

Received: 27 July 2026; **Revised:** 21 September 2026; **Accepted:** 30 September 2026; **Published:** 18 October 2026

Archs Sci. (2026) Volume 76(4), Pages 56-68, Paper ID 2026407.
<https://doi.org/10.68304/as/76407>

Learning Skill Representations from Recruitment Data for Skill Gap Analysis and Career Path Optimization

Peng Cheng

College of Art, Hubei Polytechnic University, Huangshi 435000, Hubei, China; 13886488957@163.com

Abstract To address the mismatch between educational provision and labor-market demand affecting the employment of arts graduates, this paper proposes a skill-gap analysis framework integrating recruitment-text mining, semantic representation learning, and knowledge graphs. First, a corpus of job advertisements is compiled from online recruitment data, and a Transformer-based model is used to encode job descriptions semantically, enabling the extraction of skill entities and the development of context-sensitive representations. Second, semantic skill embeddings, hierarchical clustering, and BERTopic-based topic modeling are combined to construct a skill-classification system for arts-related positions, with the aim of identifying core skill groups and emerging competency requirements in the employment market. Furthermore, the proposed framework maps job positions, skill requirements, and students' existing abilities into a shared semantic space and uses knowledge graphs to characterize structural relationships among jobs, skills, and competencies. On this basis, distributions representing market demand and educational skill supply are constructed and compared to quantify gaps affecting arts graduates. The reported experimental findings suggest that requirements for arts-related positions increasingly combine conventional professional expertise with digital technology, AI-assisted creative practices, data literacy, and transferable professional competencies. The results also identify apparent shortcomings in university preparation concerning computing skills, problem-solving, project implementation, and adaptability across disciplinary boundaries. Together, these analyses offer a structured basis for interpreting skill mismatches and informing curriculum development, individualized upskilling, and career planning.

Index Terms arts graduates; skill gap; recruitment text mining; transformer; knowledge graph; career path recommendation

1. Introduction

With the rapid expansion of the digital economy and the growth of creative industries, employment opportunities and competency expectations for arts professionals are undergoing substantial structural changes. Job requirements that previously emphasized aesthetic judgment and discipline-specific expertise increasingly include combinations of artistic, technological, and data-related capabilities [1], [2]. This is particularly true in emerging fields such as digital media, interaction design, and intelligent content generation, where employers seek artistic and creative production skills alongside technical proficiency, data-analysis capabilities, and effective interdisciplinary collaboration. Consequently, difficulties faced by arts graduates cannot be understood solely as a shortage of available positions. They also involve mismatches between the competencies graduates possess and those employers require, reflecting a structural gap between educational supply and labor-market demand [3].

From an educational perspective, many university arts programs continue to emphasize disciplinary theory and foundational practice, with comparatively limited attention to ad-

vanced digital skills, project-management capabilities, and interdisciplinary competencies [4], [5]. Such training may not adequately reflect employers' growing demand for professionals who can combine several complementary competencies. As a result, graduates may face skills mismatches and difficulty adapting their knowledge and practical abilities to changing occupational requirements. Systematically examining competencies specified in job advertisements and comparing them with the abilities developed through arts education is therefore an important research problem in educational planning and labor-market analysis [6].

The widespread use of online recruitment platforms has created extensive collections of job-advertisement text, providing an increasingly important evidence base for analyzing employers' stated skill requirements. Previous research can be divided into three broad methodological categories. The first uses statistical methods, such as word-frequency analysis and term frequency-inverse document frequency (TF-IDF), to identify frequently requested skills. The second applies machine learning and natural language processing to extract and classify skill expressions through topic models or neural

networks. The third employs knowledge graphs and network analysis to represent the relationships between occupations and competencies, allowing researchers to examine skill structures and possible career pathways. However, these studies still have several limitations [7], [8].

First, conventional statistical approaches do not consistently capture contextual meaning or the semantic relationships among skill expressions, which may limit the structure and interpretability of the resulting analysis. Second, although deep learning has improved the extraction of skills from unstructured texts, identifying individual skill mentions does not by itself provide a systematic account of categories, relationships among competencies, or gaps between the skills demanded and supplied [9]. Third, some approaches examine only one analytical level, such as individual skills or occupational titles, without integrating the relationships among positions, specific competencies, and individual abilities into a single modeling framework. Moreover, static analyses may overlook how occupational requirements evolve over time, thereby limiting their usefulness for curriculum adjustment, long-term career planning, and targeted skill-development decisions [10], [11].

These methodological limitations are especially relevant to research on employment in the arts and creative industries. Arts-related jobs draw on a diverse and interdisciplinary set of competencies, making their requirements difficult to capture accurately through narrowly defined occupational classifications. At the same time, assessments of arts graduates' abilities may depend on subjective evaluations or relatively small surveys rather than systematic comparisons with large collections of recruitment evidence [12]. Therefore, it is necessary to construct a comprehensive method that integrates semantic representation, structural modeling, and dynamic analysis to more comprehensively characterize the skill requirements of arts-related positions and identify the gap between talent cultivation and market demand.

To address these issues, the present study proposes a multi-level, data-driven framework for examining the employment prospects and competency gaps of arts graduates. Online recruitment texts provide the primary source material. Semantic representation learning, skill clustering, knowledge-graph construction, and dynamic simulation are then combined in a modeling process extending from skill extraction and classification to potential career-path recommendations. Specifically, the proposed approach first uses a Transformer architecture to encode recruitment texts and applies multi-head attention to represent contextual relationships among competencies, producing contextualized skill vectors for subsequent analysis [13], [14]. Second, a hierarchy of skill categories is generated by clustering semantic embeddings and examining the resulting groups through topic modeling. This stage is designed to distinguish broad competency domains from specialized and emerging skill combinations. Third, job positions, skills, and individual competencies are mapped to a common semantic space, allowing similarities, mismatches, and other structural differences to be examined quantitatively. A knowledge graph

is then constructed to represent hierarchical associations, co-occurrence patterns, and other interpretable relationships among competencies and occupations. Finally, the framework incorporates a dynamic evolution model to explore possible changes in skill requirements and their implications for career development. These interconnected stages are intended to support more targeted, adaptable skill-enhancement strategies.

II. Methodological Framework for Skill Representation and Data-driven Recruitment

A. Methodology for skill classification and gap quantification in arts-related positions

As shown in Figure 1, the present study develops an integrated framework for skill classification and demand analysis for arts graduates based on online recruitment data. The overall process can be summarized into four stages: candidate skill description generation, quality screening, semantic classification, and skill demand modeling. Its core objective is to transform unstructured recruitment texts into structured skill distributions, thereby providing a reliable foundation for subsequent skill gap analysis [15], [16].

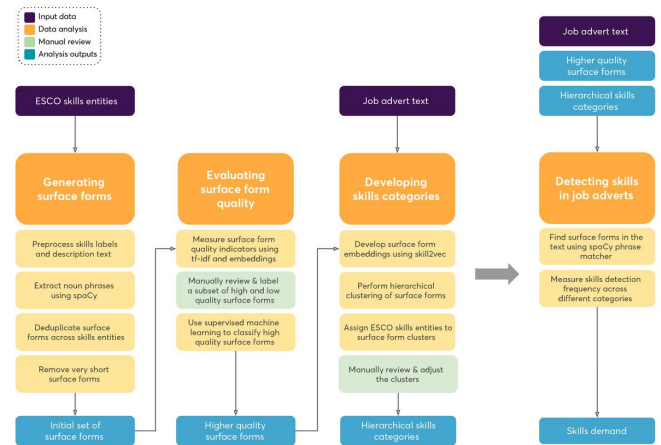


Figure 1: Framework of the proposed skill extraction, classification, and demand modeling approach

At the data-processing stage, the framework accepts a large corpus of recruitment texts and extracts candidate skill descriptions through natural language processing techniques, including tokenization, part-of-speech tagging, and phrase identification. Because vacancy descriptions can contain ambiguous expressions, such as "communication skills" and "sense of responsibility," directly treating every extracted phrase as a distinct occupational competency may introduce noise and reduce interpretability [17]. To address this concern, the framework incorporates a quality-assessment procedure for screening candidate expressions. Each candidate is evaluated using term frequency-inverse document frequency (TF-IDF), contextual semantic consistency, and occupational-domain relevance. Descriptions with clearer professional meaning can then be retained to produce a more consistent and informative skill inventory. The resulting inventory provides the

terminology used by subsequent classification and quantitative comparison stages.

