseteq \ \mathcal{S}, \tag{4}$$

where $s \preceq t_i$ denotes that at least one span in t_i realises the skill entity s.

Early approaches to skill extraction relied on exact matching, combining manual annotation with semantic clustering (e.g., Word2Vec) and ontology construction to classify skills from job-related texts [2–4]. To handle variability in skill terminology, fuzzy matching techniques were introduced, comparing extracted phrases with controlled vocabularies like ESCO using similarity metrics such as Levenshtein Distance and Jaccard Similarity [5]. Unsupervised topic modelling, particularly Latent Dirichlet

Allocation (LDA), has been used to uncover latent skill structures by applying it directly to job descriptions or using a domain-specific vocabulary [6, 7]. Deep learning further advanced the field by treating skill extraction as sequence tagging or multi-label classification, using convolutional networks and ranking-based methods [8–10]. Recently, transformer-based models have shown promising results by leveraging contextual embeddings, typically by fine-tuning BERT or SpanBERT with classification layers (e.g., CRFs) or adapting pretrained models for domain-specific recruitment tasks [11–14]. To the best of our knowledge, no prior study has conducted skill extraction over a comparably diverse set of languages or systematically evaluated the novelty of the extracted skills. To fill these gaps, we employ a rigorous expert evaluation involving 28 labour market experts, each specialized in their national contexts. This ensures extracted skills are both contextually relevant and novel.

3.2 Skills normalisation

Skills Normalisation (SN) can be cast as a specialised instance of Semantic Retrieval (SR) in the labour-market setting. In SR, items are retrieved based on semantic similarity, typically by encoding text into a vector space.

Semantic Retrieval (SR). Let $\mathcal{Q} = \{q_1, \ldots, q_n\}$ be a set of queries and $\mathcal{E} = \{e_1, \ldots, e_m\}$ a set of target elements. Assume each item is represented by an embedding, and let $\mathrm{sim}(\cdot, \cdot)$ be a similarity function (e.g., cosine). For each $q_i \in \mathcal{Q}$, SR returns the most similar target:

$$\mathrm{SR}(q_i; \mathcal{E}) \;=\; \arg\max_{e \in \mathcal{E}} \mathrm{sim}(q_i, e). \tag{5}$$

Skills Normalisation (SN). Let $\mathcal{S} = \{s_1, \ldots, s_k\}$ be the set of skill mentions extracted from OJAs, and let $\mathcal{E}_{\mathrm{ESCO}}$ denote the set of canonical skills in ESCO (each with a preferred label), represented as embeddings. SN maps each extracted mention to its canonical ESCO entry:

$$\mathrm{SN}(s_i; \mathcal{E}_{\mathrm{ESCO}}) \;=\; \arg\max_{e \in \mathcal{E}_{\mathrm{ESCO}}} \mathrm{sim}(s_i, e) \;=\; \hat{s}_i \in \mathcal{E}_{\mathrm{ESCO}}. \tag{6}$$

Aggregating over all mentions gives the normalised set

$$\mathrm{SN}(\mathcal{S}; \mathcal{E}_{\mathrm{ESCO}}) \;=\; \{\hat{s}_1, \ldots, \hat{s}_k\} \;\subseteq\; \mathcal{E}_{\mathrm{ESCO}}, \tag{7}$$

where each $\hat{s}_i$ corresponds to the ESCO entry whose *preferred label* is the closest in meaning to the extracted mention s_i.

Few works address mapping skills to ESCO. We summarize them in Table 1.

An early pre-neural system for skill extraction and normalization is SKILL [15], which established foundational benchmarks before the advent of large language models. Skill-NER [25] extracts soft skills from texts but does not normalize them to ESCO entries. Kompetencer [16] categorizes skills into 23 ESCO-aligned groups without using the ESCO skills for the mapping. The same dataset is used to evaluate ESCOXLM-R [17], a pretrained language model for skills extraction and classification. SkillGPT [18] employs a language model for skill extraction and standardization but lacks empirical validation. Among more recent LLM-based approaches, ESCOX [20] is an open-source tool that jointly extracts and maps skills and occupations to ESCO and ISCO-08 entries in a single embedding-matching step via cosine similarity; unlike `SkiLLink`, it provides no

Table 1 Comparison of Skill Normalization Approaches

Framework	Mapping Approach	Lang.	Maps to ESCO Skills?	Used?	Reproducibility	Notes
SKILL [15]	Rule-based extraction & normalization	en	No (pre-ESCO)	No	Data: ✓ Code: ✓	Foundational pre-neural system.
Kompetenc-er [16]	Rule-based classification to 23 ESCO categories	en, da	No (Only categories)	No	Data: ✓ Code: ✓	Only high-level classes.
ESCOXLM-R [17]	Pretrained LM, data from [16]	en, fr, da, de	No (Same 23 categories)	No	Data: ✓ Code: ✓	Only high-level classes.
SkillGPT [18]	LLM: summarization + vector search	en, fr, nl	Yes	No	Data: ✗ Code: ✓	No validation.
Marconi et al. [19]	Matrix sampling + bootstrapping for classifier eval	en	Yes (evaluation only)	No	Data: ✗ Code: ✗	Evaluation methodology, not extraction.
ESCOX [20]	LLMs + text embeddings to ESCO	en	Yes	No	Data: ✗ Code: ✓	Open-source; No validation.
Becker et al. [21]	LLM categorization of job ads	de	No (DigComp 2.2)	No	Data: ✗ Code: ✗	Public sector; DigComp not ESCO.
Wang and Sun [22]	LLM-supervised multilingual sentence encoder	en, da	Yes (13,896 skills)	No	Data: ✗ Code: ✓	Known skills only. Sentence-level map.
Decorte et al. [23]	Skill extraction + ESCO mapping	en	Yes	Yes	Data: ✓ Code: ✓	Granular ESCO mapping.
Clavié et al. [24]	Zero-shot LLM to ESCO entries	en	Yes	Yes	Data: ✓ Code: ✓	Granular mapping (en only).

empirical validation of either extraction or mapping accuracy.Wang and Sun address the annotation bottleneck via scalable LLM supervision to train a lightweight multilingual sentence encoder [22]; unlike `SkiLLink`, which specifically targets skill expressions absent from ESCO to enrich the taxonomy, their pipeline assumes all skill mentions can be represented within the existing 13,896-skill ESCO taxonomy, covering both explicit and implicit mentions and representing them as per-ad skill distributions. LLM-based competence classification has also been applied to public-sector job postings, specifically German government advertisements [21]. The latter applies LLM-based competence classification to public-sector job advertisements, operationalizing the DigComp 2.2 framework rather than ESCO to categorize digital skills across nearly 6,000 German government job postings. Closest to our work are [23] and [24], which map extracted skills to ESCO; the latter adopts a two-step zero-shot pipeline. Marconi et al. [19] provides the first systematic evaluation of ESCO skill classifiers at scale, finding that despite acceptable accuracy, classifiers suffer from low recall, and that simple text matching remains competitive against LLM-based methods. In Section 5, we benchmark `SkiLLink` against [23] and [24] – the only two methods that share our task formulation of mapping extracted skills to granular ESCO entries and provide fully reproducible evaluation setups – and show that `SkiLLink` outperforms both.