In the skill-classification stage, the approach moves beyond exclusive reliance on manually defined labels or simple keyword matching. Instead, a semantic representation model encodes descriptions as continuous vectors, and similarity measures are used to group related competencies into hierarchical categories. For arts-related positions, the resulting clusters may include visual-design skills, digital-media software proficiency, interaction-design capabilities, and creative-planning competencies. These groups can subsequently be examined across occupations and compared with students' educational experiences. In this way, the modeling procedure transforms initially disconnected terms into a more systematic and interpretable skill classification, while retaining the contextual relationships expressed in the original advertisements.

Building on the extracted representations, the analysis models the requirements expressed in job advertisements. For every skill category, the frequency of its occurrence is combined with contextual weights to describe its relative prominence in the observed recruitment corpus. Categories associated with several related expressions can thereby be analyzed consistently without relying solely on exact repetitions of the same words. Specifically, the estimated market-demand intensity for skill category l is defined as follows:

$$D_l = \frac{1}{N} \sum_{i=1}^N \sum_{s_k \in C_l} \mathbb{I}(s_k \in d_i) \cdot w_{i,k}, \quad (1)$$

Here, d_i denotes the i th job advertisement, C_l denotes the set of skill expressions assigned to category l , $\mathbb{I}(\cdot)$ is an indicator function, and $w_{i,k}$ is the context-dependent weight assigned to skill expression s_k in advertisement d_i . The formula aggregates weighted observations across the available recruitment records, producing a category-level measure of demand intensity. The resulting category scores can be normalized to obtain a full demand-distribution vector, $\mathbf{P}_{demand}$, for comparison with corresponding estimates of educational skill supply.

In parallel with market-demand estimation, the framework draws on university curricula, student portfolios, and graduation-project descriptions, where such information is available, to estimate a distribution of the competencies represented in educational preparation. This second distribution is denoted by $\mathbf{P}_{supply}$. To quantify the discrepancy between these two representations, the study introduces a distribution-based skill-gap measure that compares the normalized market and educational skill profiles:

$$Gap = \sum_{l=1}^L D_l \log \frac{D_l + \epsilon}{S_l + \epsilon}, \quad (2)$$

Here, D_l and S_l denote the normalized market-demand and student-supply shares for skill category l , respectively, and ϵ is a positive smoothing constant introduced to avoid undefined logarithms. The expression characterizes discrepancies between the two profiles: when a category accounts for

a relatively large proportion of demand but a smaller proportion of educational supply, its contribution can increase. Interpretation of the overall score requires the same category definitions, consistent normalization, and careful selection of the smoothing parameter. The equation should therefore be evaluated against the documented data-processing procedure before its numerical values are used to draw conclusions about absolute shortages.

B. Arts-related employment skill data and the decision chain

Figure 2 outlines the three-stage "data-model-decision" chain, which connects the proposed skill-classification and mismatch-assessment procedures with their potential use in recruitment and career guidance [18]. The first stage concerns the acquisition and preparation of recruitment and education records; the second produces semantically structured and comparable skill representations; and the final stage applies these representations to decision-support tasks. This organization makes explicit how information can progress from raw documents through quantitative modeling to interpretable recommendations, while acknowledging that final recruitment decisions require contextual and human review.

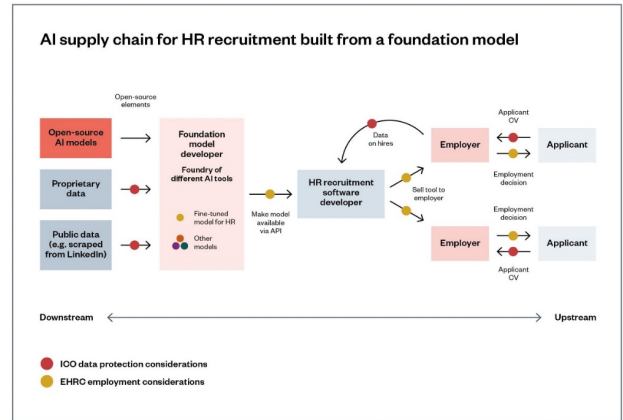


Figure 2: AI-enabled data-to-decision pipeline for skill extraction and recruitment matching

At the input stage, the framework brings together several categories of information, including publicly available vacancies, permitted data obtained through platform scraping, and structured organizational resources. These sources contain job descriptions and additional contextual details that can be used to infer employers' stated skill requirements. Because sources differ in completeness, duplication rates, access permissions, and terminology, they require careful integration and standardized processing. The combined input corpus is represented as follows:

$$\mathcal{D} = \mathcal{D}_{pub} \cup \mathcal{D}_{prop} \cup \mathcal{D}_{scr}, \quad (3)$$

where $\mathcal{D}_{pub}$ represents publicly available data, $\mathcal{D}_{prop}$ represents data from within the platform or organization, and $\mathcal{D}_{scr}$ represents data obtained through web scraping. Different data sources vary in coverage and noise levels, thus requiring integration through a unified data cleaning and standardization process.

At the semantic-modeling stage, a pretrained language model encodes recruitment documents into contextual vector representations. Descriptions of similar skills may use different vocabulary across employers, job roles, and online platforms. Encoding these expressions within a common semantic space facilitates their comparison without requiring exact lexical equivalence. The representations therefore provide a basis for extracting candidate skills, forming categories, and analyzing the relationships between advertised competencies and individual abilities. This process can be represented as:

$$\mathbf{H} = f_{\theta}(\mathcal{D}), \quad (4)$$

where $f_{\theta}(\cdot)$ represents the pre-trained language model, and $\mathbf{H}$ is the set of high-dimensional semantic representations of the recruitment text. Through this process, the system can overcome the limitations of keyword matching and achieve the recognition of implicit skill expressions.

During the intermediate stage, the semantic representations support skill recognition and the construction of a structured competency model. Candidate descriptions are clustered and mapped to a hierarchical category scheme, which allows broader competencies and specialized tool-specific requirements to be examined at different levels of detail. The resulting labels and representations can be integrated into recruitment-analysis applications through application programming interfaces (APIs). Consequently, employment analytics, curriculum evaluation, and job recommendation tools can share consistent skill definitions instead of relying on disconnected collections of keywords.

At the final stage, the framework provides analytical support for matching occupational requirements to job seekers' competencies. Employers specify requirements in vacancy descriptions, while applicants may provide evidence of their abilities through resumes, portfolios, or other approved records. Both types of descriptions are processed through compatible semantic-encoding and classification procedures. Their resulting representations can then be compared within a shared vector space, producing an interpretable indicator of potential alignment. This indicator is intended to support, rather than replace, a comprehensive evaluation of applicants and job responsibilities. Let the vector of i th job requirement be $\mathbf{d}_i$, and the vector of j th job seeker's ability be $\mathbf{s}_j$. Then the matching degree can be expressed as:

$$\text{Match}(i, j) = \frac{\mathbf{d}_i^{\top} \mathbf{s}_j}{\|\mathbf{d}_i\| \|\mathbf{s}_j\|}. \quad (5)$$

The matching function computes cosine similarity between the representation of the job requirements and that of the applicant's skills. A higher value indicates greater directional

similarity in the encoded space, but it does not independently demonstrate that a candidate meets every mandatory qualification. In practical applications, skill-specific evidence, prerequisites, and any applicable fairness constraints should also be considered.

It is important to emphasize that, compared to traditional recruitment processes, the key advantages of this framework lie in its data-driven and structured-representation capabilities. First, by uniformly modeling recruitment texts through a basic model, data from different platforms and with different forms of expression becomes comparable. By contrast, by constructing a skill space, it achieves the transformation from "unstructured descriptions" to "computable vectors," shifting employment analysis from empirical judgment to quantitative analysis.

Furthermore, this framework also implicitly contains important institutional and ethical constraints. At the data level, privacy protection and data compliance issues need to be considered; at the decision-making level, the potential impact of algorithmic bias on job fairness needs to be addressed. Therefore, in practical applications, appropriate constraint mechanisms should be introduced during model design and system deployment to ensure the rationality and interpretability of the technology application.

C. Skill semantic mapping and visualization

To reveal the underlying semantic structure among skills in arts-related positions, this paper, based on skill extraction and classification, introduces semantic mapping and visualization methods to uniformly represent and spatially analyze skill and job labels. As shown in Figure 3, skill and job descriptions are embedded in a common semantic space and projected onto a two-dimensional plane. Dimensionality-reduction methods, including UMAP, offer a way to visualize these relationships, subject to the normal limitations of projection from a high-dimensional space [19].

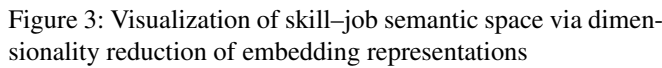
First, for the set of skill descriptions $S = s_1, s_2, \dots, s_K$ identified from recruitment texts, the study uses a pre-trained language model to semantically encode them, mapping the discrete text into a continuous vector representation:

$$\mathbf{v}_i = f_{\theta}(s_i), \quad (6)$$

where $f_{\theta}(\cdot)$ represents the semantic encoding function (such as a Transformer-based language model), and $\mathbf{v}_i \in \mathbb{R}^d$ is the high-dimensional semantic vector of skill s_i . Through this process, skills with different expressions but similar semantics (such as "Photoshop", "image processing", and "visual post-processing") can form a proximity distribution in the vector space.

Based on this, the study uses cosine similarity to measure the semantic relationships between skills:

$$\text{sim}(s_i, s_j) = \frac{\mathbf{v}_i^{\top} \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}. \quad (7)$$

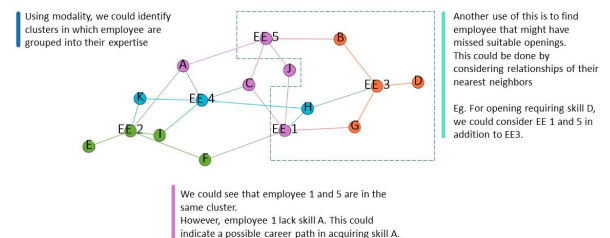


Since high-dimensional semantic spaces are difficult to interpret directly, the framework introduces a dimensionality reduction mapping function to project skill vectors onto a two-dimensional space for visualization:

where $g(\cdot)$ represents a nonlinear dimensionality reduction function (such as t-SNE or UMAP). This process compresses the high-dimensional semantic structure into a two-dimensional plane while preserving local proximity relationships as much as possible, thus forming the skill distribution structure in Figure 3.

D. Skill graph-based upskilling path identification

Finding new skills (upskilling) and employee's individual career paths



Let the observed skill collection be $S = \{s_1, s_2, \dots, s_K\}$. A weighted skill graph can be constructed by relating competency nodes according to similarities between semantic vectors, documented co-occurrence in recruitment texts, or both. The graph is represented as follows:

Here, V is the set of skill nodes, E is the set of connections between these nodes, and W denotes the edge weights used to characterize the strength or frequency of the observed relationships. A higher edge weight indicates greater semantic proximity between two skills, or a higher probability of them appearing together in similar job positions. Therefore, this graph not only reflects the static relationships between skills but also represents the observed structure of skill combinations in the job market.

where U_j represents the skills the student still needs to acquire to adapt to the target position. Furthermore, the system can determine which skills are more suitable as priority improvement targets based on the distance and edge weights between

nodes in the skill graph. If a skill to be improved is located in the same skill cluster as the student's existing skills or is close to them, this proximity may indicate a more accessible learning transition. Nevertheless, the actual difficulty of acquiring a competency also depends on formal prerequisites, instructional resources, practical experience, and the student's individual circumstances.

Based on this, the study defines a skill improvement priority score:

$$P(s_u | S_j) = \max_{s_i \in S_j} \frac{W(s_i, s_u)}{1 + \text{dist}(s_i, s_u)}, \quad (11)$$

where s_u denotes a candidate skill for improvement, conditional on the existing skill set S_j , $W(s_i, s_u)$ represents the correlation strength between the existing skill s_i and the target skill s_u , and $\text{dist}(s_i, s_u)$ represents the shortest path distance between the two in the skill graph. A higher value indicates that the target skill is closer to the student's existing abilities and is more suitable as a priority for improvement.

As shown in Figure 4, different skill nodes form several local clusters in the graph structure, indicating a clear community structure among skills. For example, if a student has mastered visual design skills but lacks interaction design or digital media production skills, the system can recommend a feasible potential learning pathway based on the nearest neighbor skill relationships. This method differs from simply listing missing skills; it further considers the learning connections between skills, thus providing students with more actionable career development suggestions.

E. Knowledge graph-based representation of job-skill relationships

To further characterize the structural relationships between job positions, skills, education levels, and enterprises, this paper constructs a job skill knowledge graph. As shown in Figure 5, the nodes in the graph include job position nodes, skill nodes, education level nodes, and enterprise nodes; edges represent the relationships between different entities, such as a job requiring a specific competency, an enterprise posting a specific occupational category, or a job requiring a specific education level. Through this graph structure, unstructured recruitment information is transformed into a computable relational network.

Let the knowledge graph be represented as:

$$\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{E}), \quad (12)$$

where $\mathcal{V}$ represents the set of entities, including job titles, skills, companies, and educational qualifications; $\mathcal{R}$ represents the set of relation types; and $\mathcal{E}$ represents the set of triples. Each relation can be represented as:

$$(h, r, t) \in \mathcal{E}. \quad (13)$$

In each relational triple, h denotes the head entity, r represents the relation type, and t denotes the tail entity. For example, "Visual Designer, requires, Photoshop" indicates that

a specific occupational category requires Photoshop skills, and "Company A, offers, Interaction Designer" indicates that Company A has posted interaction design positions.

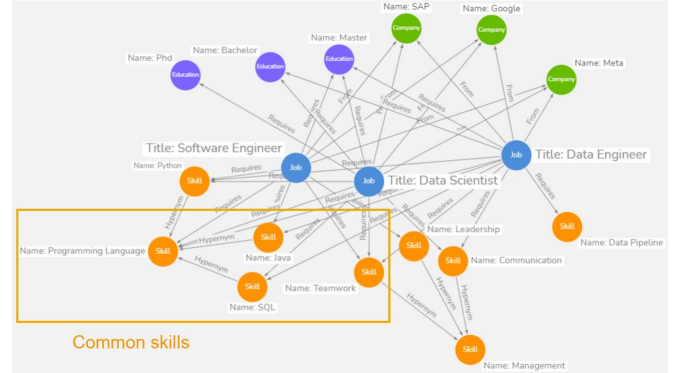


Figure 5: Knowledge graph representation of job-skill-education-company relationships

In the present study, the principal purpose of the knowledge graph is to represent different kinds of labor-market information within a common relational structure. Compared to simple text analysis, knowledge graphs can simultaneously preserve hierarchical relationships between skills, demand relationships between positions and skills, and posting relationships between companies and positions. For example, skills such as "Java," "Python," and "SQL" can be further categorized under the higher-level skill category of "programming languages"; "communication skills," "teamwork," and "project management" can be identified as cross-position common skills. The "common skills" area marked in the graph reflects these cross-position shared skills, which appear simultaneously in multiple positions and therefore have high explanatory value for graduates' employment competitiveness.