4 Building SkiLLink

We introduce SkiLLink, a multilingual, multi-stage framework designed to extract and normalise novel skill expressions from OJAs. Figure 1 illustrates how the workflow implements the two components of SkiLLink: the first two operational steps collectively realise skill discovery, while the third step performs skill mapping. The following section provides a detailed description of each stage, outlining the input and output of every step to offer a clear representation of the pipeline.

4.1 Step 1: Identification of candidate skills

The first phase of the pipeline identifies potential new skills from a large, multilingual corpus of Online Job Advertisements (OJAs). This step establishes the data foundation, trains country-specific word embeddings, and uses these embeddings to surface semantically related terms that may represent emerging skills.

Data

Our analysis relies on two primary data sources: (i) large-scale online job advertisements (OJAs) and (ii) the ESCO skill taxonomy, version 1.1.1. ESCO provides a comprehensive, multilingual classification of skills and competences, including both preferred and alternative labels, and serves as the European Commission's reference vocabulary

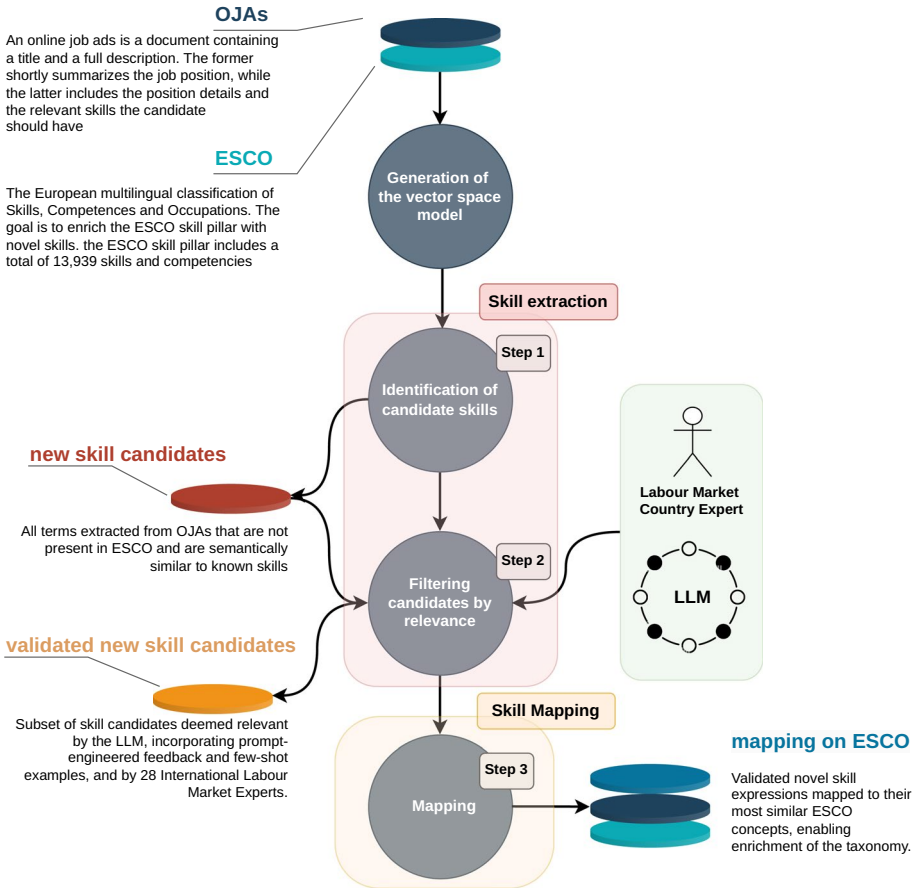


Fig. 1 Overview of the SkiLLink framework for extracting and mapping novel skill expressions from OJAs across 28 countries

for labour-market and education data.[5] We use ESCO as the semantic anchor against which newly emerging or domain-specific skill expressions from OJAs can be identified and assessed. OJAs are collected through the WIH[6] , part of Eurostat's Trusted Smart Statistics (TSS) initiative. From the full WIH-OJA dataset of over 450 million postings, we use the *NLP sample v3* (release r20240226), which covers the years 2018–2023 and includes standardized fields for text mining. For each country, we select the most recent one million advertisements where available, ensuring a focus on present-day labour-market dynamics and reducing temporal heterogeneity.

This selection results in a final multilingual dataset of roughly 18 million OJAs across EU and EFTA countries.[7] Table 2 summarizes the number of postings per country.

Embedding-based discovery of emerging skills

The goal of this step is to detect terms that are semantically close to known ESCO skills but not yet represented in the taxonomy. The input consists of raw OJA text (see example below), and the output is a list of candidate novel skills.

Input: OJAs corpus from EU countries

Output: Extracted Candidates Novel Skills

Example

Input:

> **Extract from an OJA**
> We're seeking a **detail-oriented** Data Scientist with strong quantitative skills and a knack for uncovering insights from complex datasets. You'll play a pivotal role in **analysing large volumes of data** using **Python**, **R**, or **Java**, alongside a deep understanding of machine learning algorithms. Your expertise in **SQL** and **data visualisation tools** will be crucial in communicating findings effectively to diverse stakeholders. If you're passionate about transforming data into actionable insights and **contributing to strategic decision-making**, we'd love to have you join our team.

Output:

> **Candidate Novel Skills** = {attention to detail, mathematical aptitude, quantitative reasoning, statistical analysis, Python, R, Java, machine learning, SQL, data visualisation, prompt engineering, ...}

Method We collect OJA texts from each of the 28 countries and train country-specific word embeddings. The OJA descriptions undergo standard preprocessing for feature normalisation and noise reduction, including lowercasing, removal of stopwords and punctuation, and the creation of word n-grams. For each country, word embeddings are then generated from the preprocessed corpus using FastText [26], which we adopt

Table 2 Total OJA counts by country (capped at 1M)

Country															
Total OJAs	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	792,032	888,423	
Country															
Total OJAs	651,722	371,652	496,823	631,463	665,033	317,259	285,274	289,078	174,351	198,966	139,813	120,007	113,436	59,473	

[5] https://esco.ec.europa.eu/en/classification/skill

[6] https://cros.ec.europa.eu/wih

[7] We access the WIH-OJA *NLP* dataset through the "PILLARS" project consortium. The data can also be requested from EUROSTAT, subject to approval and an NDA.

exclusively given its strong and cross-country consistent performance in previous evaluations [27, 28]. To implement the identification step of the `SkiLLink` framework, we perform an extensive grid search across 160 FastText configurations, varying the training algorithm {`SG`, `CBOW`}, embedding sizes {5, 20, 50, 100, 300}, number of epochs {10, 25, 100, 200}, and learning rates {0.01, 0.05, 0.1, 0.2}.

To select the best embedding model for each country, we use the Hierarchical Semantic Similarity (HSS) metric [29], which evaluates how well the semantic relationships encoded in the ESCO skill taxonomy are preserved in the learned embedding space. Across countries, the best-performing configuration is consistently: `model = SG`, `dim = 100`, `epoch = 5`, `lr = 0.01`.

Using this optimal model, we take the complete list of ESCO skills (approximately 14,000 preferred and alternative labels) and, for each skill s, identify the top-5 most similar terms in the embedding space based on cosine similarity. To prioritise high-precision candidates, the retrieval step focuses on the most similar neighbouring terms in the embedding space, while subsequent filtering stages remove candidates that correspond to existing ESCO skills or that do not represent meaningful emerging skills. This threshold offers a balance between semantic relevance and redundancy, as larger k values introduce repetitive variants of common skill expressions. The retrieved terms constitute an initial pool of potential new skill candidates, expressions that are semantically aligned with existing ESCO skills yet distinct in lexical form or nuance. Semantic retrieval is intentional: since the pipeline's goal is to enrich ESCO, candidates must remain grounded in its skill space. However, semantic proximity alone does not guarantee lexical novelty – a candidate may be close to an existing ESCO skill in meaning yet expressed in genuinely new terminology. To ensure this lexical novelty, each candidate is compared syntactically to all ESCO labels. We apply a lexical filtering step using the `fuzz.ratio` function from the `rapidfuzz` library, retaining only expressions with a similarity score ≤ 70. The ratio is computed using normalized Indel similarity.[8] This constraint ensures that the extracted candidates remain lexically distinct from ESCO concepts while preserving meaningful semantic proximity. This design intentionally captures novelty in a broad sense, retaining both entirely new skill concepts and new surface expressions of skills only implicitly covered by ESCO. A finer-grained distinction between these two cases is left to subsequent work.

4.2 Step 2: Filtering candidates by relevance

Input: List of candidate novel skills.

> **Candidate Novel Skills** = {attention to detail, mathematical aptitude, quantitative reasoning and problem-solving skills, expertise in statistical analysis, Python, R, Java, machine learning, SQL, data visualisation, project manager, interaction with security personnel, programming, prompt engineering}

Output: List of *relevant* candidate novel skills.

[8] Given two strings w_1 and w_2: $Indel = 1 - \frac{\text{Levenshtein}(w_1, w_2)}{\text{len}(w_1)+\text{len}(w_2)}$.

Relevant Candidate Skills = {project manager, interaction with security personnel, statistical analysis, programming, prompt engineering}

In this step, we leverage the ESCO skills as semantic anchors to discover additional, potentially relevant skill expressions in the OJA corpus. The goal is to identify terms that are both semantically similar to existing ESCO skills and frequently used in real job postings. To achieve this, we adopt the weighted relevance metric proposed by [27], which combines cosine similarity in the embedding space with corpus frequency. For each ESCO skill *s* and candidate term *c*, we compute:

$$S(c, s) = \alpha \cdot \cos_\mathrm{sim}(c, s) + (1 - \alpha) \cdot \mathrm{freq}(c),$$

where $\cos_\mathrm{sim}(c, s)$ denotes the cosine similarity derived from the country-specific FastText embeddings (Section 4), and $\mathrm{freq}(c)$ represents the relative frequency of *c* in the national OJA corpus. Following [27], we set $\alpha = 0.85$, giving priority to semantic proximity while still accounting for practical relevance in the labour market. The score *S*(*c*, *s*) is then used to rank candidate terms with respect to each ESCO skill, retaining the top-ranked neighbours as the pool of potentially relevant expressions. We further remove infrequent or spurious terms through a cumulative-frequency filter: for each skill, candidates are ranked by $\mathrm{freq}(c)$, and only those contributing to the top 95% of cumulative frequency among the retained neighbours are kept. This dual filtering process guarantees that selected terms are (i) semantically grounded in the ESCO taxonomy, (ii) commonly used in OJAs, and (iii) likely to represent practically relevant skills rather than noise. Further details on the filtering methodology, including a sensitivity analysis of the lexical threshold and the relevance weight α, are provided in Appendix A.

Candidates validation

The candidate validation stage addresses a different objective than the lexical filtering step applied during candidate extraction. While the extraction phase removes expressions that correspond to existing ESCO skills or close lexical variants, the validation phase evaluates whether the remaining candidates represent semantically coherent and meaningful skill concepts in the context of labour market demand.

We perform a two-part validation process to ensure the quality and relevance of the extracted novel skill candidates before proceeding to their mapping onto ESCO.

Input: List of relevant candidate novel skills.

Relevant Candidate Skills = {project manager, interaction with security personnel, statistical analysis, programming, prompt engineering}

Output: Validated list of candidate novel skills.

Automated Validation via LLM

To assess the plausibility of new skill candidates, we first employ an LLM validation step based on GPT-4. The complete prompt template adopted for this validation step is provided in Appendix B for reproducibility. The prompt is instantiated for each candidate term together with an example of its usage in a job advertisement. When using LLMs for classification-like tasks, carefully designed prompts can significantly enhance

the consistency and reliability of the responses. In this stage, we apply the following strategies:

- *Contextual framing* We provide rich contextual information within the prompt to improve the model's comprehension. This includes:
  - A definition of the concept of *skill*, aligned with the one used in ESCO and the broader literature.
  - Instructions on the expected format of the response, explicitly asking the model to begin with "yes" or "no" to ensure machine-readable outputs.
  - A representative example of an OJA that contains the candidate skill term in a natural context, helping the model assess whether the term functions as a skill in real usage.
  - Contextual grounding in the linguistic and institutional setting of the job advertisement, as both the candidate term and the example OJA are drawn from the same country. This allows the model to interpret the term within the appropriate local terminology, accounting for country-specific vocabulary, conventions, and nuances in skill expression.
- *Few-shot learning* We prepend the actual query with a curated set of example prompts and their corresponding responses. Each example includes five candidate terms along with their usage in OJAs. The model is asked to determine whether each term constitutes a skill, beginning its answer with "yes" or "no" and justifying the choice with a short explanation. This calibration helps guide the model toward more accurate and coherent responses. The model outputs are interpreted by considering the leading "yes/no" token as the classification decision, while the accompanying explanation is used for qualitative inspection and is also provided to domain experts in the subsequent validation stage. Candidates classified as "yes" are retained for further review, whereas those classified as "no" are discarded.

This LLM-based validation serves as a crucial filtering step before expert evaluation. The use of LLMs was necessary due to the large volume of candidate terms (approximately 1,000 per country) which made manual expert validation of all entries impractical. To address this, we designed the prompt to be deliberately conservative: in cases of uncertainty, the model was instructed to flag terms as potentially relevant rather than discard them. This conservative behaviour is reflected in the overall expert-validated precision of around 70%, indicating that we prioritized high recall over precision to ensure that relevant terms were retained. Although a full quantitative evaluation of LLM filtering is out of scope, we conducted random checks on filtered-out terms for two markets (Italy and the UK), and these showed that the recall of valid skills was above 98%, confirming that very few relevant candidates were erroneously excluded. This two-stage design ensures that the LLM acts as a high-recall pre-filter, while final validation is delegated to domain experts. Detailed country-level statistics on the proportion of candidates discarded during the LLM filtering step are reported in Appendix D.