To measure the universality of a skill across different positions, this paper defines a skill universality score:

$$C(s) = \frac{|\{j \in J : (j, \text{requires}, s) \in \mathcal{E}\}|}{|J|}, \quad (14)$$

Here, J denotes the set of observed job positions, and s is the competency being evaluated. As $C(s)$ increases, the share of positions requiring that competency also increases, indicating broader relevance across the observed occupations. This indicator can be used to identify the fundamental abilities that arts graduates should prioritize mastering, such as visual expression, digital tool usage, communication and collaboration skills, and project execution abilities.

III. Semantic representation model for recruitment text

As illustrated in Figure 6, the proposed system uses a Transformer architecture to encode online vacancy descriptions and produce contextual representations of job titles, competencies, educational qualifications, and employers' stated requirements. In contrast to conventional keyword matching, a Transformer can represent associations among related expressions, including terms separated by several other words in the

same recruitment document, through a self-attention mechanism. For example, in the phrase "proficient in Photoshop and possessing visual expression skills," "Photoshop" is not only a software skill but also implicitly related to arts-related job skills such as "visual expression skills," "graphic design," and "portfolio creation."

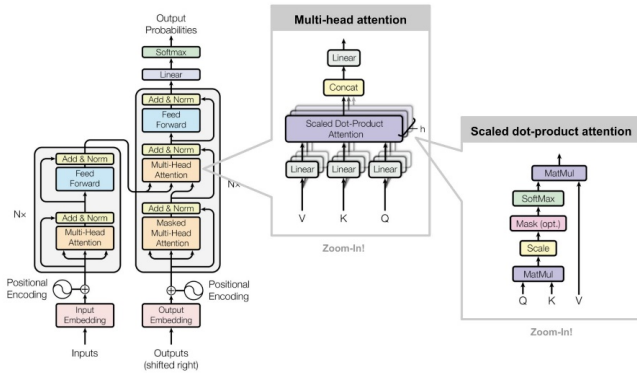


Figure 6: Semantic representation model for recruitment text and skill-demand extraction

Suppose a job posting text is represented as a sequence of terms:

$$X = (x_1, x_2, \dots, x_n). \quad (15)$$

The initial representation is obtained after word embedding and positional encoding:

$$\mathbf{H}^{(0)} = \mathbf{E}(X) + \mathbf{P}, \quad (16)$$

Here, $\mathbf{E}(X)$ is the matrix of token embeddings, and $\mathbf{P}$ represents positional information added to the initial token representations. These quantities are passed to successive attention layers to learn contextual associations between words and phrases. The multi-head attention operation is expressed as follows:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}^O. \quad (17)$$

Each attention head is defined as follows:

$$\text{head}_r = \text{Softmax} \left(\frac{\mathbf{Q} \mathbf{W}_r^Q (\mathbf{K} \mathbf{W}_r^K)^T}{\sqrt{d_k}} + \mathbf{M} \right) \mathbf{V} \mathbf{W}_r^V, \quad (18)$$

In this expression, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices, respectively; $\mathbf{M}$ is an optional attention mask; and d_k is the dimension of the projected key vectors, used in the scaling term. Through a multi-head attention mechanism, the model can simultaneously capture the relationships between skill terms, job descriptions, and competency requirements from multiple semantic subspaces.

After applying the successive Transformer layers, a pooling operation aggregates the contextual token vectors into a single representation of the complete recruitment document:

$$\mathbf{z}_i = \text{Pool}(\mathbf{H}_i^{(L)}), \quad (19)$$

Here, $\mathbf{H}_i^{(L)}$ is the contextual representation of the i th recruitment document after the L th Transformer layer, and $\mathbf{z}_i$ is the pooled, document-level semantic vector. This vector is subsequently used for skill extraction, skill category identification, job requirement modeling, and skill gap calculation.

The framework also describes a classification stage for identifying skill categories within a recruitment text. For the i th posting, the reported formulation estimates the probability associated with the l th skill category using a softmax function. Because softmax normalizes probabilities across mutually exclusive categories, its use for a genuinely multilabel task requires further methodological clarification.

$$p_{i,l} = \frac{\exp(\mathbf{w}_l^T \mathbf{z}_i + b_l)}{\sum_{r=1}^L \exp(\mathbf{w}_r^T \mathbf{z}_i + b_r)}, \quad (20)$$

Here, L denotes the number of categories, while $\mathbf{w}_l$ and b_l are the learned weight vector and bias associated with skill category l , respectively. This probability value can be used to measure the demand intensity of different skill categories in recruitment texts and serve as the basis for constructing the subsequent market skill demand distribution.

A. Skill-demand evolution and career pathways

Figure 7 presents the proposed knowledge-evolution framework, which extends the semantic representation of recruitment texts to a dynamic graph of occupations, required competencies, employers, and students' abilities. The model considers both the relationships present at a particular time and possible changes in those relationships as the employment market evolves. This approach is intended to support assessments of emerging skill requirements and the planning of potential upskilling pathways for arts graduates [21].

This paper models the employment market as a dynamic heterogeneous graph:

$$\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t, \mathcal{R}, \mathbf{X}_t), \quad (21)$$

Here, $\mathcal{V}_t$ is the set of nodes at time t , including job, skill, employer, and student-competency nodes; $\mathcal{E}_t$ contains the relationships between these entities; $\mathcal{R}$ represents the relation type, such as "job requires skills," "company posts job," "skill belongs to a higher-level category," "student possesses skills," etc.; and $\mathbf{X}_t$ represents the node feature matrix.

For any two nodes v_i and v_j , the strength of a relation of type r at time t is modeled by:

$$a_{ij}^{(r,t)} = \sigma \left(\mathbf{u}_r^\top [\mathbf{x}_{i,t} \parallel \mathbf{x}_{j,t} \odot \mathbf{x}_{j,t}] \right), \quad (22)$$

Here, r is the relation type; $\mathbf{x}_{i,t}$ and $\mathbf{x}_{j,t}$ are the feature vectors of nodes i and j at time t ; $\parallel$ denotes vector concatenation; $\odot$ represents elementwise multiplication; and $\sigma(\cdot)$ denotes the sigmoid activation function. This formula is used to characterize the strength of association between different entities, such as the demand for a specific competency in a job, or the closeness between a student's existing skills and the target skill.

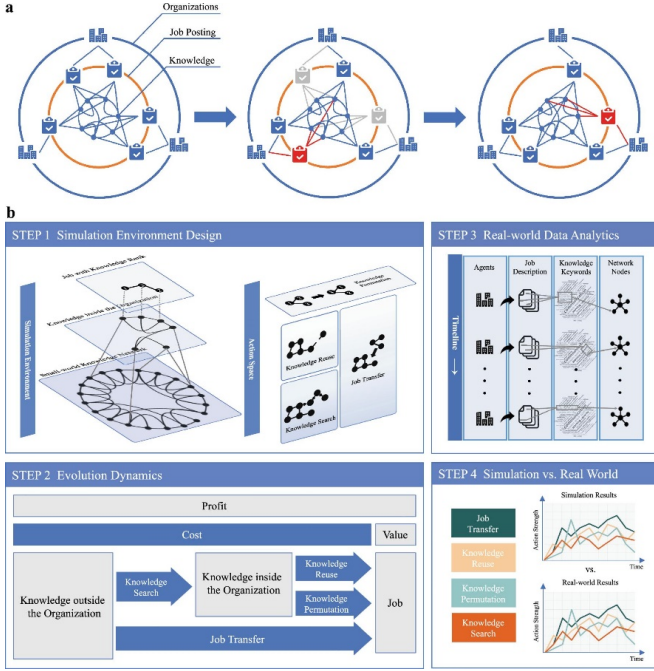


Figure 7: Knowledge-evolution simulation framework for employment and skill relationships

In the process of knowledge evolution, the node state is updated as the job market changes. This paper uses a graph neural network to describe this propagation process:

$$\mathbf{x}_{i,t+1} = \sigma \left(\mathbf{W}_0 \mathbf{x}_{i,t} + \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{a_{ij}^{(r,t)}}{\sqrt{|\mathcal{N}_i^r| |\mathcal{N}_j^r|}} \mathbf{W}_r \mathbf{x}_{j,t} \right), \quad (23)$$

Here, $\mathcal{N}_i^r$ denotes the neighbors of node i connected through relation type r , while $\mathbf{W}_0$ and $\mathbf{W}_r$ are learnable transformation matrices. This formula reflects the propagation process of skill demand, job changes, and corporate recruitment preferences in the graph structure. For example, when the demand for skills such as "interaction design," "AIGC visual generation," and "digital media production" increases in the market, the related job nodes and student ability nodes will also undergo state changes accordingly.

To represent personalized skill-development pathways for arts graduates, let $\mathbf{s}_{u,t}$ denote student u 's competency vector at time t and $\mathbf{d}_{j,t}$ denote the requirements associated with target occupation j at the same time. The model evaluates their similarity as follows:

$$\text{Match}_{u,j,t} = \frac{\mathbf{s}_{u,t}^\top \mathbf{d}_{j,t}}{\|\mathbf{s}_{u,t}\| \|\mathbf{d}_{j,t}\|}. \quad (24)$$

Based on this, the system selects skill enhancement actions from the candidate skill set that maximize future job suitability. Let the available actions be:

$$a_t \in \mathcal{A} = \{\text{learn, reuse, transfer, combine}\}, \quad (25)$$

where "learn" represents learning new skills, "reuse" represents reusing existing skills, "transfer" represents skill transfer, and "combine" represents combining multiple existing skills to form a composite ability. The skill enhancement strategy is defined as:

$$\pi(a_t | \mathbf{s}_{u,t}, \mathcal{G}_t) = \text{Softmax}(\mathbf{W}_\pi [\mathbf{s}_{u,t} \| \text{Readout}(\mathcal{G}_t)] + \mathbf{b}_\pi). \quad (26)$$

The policy assigns probabilities to possible learning actions on the basis of the student's current abilities and the structure of the modeled skill graph. The choice of a particular action should additionally consider prerequisites, available training, and the resources required for successful completion.

Ultimately, the proposed framework models career path optimization as a long-term utility maximization problem:

$$\max_{\pi} \mathbb{E}_{\pi} \sum_{t=0}^T \gamma^t \left[\alpha \text{Match}_{u,j,t} + \beta \Delta \text{Skill}_{u,t} \lambda \text{Cost}(a_t) \right], \quad (27)$$

In this objective, γ denotes the temporal discount factor, $\Delta \text{Skill}_{u,t}$ is the skill improvement associated with student u at time t , and $\text{Cost}(a_t)$ represents the learning cost associated with action a_t . The coefficients α , β , and λ are weighting parameters. The operation connecting skill gain and learning cost in the source equation remains ambiguous and requires confirmation from the author. This objective function embodies the core idea of the model in this paper: the skill improvement path should not only pursue short-term matching, but should simultaneously consider skill growth, learning cost, and future employment adaptability.

IV. Experiments and Results

A. Data sources and experimental setup

The experiments described in this study use online recruitment information from multiple sources, primarily from mainstream recruitment platforms, covering art design, digital media, interaction design, and related technical positions. The original records include fields such as job descriptions, skill requirements, company information, and job categories. After data cleaning, deduplication, and standardization, these observations are assembled into a structured dataset containing job descriptions, extracted skill tags, and candidate-ability descriptions. Similar approaches to deriving skill-demand indicators from job advertisements have been reported in labor-market research [22], [23]; these citations provide methodological context rather than independent verification of the dataset used in this study.

To evaluate the proposed method under differences in sample composition, the study describes a dissimilarity-based partitioning procedure rather than relying exclusively on a

random train–test split. The stated aim is to preserve broad comparability between the training and test collections while allowing meaningful differences in their semantic distributions. Such a design may be useful for examining generalization to changing recruitment requirements, although the exact partitioning criterion and sampling parameters must be specified for reproducible evaluation.

The experimental design comprises three complementary categories of model components:

- A Transformer-based semantic encoder for recruitment texts and contextual skill identification [25].
- A clustering- and topic-modeling-based method for analyzing skill structures and categories.
- A knowledge-graph-based approach to relationships among skills and potential career-development pathways.

The reported evaluation considerations include consistency between skill distributions, interpretability of the modeled semantic structure, and the utility of the resulting representations for comparing occupational requirements with candidate competencies.

B. Results analysis

As shown in Figure 8, the analysis further verifies the differences and clustering characteristics of skill demand structures across different job types using hierarchical clustering heatmaps. In Figure 8, the horizontal axis represents the extracted and standardized skill categories, the vertical axis represents different job roles, and the color intensity indicates the demand intensity for specific skills in a particular job. The visualization suggests that different jobs do not randomly require skills, but rather exhibit a clear structured distribution: some skills maintain high demand across multiple jobs, indicating strong versatility; while other skills are concentrated in specific job categories, reflecting strong professional orientation. Corresponding to the skill semantic representation, skill classification, and skill gap quantification methods proposed earlier in this paper, this result demonstrates that the skill demand matrix constructed based on recruitment texts can effectively reveal the capability boundaries and similarity relationships between jobs. In particular, when certain jobs are close together in the clustering tree, it indicates that they share similar skill combinations, providing empirical evidence for cross-job employment paths for arts graduates, curriculum module adjustments, and personalized skill enhancement suggestions.

As shown in Figure 9, the analysis further compares the differences between the skills provided by higher education institutions and the actual requirements of the labor market. In the radar chart, the dark line represents the skill level covered in school training, while the light line represents the skill level required in market recruitment. The results show significant discrepancies in most ability dimensions, particularly in computer skills, time management, problem-solving, planning, creativity, resilience, initiative, and adaptability. Market demand generally exceeds school supply, indicating

a structural gap between the training system for arts graduates and the actual needs of the job market. In comparison, schools perform relatively well in terms of traditional courses such as professional theoretical knowledge and professional practical knowledge, but they are still insufficient in digital tool skills, comprehensive professional abilities, and cross-job transfer capabilities.