Manual Validation by Experts

In the second validation stage, the filtered list of candidate skills is reviewed by labour market experts from the 27+1 EU countries. Experts were recruited through the

European Network of Regional Labour Market Monitoring[9], a community of practitioners specialised in national and regional labour market analysis. Each expert assessed the plausibility, contextual relevance, and linguistic appropriateness of the candidate expressions with respect to real job descriptions in their native language. Experts performed a binary validation task, deciding whether each candidate could be considered a valid skill in the context of national job advertisements. In addition, they were asked to provide a short motivation for their decision, supporting qualitative analysis of disagreements with the LLM-based filtering. Table 3 reports the number of candidate skills evaluated per country.

The validation was conducted by one expert per country, each evaluating the candidate skills extracted from job advertisements related to their national labour market. These experts are official specialists collaborating with the ESCO initiative and involved in ESCO-related activities, which ensures strong domain knowledge of the taxonomy and the evolving skill landscape. Because each expert assessed country-specific candidate skills, the annotated sets were non-overlapping across experts, making standard inter-annotator agreement measures such as Cohen's Kappa not directly applicable in this setting. The goal of the validation phase was therefore not to measure annotation consistency across annotators, but rather to provide domain-level validation of the plausibility and relevance of the extracted skills. To nonetheless provide empirical evidence of the reliability of the annotation task, we conducted a double-annotation study on a 100-item subset for two countries (UK and IT), in which a second independent expert evaluated the same candidate skills. The resulting Cohen's κ values of 0.81 and 0.71 on the inclusion decision indicate substantial to almost perfect agreement, suggesting that the task is consistently interpretable across judges and that individual annotator bias is unlikely to be the dominant driver of the observed cross-country variation in validation rates.

Overall, experts validated 3,552 out of the 4,941 proposed candidates (71.9%), although the proportion varies substantially across countries due to differences in the linguistic characteristics of the OJAs, the quality of country-specific embeddings, the amount of available data, and subjective expert judgments (see Fig. 2). This cross-country variation can be attributed to several factors. First, differences in linguistic resources and model coverage may affect the quality of text representations and translations across languages. Although we adopt a translation-based strategy to obtain more stable embeddings, languages with fewer available resources or less representation in pretraining data may still lead to worse text representations and translation. Second, the amount of available data varies significantly across countries. For countries with a smaller number of OJAs, the embedding space is less well-defined, which can result in a higher proportion of noisy or less relevant candidate terms. Finally, expert judgments may reflect genuine differences

Table 3 Number of new skill candidates evaluated by each expert per country

Country														
Num. candidates	400	395	350	316	292	287	279	242	241	229	224	215	209	207
Country														
Num. candidates	191	163	145	142	130	107	90	89	76	62	45	33	30	19

[9] https://www.regionallabourmarketmonitoring.net

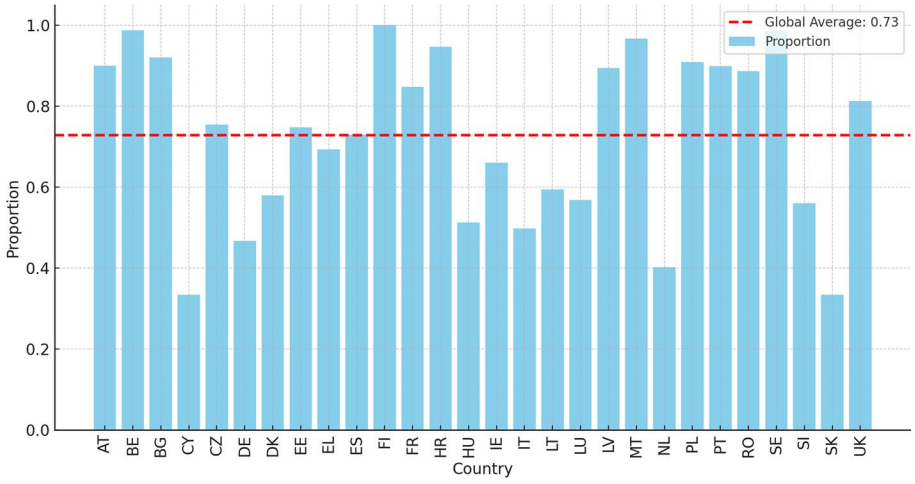


Fig. 2 Proportion of candidate skills evaluated as novel and valid by experts

Table 4 An example of candidate novel skill validation

Extracted Skill	LLM Validation	LLM Motivation	Experts Validation	Experts Motivation
Project manager	×	No, it is an occupation	*null*	*null*
Interaction with security personnel	✓	Yes, because it requires effective communication	×	this is 'experience' not skills
Statistical analysis	✓	Yes, it is a core methodological competence	✓	
Programming	✓	Yes, it is a specialized skill.	×	too broad... programming activities, computer programming, etc.
Prompt engineering	✓	Yes, it is a learned competence in AI.	✓	

in national labour market structures and terminology, as certain skill expressions may be considered more or less relevant depending on local practices and institutional contexts. Taken together, these factors contribute to the observed variability in validation rates across countries.

It is important to note that DE, CY, and NL did not consider ESCO knowledge concepts during validation, which affects their final scores.

Table 4 illustrates an example of this two-step validation pipeline: an initial LLM-based pre-validation followed by expert review. The table shows how certain terms may be accepted or rejected depending on their conceptual specificity, their semantic type, such as skill, occupation, or experience, and their contextual relevance.

4.3 Step 3: Mapping

This phase aims to map the extracted new skill expressions to their most appropriate position within the ESCO taxonomy, ensuring their standardisation and enabling the potential enrichment of the taxonomy with relevant, emerging skills.

Sentence embedding mapping to ESCO via LMs

Input: Validated list of candidate novel skills.

Output: List of n recommendations of mapping for each candidate.

We map novel skill candidates to the ESCO taxonomy through a two-stage process combining embedding-based retrieval and LLM-based refinement.

First, we compute semantic similarity between each candidate skill and ESCO entries using sentence embeddings. Multiple models, including Sentence-BERT and multilingual transformer encoders, were explored. For each candidate, we retrieve the top-3 most similar ESCO skills based on cosine similarity.

Second, we apply an LLM-based refinement step to disambiguate among these top candidates. In this stage, the model is provided with the candidate skill together with the three most similar ESCO preferred labels and their corresponding descriptions, and is required to select the most appropriate match among the three options only. This constrained selection preserves the high-recall properties of embedding-based retrieval while enabling a more precise semantic alignment through contextual reasoning. The complete prompt template used in this stage is reported in Appendix C. To evaluate the effectiveness of this mapping strategy, we construct a multilingual benchmark based on ESCO data. Specifically, we sample 300 ESCO skill labels (from the 13,939 corpus of preferred terms) and their associated alternative labels for all 22 languages represented in the new skills corpus[10]. The task is to correctly match each alternative label to its corresponding preferred label. This benchmark serves as a proxy to evaluate the ability of our method to correctly place out-of-taxonomy terms in their most semantically appropriate location within the existing skill hierarchy. Overall, this hybrid approach combines the strengths of embedding-based similarity (high recall) with LLM-based reasoning (improved precision), particularly in cases where multiple ESCO entries exhibit comparable semantic proximity.

Implementation

Given the cross-lingual nature of our corpus, we experimented with two alternatives for obtaining sentence embeddings: (i) directly using multilingual embedding models, and (ii) translating all texts into English before encoding. The latter strategy yielded more stable and higher-quality representations. Translations were produced using `DeepSeek v3`.

For embedding-based skill normalisation, we selected models using guidance from the MTEB leaderboard [30]. We excluded commercial or restricted-weight models and additionally incorporated three widely used open alternatives commonly adopted in NLP libraries[11]. In total, we evaluated a diverse pool of open-weight models spanning different architectures and embedding sizes.[12] To assess robustness, we computed similarity using not only the well-known cosine similarity but also twelve additional distance metrics, including Bray–Curtis, Canberra, Chebyshev, Dice (continuous), dot product, Euclidean, Jaccard (continuous), Jensen–Shannon divergence, Kullback–Leibler divergence, Manhattan, Pearson and Spearman.

While cosine similarity is commonly used for embedding comparison, we employ Spearman correlation to capture rank-based relationships between vector components. Although cosine similarity is invariant to global rescaling, Spearman correlation offers a stronger invariance: it depends only on the rank order of components and is therefore

[10] In ESCO, alternative labels can include synonyms (words with similar or identical meanings), spelling variants, declensions, and abbreviations. These labels help connect ESCO concepts to the real labour market by providing alternative ways of referring to them.

[11] all-MiniLM-L6-v2, all-mpnet-base-v2, and gte-large-1.5b.

[12] Evaluated models: thenlper/gte-large; Qwen/Qwen3-Embedding-0.6B, 4B, 8B; Alibaba-NLP/gte-Qwen2-1.5B-instruct, 7B-instruct; google/embeddinggemma-300m; infly/inf-retriever-v1, v1-1.5b; sentence-transformers/all-MiniLM-L6-v2; sentence-transformers/all-mpnet-base-v2; plus the three additional models noted above.

Table 5 Best similarity metric per embedding model, ranked by mean score across all 22 languages. The selected model–metric combination is highlighted in textbi

Model	Best Metric	Mean Score
Qwen3-Emb-8B	spearman	0.7513
Qwen3-Emb-4B	spearman	0.7446
GTE-large	kl divergence	0.7401
Emb-Gemma-300M	spearman	0.7285
all-mpnet-v2	js divergence	0.7172
Qwen3-Emb-0.6B	braycurtis	0.6945
all-MiniLM-L6-v2	cos	0.6942
inf-retriever-v1	braycurtis	0.6914
inf-retriever-v1-1.5B	spearman	0.6772
GTE-Qwen2-1.5B	spearman	0.5374
GTE-Qwen2-7B	canberra	0.5101

Table 6 Spearman similarity for `Qwen/Qwen3-Embedding-8B` across all languages

Metric	Mean	bg	cs	da	de	el	en	es	et	fi	fr	hr
Spearman Sim	0.751	0.744	0.801	0.748	0.808	0.696	0.890	0.803	0.569	0.748	0.801	0.847
Metric	**hu**	**it**	**lt**	**lv**	**nl**	**pl**	**pt**	**ro**	**sk**	**sl**	**sv**	
Spearman Sim	0.750	0.772	0.718	0.720	0.790	0.704	0.769	0.577	0.675	0.789	0.809	

unaffected by any monotonic transformation of their values. This makes the similarity measure robust to heterogeneous per-dimension scales and to heavy-tailed components that would otherwise dominate cosine-based comparisons, shifting the focus to the relative ordering of dimensions rather than their absolute magnitudes.

For each model–metric pair, we calculated the similarity values across all languages in our dataset. This broader evaluation framework enabled us to observe how different metrics behave under varying linguistic conditions and to identify, in an empirical and model-agnostic manner, the embedding–metric combinations that consistently provide stable and interpretable results. Table 5 summarises, for each model, the best-performing metric and the corresponding mean score across all 22 languages. `Qwen3-Embedding-8B` with Spearman similarity achieves the highest mean score (0.751) and is therefore selected. Table 6 reports the per-language breakdown for this combination. Having identified the optimal model-metric pair, we proceed with the mapping to ESCO. The complete results for all model- metric combinations are provided in Table S1 of the Supplementary Material.

A clear cross-lingual pattern emerges: while Qwen3-Embedding-8B maintains strong overall consistency (mean 0.751), performance varies notably across languages, with English achieving the highest correlation (0.890), followed by German, French, and Croatian, whereas smaller or morphologically complex languages such as Estonian (0.569) and Romanian (0.577) show lower alignment. This suggests that the model's pretraining distribution and linguistic coverage still influence embedding quality despite its multilingual design.

5 Evaluation

In this section, we present our evaluation framework, designed to assess the ability of the proposed model to map extracted skills onto the ESCO taxonomy. After identifying the optimal embedding configuration and relevance metric, as described in Section 4, we compare our pipeline against two existing approaches.

Table 7 Comparison of skill-mapping performance on the *tech* and *house* datasets. Best performances shown in bold.

Method	RP@1	RP@5	RP@10	MRR
Tech				
Decorte et al. [23]	n/a	0.3171	0.3919	0.339
Clavié and Soulié [24]	0.465	0.6150	0.6890	0.537
`SkiLLink` (ours)	**0.608**	**0.869**	**0.895**	**0.707**
House				
Decorte et al. [23]	n/a	0.3080	0.3870	0.299
Clavié and Soulié [24]	0.630	0.5670	0.6100	0.507
`SkiLLink` (ours)	**0.662**	**0.875**	**0.919**	**0.755**

Notes: RP@k = Recall@k; MRR = Mean Reciprocal Rank. Values are reported as fractions (e.g., 0.608 = 60.8%)

Table 8 Precision@1 for prompting-based best-match selection on *tech* and *house* datasets

Model	P@1 Tech	P@1 House	Input Cost	Output Cost	Latency /TPS
No LLM (baseline)	60.8	66.2	–	–	–
DeepSeek	62.1	69.1	0.27	1.00	1.48 / 32.8
Qwen3-235B	62.8	69.1	0.09	0.57	0.66 / 26.6
Claude 3.5 Sonnet	64.7	70.6	3.00	15.00	1.35 / 48.1
GPT-5-Mini	67.6	69.3	0.25	0.92	6.11 / 43.3
Anthropic/claude-opus-4.5	64.7	69.9	5.00	25.00	2.79 / 63.23

Input/Output costs in USD per 1 million tokens. Latency in seconds, throughput in transactions per second (TPS)

Comparison against state-of-the-art

We compared our approach with the methods in [23] and [24], both of which focus on the English language, the only language analysed in their studies. Specifically, we replicated their task of mapping soft skills to the ESCO taxonomy, rather than classifying skills in job posts, which falls outside the scope of this paper, as our focus is solely on novel skills. Table 7 presents the comparison results for English. The authors evaluated their approaches on two datasets: 'house' and 'tech.' Testing the same datasets, our method outperforms theirs in terms of RP@1, RP@5, RP@10, and MRR, the key metrics used in their studies.

Beyond the embedding-based comparison with existing methods, we further explored whether a final LLM-mediated mapping step could yield additional improvements. After obtaining the top-3 most semantically similar ESCO candidates using our best-performing embedding configuration, we passed these candidates to a large language model, which selected the most appropriate match through a structured prompt containing task instructions and few-shot examples (see the prompt in Appendix C).