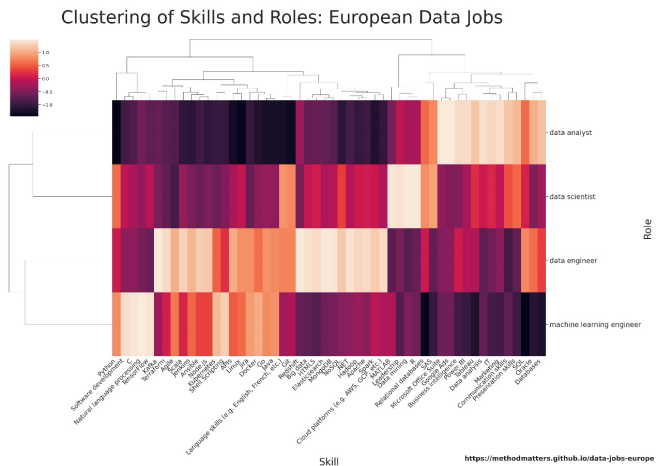


Figure 8: Hierarchical clustering heatmap of skill demand patterns

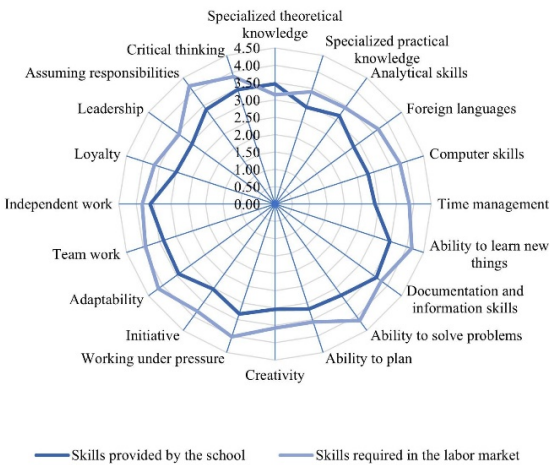


Figure 9: Comparison between school-provided skills and labor-market-required skills among arts graduates

As shown in Figure 10, the analysis further examines the changes in market demand for different job categories over the observation period. The results show that the demand for data science positions grew rapidly after 2023, reaching its peak in 2026, indicating that digital analysis capabilities and intelligent technology application skills are becoming important growth points in the job market. Software engineering positions maintained a steady upward trend overall, indicating that computer technology-related skills remain in continuous demand. Graphic design positions saw relatively

moderate growth, suggesting that traditional art and design positions still have market space, but the growth rate is significantly lower than that of technology-intensive positions such as data science and software engineering. In contrast, accounting positions showed a downward trend, while HR positions experienced less fluctuation, indicating limited market expansion. These results corroborate the analysis of the skill gap among arts graduates in this paper: current art positions do not necessarily rely solely on traditional aesthetic and design skills, but are gradually expanding towards digital tool usage, data understanding, cross-media expression, and technological integration capabilities.

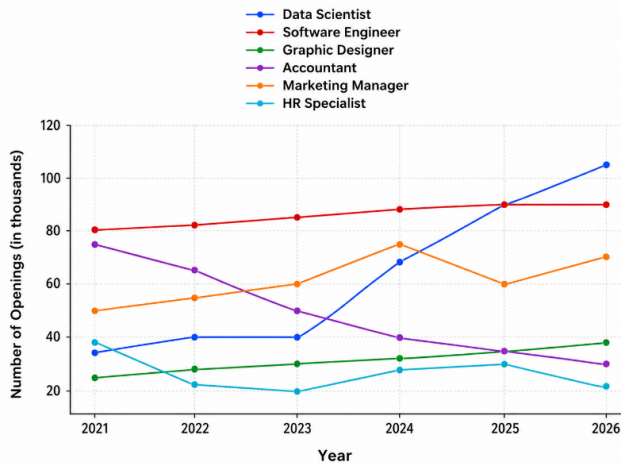


Figure 10: Temporal trends of job demand

As shown in Figure 11, the study examines the skill requirements in recruitment texts from two perspectives: temporal evolution and overall frequency. Panels (a) and (b), respectively, show the changes in demand for theoretical hard skills and tool-based hard skills in different years. The results show that computer-related skills such as machine learning, data science, artificial intelligence, Python, and SQL have shown a continuous upward trend in recent years, indicating that digital technology capabilities are becoming a core component of job requirements. Panel (c) shows the changes in soft skills. It can be seen that communication skills, problem-solving skills, teamwork, and project management skills, although fluctuating relatively little, have always maintained high demand, indicating that the job market has not simply shifted to a technology-oriented approach, but rather emphasizes a composite ability structure of "technical skills + comprehensive professional qualities". Panel (d) additionally presents a ranking of high-frequency skills in recorded job postings, with hard skills such as machine learning, Python, C++, Java, SQL, and data analysis occupying major positions, while soft skills such as research ability, project management, and public speaking have also entered the important skill set [24].

As shown in Figure 12, the study uses BERTopic to perform semantic clustering and visualization analysis of skill requirement topics in recruitment texts. Different colored clusters

represent skill topics formed by semantically similar texts, and the distance between clusters reflects the strength of semantic association between topics. The labeled text indicates the main meaning of each topic. The results show that skill requirements in the recruitment market are not scattered but form several stable topic clusters, including digital design tools, AI-assisted creation, data analysis, user experience, project management, cross-media communication, and general professional skills. This result supports the skill semantic representation and hierarchical classification methods proposed earlier in this paper, demonstrating that topic modeling based on recruitment texts can effectively reveal the potential structure of skill requirements for arts-related positions.

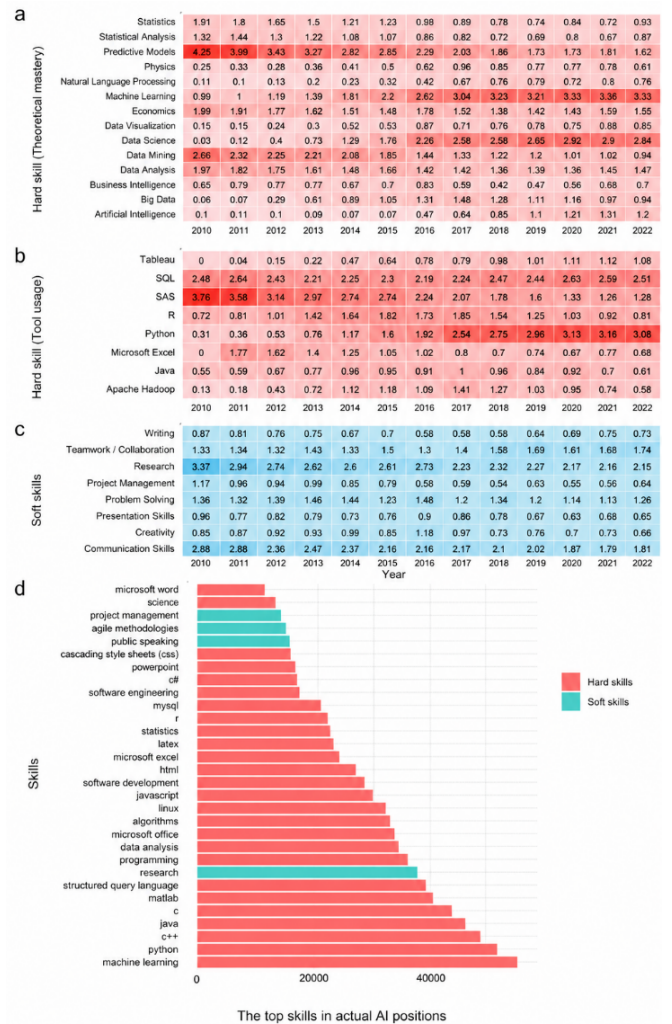


Figure 11: Temporal evolution and frequency distribution of hard and soft skills

As shown in Figure 13, this paper maps job roles, skill tags, and student/candidate ability representations into the same semantic space to observe the matching relationship among the three. In Figure 13, blue nodes represent job roles, orange nodes represent skills, and black nodes represent individual students or candidates. The results show that most candidates

are concentrated in the central region of the semantic space, indicating that their ability structure mainly covers general skills and basic professional abilities; while some skill nodes and job nodes are distributed in the peripheral region, indicating that these abilities are more specialized or emerging, and are currently under-represented by the student population. This result is consistent with the skill gap model proposed in this paper, that is, arts graduates are not completely lacking in employability, but rather there is a spatial distance between their ability distribution and the skills of some high-demand jobs.