This hybrid procedure leverages both the semantic accuracy of dense embeddings and the reasoning capabilities of LLMs, enabling finer-grained disambiguation for ambiguous or context-dependent skills. To assess its robustness, we evaluated several LLMs: DeepSeek, Qwen3-235B, Claude 3.5 Sonnet, and OpenAI GPT-5-Mini, across the same SOTA datasets. As reported in Table 8, the prompting-based reranking is stable across architectures and computational budgets, consistently improving Precision@1 over the embedding-only baseline. These results demonstrate that adding a lightweight LLM refinement stage provides measurable gains while remaining compatible with scalable LMI deployment settings.

Performance correlates strongly with model class and cost. The open-access efficient models, DeepSeek and Qwen3, perform almost identically, offering a cost-effective

baseline. Meanwhile, the larger proprietary models (Claude Sonnet/Opus, GPT-5-Mini) provide a clear step up in capability, with GPT-5-Mini achieving the highest P@1.

Finally, we note that the evaluation is more extensive for the mapping component, as it allows for direct comparison with a structured taxonomy. In contrast, the extraction and filtering stages involve identifying potentially novel skill expressions, for which ground truth labels are inherently limited, making standard quantitative evaluation less straightforward.

6 Conclusion

Our empirical evaluation shows that the extracted skills exhibit strong alignment with human judgement: more than 70% of candidates are recognised as valid by national labour-market experts. In the mapping task, `SkiLLink` consistently outperforms existing baselines across all evaluation metrics, demonstrating the effectiveness of combining dense semantic representations with LLM-based reasoning. At the same time, the results reveal meaningful cross-country variation, particularly in low-resource or morphologically complex languages, suggesting that further advances in cross-lingual modelling remain essential. Moreover, instead of relying on cosine similarity as a default choice, we systematically evaluated 13 different similarity metrics. This comprehensive comparison enabled us to identify the most suitable metric for the ESCO mapping task and avoid arbitrary assumptions about similarity behaviour across languages. Beyond its technical contributions, `SkiLLink` has significant potential for policy applications. The structured, ESCO-aligned outputs it produces can be directly integrated into interactive labour market intelligence tools enabling policymakers to monitor skill demand dynamics across sectors, regions, and countries in near real-time. More concretely, the emerging skills identified by `SkiLLink` can inform targeted policy interventions: by detecting skill gaps between workforce supply and employer demand, the framework can support the design of upskilling and reskilling programmes, guide education-employment alignment initiatives, and help statistical agencies keep the ESCO taxonomy current with labour market evolution. A further avenue for future work concerns the deterministic nature of the current mapping mechanism. At present, `SkiLLink` returns the most similar ESCO entry based on embedding similarity, without quantifying the confidence of each match. Incorporating uncertainty estimates – for instance through ensemble-based similarity scoring or confidence calibration – would allow decision-makers to distinguish high-confidence mappings from ambiguous ones. This is particularly relevant when the framework is used to inform labour market policy, where the reliability of skill signals directly affects the quality of interventions. Looking ahead, we plan to extend `SkiLLink` in several directions. First, we aim to incorporate iterative feedback loops that allow experts and statistical agencies to feed newly validated skills back into the taxonomy. Moreover, the single-annotator design, while necessary given the country-specific nature of the task, means that individual differences in annotation style cannot be fully disentangled from the observed cross-country variation in validation rates; this should be considered in future work. Second, we will explore domain-adaptive fine-tuning, multilingual prompt optimisation, and improved alignment of sentence embeddings to enhance robustness in challenging linguistic settings. Finally, incorporating techniques that analyse how skill usage evolves over time would allow us to identify

emerging or fast-growing skills as soon as they begin to appear in job postings, further supporting evidence-based labour-market intelligence and taxonomy maintenance.

Details of filtering thresholds

This appendix provides further details on the two main hyperparameters governing the candidate filtering stage: the lexical similarity threshold of 70 against ESCO entries, and the relevance weight $\alpha = 0.85$ in the combined cosine–frequency score. For each, we present both the conceptual rationale and the empirical evidence supporting the chosen value.

Lexical similarity threshold

The lexical threshold of 70 against ESCO terms was selected to filter out candidates that constitute mere repetitions or superficial variants of existing taxonomy entries — for example, candidates such as *project managing* with respect to the ESCO entry *project management*, or *data analyst skills* with respect to *data analysis*, which share substantial lexical material with existing entries but do not necessarily contribute genuine novelty. Setting the threshold at this level helps ensure that retained candidates represent genuinely novel lexical contributions rather than redundant restatements of existing ESCO skills.

To validate this choice, we conducted a brief sensitivity analysis by varying the cutoff and examining the resulting candidate pool. Lowering the value produced a larger pool but also admitted more candidates differing from existing ESCO entries only through inflectional, morphological, or minor syntactic variation. Raising it made the filter more restrictive and reduced redundancy, but at the cost of excluding candidates that, while lexically close to ESCO terms, still expressed potentially meaningful specialisations or emerging skill formulations. To quantify this trade-off, we relied on an external LLM judge to estimate, for each similarity band, how many candidates would denote the same skill as their most similar ESCO entry — i.e., how many redundant entries would be admitted at different cutoffs. The results (Table 9) show that the proportion of candidates judged equivalent to an existing ESCO entry crosses 50% precisely at the 70–75 band, confirming 70 as the natural inflection point between novelty and redundancy.

Table 9 LLM-judged equivalence between candidates and their most similar ESCO entry, by lexical-similarity band

Band	Sample size	Same skill	Different skill	% Same
60–65	100	31	69	31.0
65–70	100	40	60	40.0
70–75	100	49	51	51.0
75–80	100	55	45	55.0
80–85	100	66	34	66.0

Relevance weight α

The thresholds for cosine similarity and frequency are adopted from [27], whose methodological setting is closely aligned with ours: both works address the extraction of novel candidates from a structured labour-market taxonomy — new occupations in their case, and new skills in ours. This parallel justifies the transfer of their calibrated values, since the underlying statistical regime is comparable: emerging terms are extracted from a domain-specific labour-market corpus and filtered against a reference taxonomy.

Within this scheme, we assign a high relevance weight, $\alpha = 0.85$, to cosine similarity. High-frequency terms in job advertisements often correspond to generic vocabulary, boilerplate expressions, or widely shared soft attributes — for instance *team*, *experience*, or *motivated* — that, despite their statistical prominence, carry limited discriminative value as skill candidates. Cosine similarity, by contrast, captures semantic proximity to the skill domain and is therefore a more reliable indicator of candidate relevance. This consideration is particularly important in our multilingual setting, where raw frequency distributions vary substantially across languages and corpora, whereas embedding-based similarity provides a more comparable signal across linguistic varieties.

We further validated this choice by varying α and inspecting the resulting candidate pools. Lower values increased the influence of frequency and admitted more borderline candidates, while higher values made filtering more conservative but reduced the diversity of retained candidates. The chosen value of 0.85 provided the most reasonable qualitative balance, preserving meaningful candidate–neighbour relations with clear topical, lexical, or contextual links while reducing semantically weak or noisy candidates before LLM validation. Together, the lexical threshold of 70 and $\alpha = 0.85$ act as conservative pre-validation filters, reducing the burden on the LLM stage without making the candidate pool overly sparse.

Prompt for LLM-based skill validation

This appendix reports the prompt template used for the automated validation of candidate skills via GPT-4.

LLM Prompt Template

System:
You are an expert in labour market, trying to detect new skills in job ads.
Skills can be defined as: "the ability to apply knowledge and use know-how to complete tasks and solve problems. They can be described as cognitive (involving the use of logical, intuitive and creative thinking) or practical (involving manual dexterity and the use of methods, materials, tools and instruments)."
Candidate new skills have been extracted from job descriptions using automated methods, therefore they could contain typos, they could be words that are not skills, like occupations, requirements, description of the company or the benefits, etc. Your role is to find what can be effectively considered a required skill (even if badly written). Remember that occupations are not skills. Moreover too generic words, like just "installing" or "drawings", should not be considered a skill because they are too generic.

Few-shot examples:
Example 1
User: Do you think the term <CANDIDATE_1> can be considered a skill? Answer starting with yes or no.
Context: <OJA_EXCERPT_1>
Assistant: Yes/No + short justification.

Example 2
User: Do you think the term <CANDIDATE_2> can be considered a skill? Answer starting with yes or no.
Context: <OJA_EXCERPT_2>
Assistant: Yes/No + short justification.

Example 3
User: Do you think the term <CANDIDATE_3> can be considered a skill? Answer starting with yes or no.
Context: <OJA_EXCERPT_3>
Assistant: Yes/No + short justification.

Example 4
User: Do you think the term <CANDIDATE_4> can be considered a skill? Answer starting with yes or no.
Context: <OJA_EXCERPT_4>
Assistant: Yes/No + short justification.

Example 5
User: Do you think the term <CANDIDATE_5> can be considered a skill? Answer starting with yes or no.
Context: <OJA_EXCERPT_5>
Assistant: Yes/No + short justification.

Final query:
User: Do you think the term <CANDIDATE_TERM> can be considered a skill? Answer starting with yes or no.
Context: <OJA_EXCERPT>

The prompt is instantiated by replacing each placeholder with the corresponding candidate term and its associated job advertisement excerpt.

Prompt for LLM-based ESCO mapping refinement

This appendix reports the prompt template used for the LLM-based refinement step applied to select the most appropriate ESCO mapping among the top-3 candidates retrieved via sentence embeddings.

LLM Prompt Template

System:
You are a careful assistant that follows instructions exactly.

User:
I need to select the correct match for the following skill: `<CANDIDATE_TERM>`. The correct match should be the most similar term.
Now I will give you three possible terms and their descriptions:

- term 1: `<TERM_1>`
  description 1: `<DESCRIPTION_1>`
- term 2: `<TERM_2>`
  description 2: `<DESCRIPTION_2>`
- term 3: `<TERM_3>`
  description 3: `<DESCRIPTION_3>`

Tell me which is the most similar term to the candidate. Give me only the term without comments. Always give me an answer among the three. If you think none match, try to guess anyway. Never give me a different choice from one of the three options.

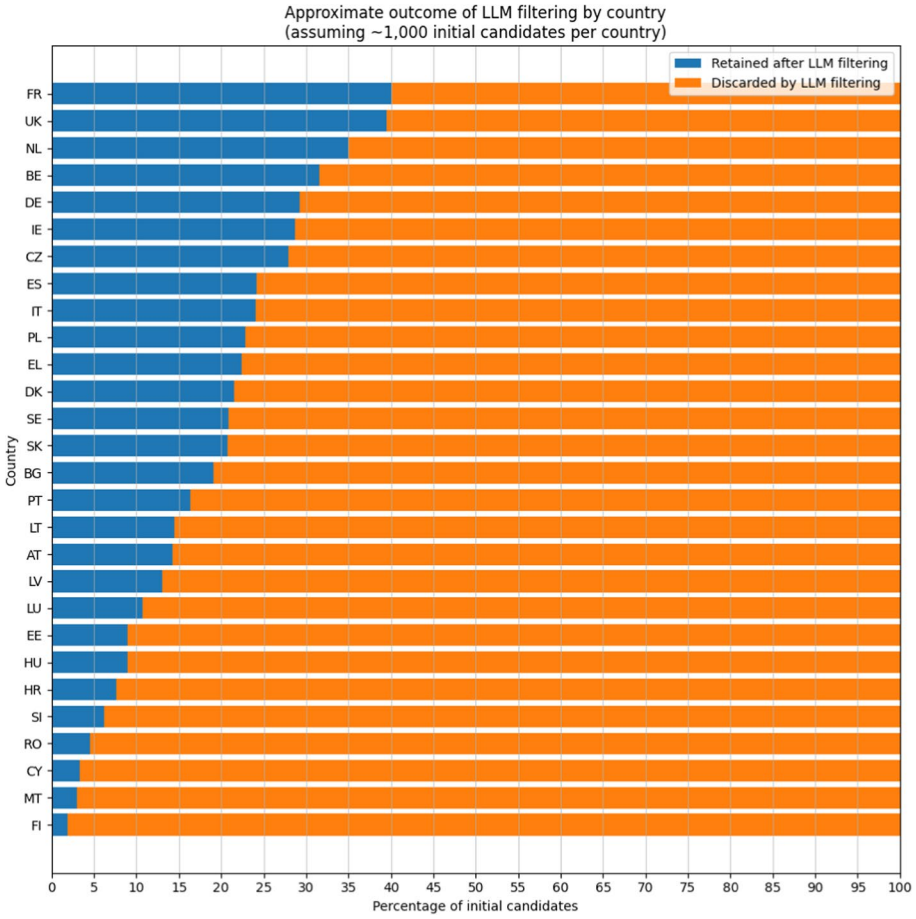


Fig. 3 Outcome of the LLM filtering step by country. Blue bars indicate candidates retained after LLM filtering and sent for expert validation, while orange bars indicate candidates discarded by the LLM

The prompt is instantiated by replacing the placeholders with the candidate skill and the top-3 ESCO preferred labels retrieved via embedding similarity, along with their corresponding descriptions.

LLM filtering statistics by country

This appendix reports the approximate proportion of candidate skills retained and discarded during the LLM-based filtering stage, broken down by country. Since the exact number of initial candidates varied slightly across countries, the percentages reported here are computed assuming approximately 1,000 initial candidates per country, which corresponds to the candidate set size used before expert validation (Fig. 3).

The figure shows substantial cross-country variation in retention rates. This variability is consistent with the broader discussion in the manuscript regarding differences in language coverage, data availability, and the robustness of country-specific representations.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s44163-026-01609-1.

Supplementary file 1 (xlsx 25 KB)

Author contributions

A.D.S. contributed to writing the original draft, data curation, methodology, formal analysis, software development, and resource collection. L.M. contributed to writing the original draft, data curation, methodology, formal analysis, and validation. F.M. contributed to conceptualization, investigation, methodology, formal analysis, project administration, supervision, and writing -review and editing. M.M. contributed to conceptualization, investigation, methodology, formal analysis, supervision, and writing - review and editing. N.N. contributed to writing the original draft, methodology, formal analysis, validation, and resource collection. All authors reviewed and approved the final manuscript.