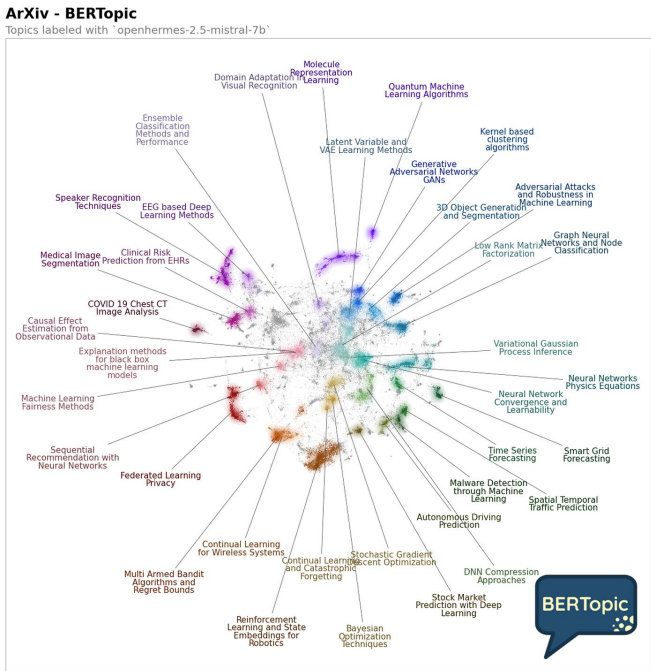


Figure 12: Topic modeling and semantic clustering of recruitment skills

As shown in Figure 14, this paper employs a training and test set partitioning method based on dissimilarity to examine the robustness of the model under different sample distributions. The upper two-dimensional scatter plot displays the distribution relationship between the training and test samples through principal component space. The visualization suggests that there is some overlap between the two sets, while retaining some structural differences, indicating that the test set is not a simple repetition of the training set, but rather a more realistic test of the model's generalization ability. The lower box plot further compares the distribution of the training and test sets across multiple variable dimensions. The results show that most variables maintain similar medians and dispersion ranges in both datasets, but some variables still exhibit differences, providing more stringent validation conditions for subsequent model evaluation.

As shown in Figure 15, employers' assessments of future skill requirements exhibit a clear trend of simultaneous growth in "digital technology skills + higher-order cognitive skills

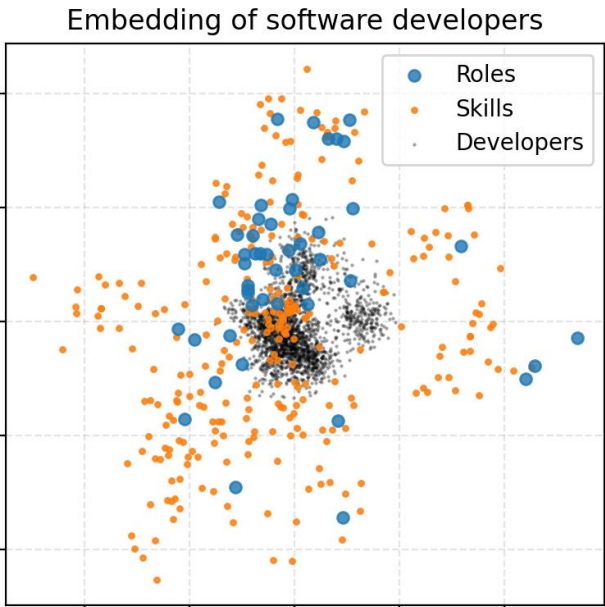


Figure 13: Joint embedding visualization of roles, skills, and candidates

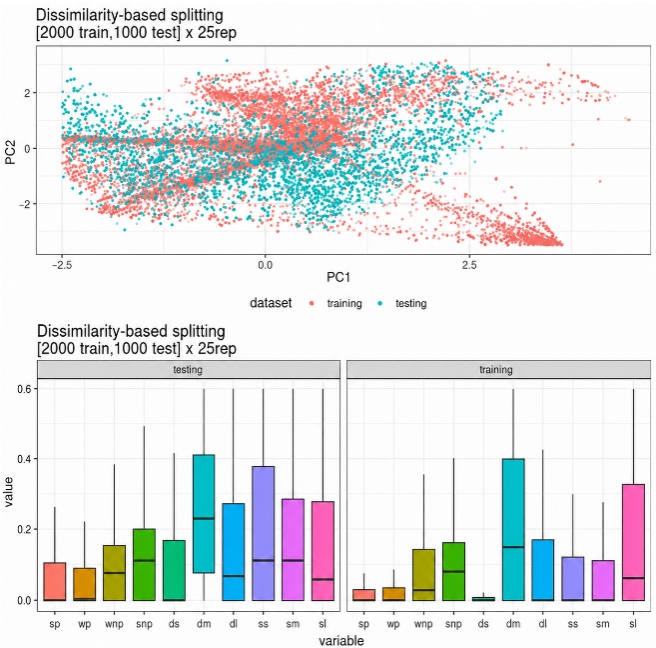


Figure 14: Dissimilarity-based train-test splitting and feature distribution validation

+ career adaptability." Among these, digital skills such as AI and big data, network and cybersecurity, and technical literacy showed the highest net growth, indicating that computer technology is becoming a core driver of changes in the job market's skill structure. Simultaneously, abilities such as creative thinking, resilience and flexibility, curiosity and life-

long learning, leadership, and analytical thinking also showed significant growth, suggesting that the market is not merely requiring graduates to master tools, but rather emphasizing comprehensive abilities such as problem-solving in complex task situations, continuous learning, and cross-job transfer.

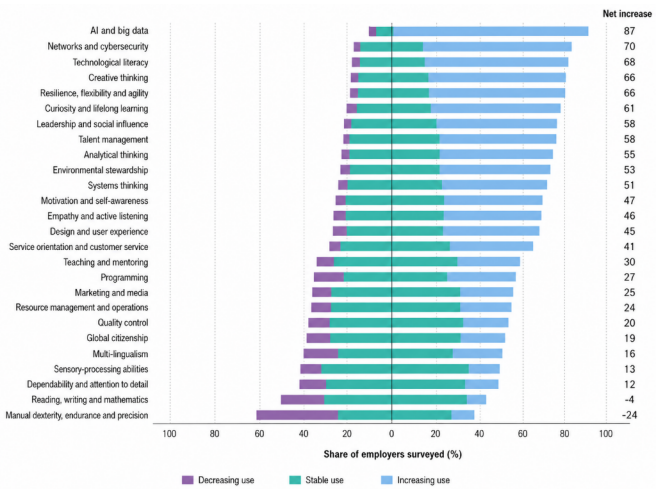


Figure 15: Employer-perceived changes in skill demand and net growth

V. Conclusion

This paper addresses the structural mismatch between "skills supply" and "job demand" in the employment of arts graduates, proposing a comprehensive framework integrating semantic representation, skills modeling, and dynamic analysis. Based on large-scale online recruitment data, the paper first utilizes the Transformer model to achieve deep semantic encoding of recruitment texts, and then constructs a skills classification system through skill embedding, hierarchical clustering, and topic modeling. Building upon this, the paper maps job positions, skills, and individual abilities to a unified semantic space, and combines knowledge graph methods to characterize their structural relationships. Furthermore, it quantifies the skill gap through distribution difference analysis.