Funding

Funding: not applicable.

Data availability

The data used in this study were accessed through the WIH-OJA NLP dataset within the PILLARS project consortium. The dataset can also be requested from EUROSTAT, subject to approval and the signing of a non-disclosure agreement (NDA).

Code availability

The code will be made available upon publication.

Declarations

Ethical approval

The study did not involve experiments, interventions, or observations of human research participants, nor the collection of personal, sensitive, or biological data. As stated in the manuscript, contributions from labor market experts were provided only to support the validation of the proposed methodological process. As the study did not constitute human subjects research, ethical approval was not required. The research was conducted in accordance with accepted ethical standards for academic research.

Consent to participate

Informed consent to participate was not required, as the study did not involve human subjects as research participants.

Consent to publish

Consent to publish was not required, as no identifying information about individuals was collected, analysed, or reported in this study.

Competing Interests

The authors declare no competing interests.

Received: 12 January 2026 / Accepted: 4 June 2026

Published online: 29 June 2026

References

1. Mooney RJ, Bunescu R. Mining knowledge from text using information extraction. ACM SIGKDD Explorations Newsl. 2005;7(1):3–10.

2. Calanca F, Sayfullina L, Minkus L, Wagner C, Malmi E. Responsible team players wanted: an analysis of soft skill requirements in job advertisements. EPJ Data Science. 2019;8(1):1–20.
3. Gugnani A, Kasireddy VKR, Ponnalagu K. *Generating unified candidate skill graph for career path recommendation*, in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (IEEE) 2018;328–333
4. Javed F, Hoang P, Mahoney T, McNair M. *Large-scale occupational skills normalization for online recruitment*, in *Proceedings of the AAAI Conference on Artificial Intelligence* 2017;31:4627–4634
5. Boselli R, Cesarini M, Mercorio F, Mezzanzanica M. Classifying online job advertisements through machine learning. Futur Gener Comput Syst. 2018;86:319–28.
6. Gurcan F, Cagiltay NE. Big data software engineering: Analysis of knowledge domains and skill sets using lda-based topic modeling. IEEE access. 2019;7:82541–52.
7. De Mauro A, Greco M, Grimaldi M, Ritala P. Human resources for big data professions: A systematic classification of job roles and required skill sets. Information Processing & Management. 2018;54(5):807–17.
8. Li S, Shi B, Yang J, Yan J, Wang S, Chen F, He Q. *Deep job understanding at linkedin*, in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* 2020;2145–2148
9. Jiechieu KFF, Tsopze N. Skills prediction based on multi-label resume classification using cnn with model predictions explanation. Neural Comput Appl. 2021;33(10):5069–87.
10. Goyal N, Kalra J, Sharma C, Mutharaju R, Sachdeva N, Kumaraguru P. JobXMLC: EXtreme multi-label classification of job skills with graph neural networks. Findings of the Association for Computational Linguistics EACL. 2023;2023:2181–91.
11. Bhola A, Halder K, Prasad A, Kan MY. *Retrieving skills from job descriptions: A language model based extreme multi-label classification framework*, in *Proceedings of the 28th international conference on computational linguistics* 2020;5832–5842
12. Zhang M, Jensen KN, Sonniks SD, Plank B. Skillspan: Hard and soft skill extraction from english job postings. 2022 arXiv preprint arXiv:2204.12811
13. Barducci A, Iannaccone S, La Gatta V, Moscato V, Sperlì G, Zavota S. An end-to-end framework for information extraction from italian resumes. Expert Syst Appl. 2022;210:118487.
14. Fang C, Qin C, Zhang Q, Yao K, Zhang J, Zhu H, Zhuang F, Xiong H. *Recruitpro: A pretrained language model with skill-aware prompt learning for intelligent recruitment*, in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 2023;3991–4002
15. Zhao M, Javed F, Jacob F, McNair M. Skill: A system for skill identification and normalization. Proceedings of the AAAI Conference on Artificial Intelligence. 2015;29:4012–7. https://doi.org/10.1609/aaai.v29i2.19064.
16. Zhang M, Jensen KN, Plank B. *Kompetencer: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning*, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference* 2022;436–447
17. Zhang M, van der Goot R, Plank B. *ESCOXLM-R: Multilingual Taxonomy-driven Pre-training for the Job Market Domain*, in *The 61st Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics) 2023;11871–11890
18. Li N, Kang B, Bie TD. Skillgpt: a restful api service for skill extraction and standardization using a large language model. 2023 arXiv preprint arXiv:2304.11060
19. Marconi G, Baer I, Silveira MD, Gallais M, Pruski C. Evaluating esco skill classifiers. Stat J IAOS. 2025;41(4):1038–57. https://doi.org/10.1177/18747655251395229.
20. Kavargyris DC, Georgiou K, Papaioannou E, Petrakis K, Mittas N, Angelis L. Escox: A tool for skill and occupation extraction using llms from unstructured text. Software Impacts. 2025;25:100772. https://doi.org/10.1016/j.simpa.2025.100772 (https://www.sciencedirect.com/science/article/pii/S2665963825000326).
21. Becker T, Koddebusch M. *In Search of Digital Talent: Evidence from Public Sector Job Advertisements in Germany Using LLM-Enabled Competence Classification*, in *Electronic Participation: 17th IFIP WG 8.5 International Conference, EPart 2025, Krems, Austria, August 31–September 4, 2025, Proceedings* (Springer-Verlag, Berlin, Heidelberg) 2025;344–359 https://doi.org/10.1007/978-3-032-02515-9_21.
22. Wang JM, Sun Z. *LLM-Supervised Multilingual Skill Extraction and Classification from Job Ads*, in *International Conference on Applications of Natural Language to Data Bases* 2025; https://api.semanticscholar.org/CorpusID:280070185
23. Decorte JJ, Van Hautte J, Deleu J, Develder C, Demeester T. *Design of negative sampling strategies for distantly supervised skill extraction*, in *RecSys in HR* (CEUR) 2022;3218
24. Clavié B, Soulié G. Large language models as batteries-included zero-shot esco skills matchers. 2023.
25. Fareri S, Melluso N, Chiarello F, Fantoni G. Skillner: Mining and mapping soft skills from any text. Expert Syst Appl. 2021;184:115544.
26. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the association for computational linguistics. 2017;5:135–46.
27. Giabelli A, Malandri L, Mercorio F, Mezzanzanica M, Seveso A. *NEO: A tool for taxonomy enrichment with new emerging occupations*, in *International Semantic Web Conference* (Springer) 2020;568–584
28. D'Amico S, De Santo A, Mercorio F, Mezzanzanica M. *Enriching Skill Taxonomies through Vector Space Models*, in *2024 IEEE International Conference on Big Data (BigData)* (IEEE) 2024;2297–2302
29. Giabelli A, Malandri L, Mercorio F, Mezzanzanica M, Nobani N. Embeddings evaluation using a novel measure of semantic similarity. Cogn Comput. 2022;14(2):749–63.
30. Muennighoff N, Tazi N, Magne L, Reimers N. *MTEB: Massive Text Embedding Benchmark*, in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* 2023;2014–2037

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.