Experimental results show that the skills required for arts-related jobs are exhibiting a significant trend towards greater complexity: on the one hand, computer technology skills, data analysis skills, and AI-related skills are rapidly increasing, becoming key factors influencing employment competitiveness; on the other hand, soft skills such as communication skills, problem-solving skills, teamwork, and adaptability remain in high demand. Simultaneously, while the school's training system is relatively close to market demands in terms of basic professional skills, it shows significant deficiencies in digital skills, comprehensive professional abilities, and cross-disciplinary capabilities, validating the effectiveness of the proposed skill gap model.

Furthermore, through semantic space analysis and knowledge graph modeling, the analysis further reveals the struc-

tural relationships and transfer paths among skills, providing an interpretable basis for the personalized career development of arts graduates. The proposed dynamic simulation provides a basis for examining progressive learning pathways informed by skill proximity and changing occupational requirements. However, the supplied Results section does not report a controlled numerical comparison against conventional static skill-completion strategies. Claims of demonstrated superiority therefore require additional experimental evidence and supporting evaluation results.

Overall, this paper not only systematically characterizes the skill demand structure of the art job market from a data-driven perspective, but also achieves unified modeling from skill identification and gap analysis to path optimization, providing theoretical and methodological support for the optimization of university curriculum systems and employment guidance. Future research could incorporate larger-scale cross-regional data, incorporate real-time recruitment dynamics information, and integrate individual learning behavior data to enhance the model's application capabilities in actual job recommendations and educational decision-making.

Conflicts of Interest

The author declares no conflicts of interest concerning the research reported in this manuscript.

Data Availability

The data supporting the reported analyses may be requested from the corresponding author, subject to applicable access requirements and privacy restrictions.

Funding Information

The author declares that no specific funding was received to support this study.

References

- [1] Decorte, J.-J., Van Haute, J., Deleu, J., Develder, C., and Demeester, T. (2023). Career path prediction using resume representation learning and skill-based matching. arXiv preprint, arXiv:2310.15636. <https://arxiv.org/abs/2310.15636>
- [2] World Economic Forum (2025). The Future of Jobs Report 2025. World Economic Forum. <https://www.weforum.org/publications/the-future-of-jobs-report-2025/>
- [3] Lai, J. W., Zhang, L., Sze, C. C., and Lim, F. S. (2025). Learning analytics for bridging the skills gap: A data-driven study of undergraduate aspirations and skills awareness for career preparedness. Education Sciences, 15(1), Article 40. <https://doi.org/10.3390/educsci15010040>
- [4] Brasse, J., Förster, M., Hühn, P., Klier, J., Klier, M., and Moestue, L. (2024). Preparing for the future of work: A novel data-driven approach for the identification of future skills. Journal of Business Economics, 94, 467–500. <https://doi.org/10.1007/s11573-023-01169-1>
- [5] Sun, Y., Ji, Y., Zhu, H., Zhuang, F., He, Q., and Xiong, H. (2025). Market-aware long-term job skill recommendation with explainable deep reinforcement learning. ACM Transactions on Information Systems, 43(2), Article 4998. <https://doi.org/10.1145/3704998>
- [6] Lábaj, M., Oleš, T., and Procházka, G. (2025). Impact of robots and artificial intelligence on labor and skill demand: Evidence from the UK. Eurasian Business Review, 15, 953–1001. <https://doi.org/10.1007/s40821-025-00314-w>
- [7] 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), 1–15. <https://doi.org/10.18653/v1/2024.nlp4hr-1.1>
- [8] Oleš, T. (2026). In-demand skills: A shield against automation—evidence from online job vacancies. *Journal for Labour Market Research*, 60, Article 5. <https://doi.org/10.1186/s12651-026-00424-6>
 - [9] Zhang, M., Jensen, K., Sonniks, S., and Plank, B. (2022). SkillSpan: Hard and soft skill extraction from English job postings. *Proceedings of NAACL-HLT 2022*, 4962–4984. <https://doi.org/10.18653/v1/2022.naacl-main.366>
 - [10] Frazzetto, P., Haq, M. U. U., Fabris, F., and Sperduti, A. (2026). Graph neural networks for candidate-job matching: An inductive learning approach. *Data Science and Engineering*, 11, 360–377. <https://doi.org/10.1007/s41019-025-00293-y>
 - [11] Kavargyris, D. C., Georgiou, K., Papaioannou, E., Moysiadis, T., Mittas, N., and Angelis, L. (2025). Future skills in the GenAI era: A labor market classification system using Kolmogorov–Arnold networks and explainable AI. *Algorithms*, 18(9), Article 554. <https://doi.org/10.3390/a18090554>
 - [12] Çelik Ertuğrul, D., and Bitirim, S. (2025). Job recommender systems: A systematic literature review, applications, open issues, and challenges. *Journal of Big Data*, 12, Article 140. <https://doi.org/10.1186/s40537-025-01173-y>
 - [13] Lee, H., and Hong, I. (2025). Quantifying the influence of vocational education and training with text embedding and similarity-based networks. *PLOS ONE*, 20(8), Article e0329405. <https://doi.org/10.1371/journal.pone.0329405>
 - [14] Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP 2019*, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
 - [15] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint, arXiv:2203.05794*. <https://arxiv.org/abs/2203.05794>
 - [16] Kavas, H., Serra-Vidal, M., and Wanner, L. (2025). Multilingual skill extraction for job vacancy–job seeker matching in knowledge graphs. *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, 146–155. <https://aclanthology.org/2025.genaik-1.15/>
 - [17] Nguyen, K., Zhang, M., Montariol, S., and Bosselut, A. (2024). Rethinking skill extraction in the job market domain using large language models. *Proceedings of NLP4HR 2024*, 27–42. <https://doi.org/10.18653/v1/2024.nlp4hr-1.3>
 - [18] Yang, B., and Shen, Z. (2025). Knowledge graph construction and talent competency prediction for human resource management. *Alexandria Engineering Journal*, 121, 223–235. <https://doi.org/10.1016/j.aej.2025.02.043>
 - [19] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), Article 861. <https://doi.org/10.21105/joss.00861>
 - [20] De Santo, A., Nobani, N., Malandri, L., Mercorio, F., and Mezzanzanica, M. (2026). Using language models to extract and map emerging skills to the ESCO taxonomy. *Discover Artificial Intelligence*, 6, Article 1150. <https://doi.org/10.1007/s44163-026-01609-1>
 - [21] Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. (2018). Modeling relational data with graph convolutional networks. *The Semantic Web (ESWC 2018)*, 593–607. https://doi.org/10.1007/978-3-319-93417-4_38
 - [22] Aubert, B. A., Bourdeau, S., Warin, T., Schneider, J., and Senegas, T. (2026). Digital skills assessment in jobs in Canada. *Journal for Labour Market Research*, 60, Article 18. <https://doi.org/10.1186/s12651-026-00440-6>
 - [23] Wunsch, C., Rochlitz, F., and Arni, P. (2025). Jobseekers’ skills and job search behaviour. *Swiss Journal of Economics and Statistics*, 161, Article 11. <https://doi.org/10.1186/s41937-025-00142-9>
 - [24] Calanca, F., Sayfullina, L., Minkus, L., Wagner, C., and Malmi, E. (2019). Responsible team players wanted: An analysis of soft skill requirements in job advertisements. *EPJ Data Science*, 8, Article 13. <https://doi.org/10.1140/epjds/s13688-019-0190-z>
 - [25] Qin, C., Fang, C., Yao, K., Chen, X., Zhuang, F., and Zhu, H. (2025). COTR: Efficient job task recognition for occupational information systems with class-incremental learning. *ACM Transactions on Management Information Systems*, 16(2), Article 17. <https://doi.org/10.1145/3712306>

...