Translating Job Advertisements into Competency Taxonomy: An Interpretable Approach for Robotics Recruitment Analysis

Zhiyan Xue[1], Yang Zhou[2]*
School of Electronic Information Engineering, Suzhou Polytechnic University, Suzhou, China [1]
School of Management, Suzhou Polytechnic University, Suzhou, China [2]

***Abstract*—To translate robotics recruitment texts into an interpretable competency taxonomy, this study explored how latent topics can be induced from large-scale job advertisements and used to construct a structured competency taxonomy. A domain-specific preprocessing pipeline was applied to a corpus of robotics job advertisements, and latent topics were subsequently induced using Latent Dirichlet Allocation (LDA). Building on the extracted topic evidence, a procedure was applied to organize topic summaries into task domains and competencies grounded in topic keywords. Evaluation was conducted using a set of quantitative measures to assess semantic consistency and structural quality. The results revealed recurring competency patterns encompassing on-site operation and service, engineering design and integration, software and algorithm development, and system verification and reliability assurance. The resulting competency taxonomy captures the underlying structure of employer demand and provides an interpretable basis for robotics skill demand analysis, supporting role profiling and serving as an empirical reference for Vocational Education and Training (VET).**

Keywords*—Competency taxonomy; robotics industry; Latent Dirichlet Allocation (LDA); Vocational Education and Training (VET)

I. Introduction

A. Research Background

Robotics has expanded from laboratory research into broad industrial and service deployments, including manufacturing, logistics, healthcare, and consumer-facing applications. This expansion has diversified robotics job roles and increased demand for talent with cross-disciplinary skill profiles. At the same time, rapid iteration of toolchains and deployment practices shortens the validity period of static skill descriptions. These trends create a practical need for methods that can characterize robotics competency requirements in a scalable and updatable manner [1][2].

Online job advertisements provide a timely and large-scale source of evidence for this purpose. Recruitment texts serve as explicit, employer-authored descriptions of hiring requirements and are typically high in volume, frequently updated, and broad in coverage. As a result, job advertisements can be treated as a near real-time signal of skill demand, enabling analysis of the structure of required competencies and their evolution across job families and time windows [3][4]. Compared with interviews or small-sample approaches, recruitment texts support repeatable analysis at scale, which is particularly valuable in fast-changing technical domains [5][6]. Robotics recruitment introduces additional complexity that makes competency modeling nontrivial. Robotics roles commonly combine capabilities spanning electromechanical systems, software control, as well as operation and maintenance. Skill evidence is distributed across a chain of job families that involve mechanical design, electrical engineering, embedded development, systems integration, and on-site support. Consequently, skill expressions in job advertisements are fragmented, terminology is not standardized across enterprises, and the granularity of requirements varies from broad capability statements to tool-specific and task-level descriptions [7][8]. Traditional interview-based or small-sample approaches often fail to yield a competency structure that generalizes across diverse job roles [9][10][11].

Against this background, two research questions guide the present work. What latent competency domains and skill patterns are reflected in robotics recruitment texts? How can unsupervised topic evidence from recruitment texts be translated into an interpretable and well-structured competency taxonomy?

To address these two questions, an approach is proposed that translates job advertisements into a structured competency taxonomy. The approach proceeds from recruitment text preprocessing to LDA-based topic induction, followed by translation from topic summaries to task domains and competencies grounded in topic keywords. Evaluation is designed for unlabeled settings and emphasizes interpretability and structural validity. Topic coherence is used to assess the semantic consistency of topic summaries, and a set of quantitative measures is used to quantify representativeness, noise, and redundancy of the induced taxonomy. Together, these components provide an interpretable basis for robotics skill demand mining from job advertisements, while supporting comparison against standard baselines in applied text mining.

B. Related Work

With the rapid development of the robotics industry, Vocational Education and Training (VET) systems have been under sustained pressure to update talent cultivation promptly. Existing studies have shown that educational and training responses often lag behind technological progress and shifts in market demand. Accordingly, scholars have proposed a range of strategies at different levels to improve talent development

*Corresponding author.

in robotics-related fields [12][13][14][15]. Existing studies have mainly approached this issue from two closely related directions. The first strand focuses on curriculum systems and talent cultivation models. Studies in this line of research primarily examine curriculum design, training mechanisms, and pathways for industry-education integration in robotics-related programs to enhance the alignment between educational provision and industrial demand. For example, Wang et al. pointed out that school-enterprise cooperation is an effective approach to improving teaching quality, thereby highlighting industrial demand as an important reference for curriculum design and educational improvement [12].

The second strand concentrates on specific skill demands and occupational competency requirements. This line of research mainly analyzes particular technical domains or concrete occupational contexts. Specifically, Yu-Shen et al. [13] focused on specific skill demands in the robotics field; Olukanni et al. [14] further situated the analysis in concrete work settings, such as human-robot collaboration, and Do et al. [15] through an Importance-Performance Analysis (IPA), suggested that university curricula should place greater emphasis on courses such as Artificial Intelligence, Machine Vision, and Robot Structure Design. These studies have helped identify key competency dimensions required in robotics-related occupations and have provided relatively direct practical implications for the optimization of course content.

Overall, existing research on robotics talent cultivation and competency development has largely proceeded from the supply side, with emphasis on educational planning and curriculum reform. While such studies are valuable for defining training objectives and competency expectations, they provide only limited insight into how firms specify skill requirements in actual recruitment and how these requirements evolve with industrial change. In parallel, studies addressing specific robotics skills or competency dimensions often focus on a single technical direction or a limited set of occupational contexts. Sample scopes are therefore typically constrained, and the resulting outputs frequently take the form of partial skill lists tied to local tasks or particular positions. Such a pattern makes it difficult to cover the major upstream and downstream positions across the robotics industrial chain.

Beyond the substantive findings of robotics-specific studies, a related body of research has approached the issue from a methodological perspective by examining how occupational skills can be identified, constructed, and organized. From this perspective, existing studies can be broadly grouped into three methodological approaches. The first perspective relies on expert-based methods, such as Delphi studies, expert interviews, and literature synthesis [15][16]. A notable strength of this line of work is that the resulting frameworks are usually grounded in expert understanding of occupational practice and therefore tend to remain highly interpretable and closely aligned with domain knowledge [17]. However, such methods are often costly to update and dependent on expert judgment. Consequently, they are well suited to normative framework construction, but less suited to tracking the changing skill requirements of large numbers of job advertisements. The second perspective uses recruitment texts as a demand evidence source for skill analysis. Online job advertisements can be treated as employer demand statements that contain explicit signals about skills, work activities, and job requirements [18][19]. Some studies have used rule extraction and statistical representation tools to identify terms from job advertisements [20]. Findings from this stream of research indicate that recruitment texts can support large-scale labor market and skill analysis and can reveal meaningful patterns in employer demand. However, the resulting outputs often remain fragmented because they are typically expressed as keyword lists. The third perspective extends recruitment text analysis toward the discovery of latent structure. Topic modeling, including LDA and other related approaches, has been used to identify recurring themes [21]. In recruitment analytics, these methods have been applied to domains such as big data analytics and software engineering to derive topic-based representations of knowledge domains, skill combinations [22][23][24]. Compared with direct term extraction, topic modeling provides a more effective basis for organizing dispersed skill expressions into coherent thematic structures. Nevertheless, a topic provides a meaningful set of salient terms, but it does not explain how those terms should be translated into hierarchical competency domains [25]. In other words, existing recruitment text studies have shown that skill signals can be extracted and that latent themes can be discovered, yet they have not fully resolved how topic-level evidence should be converted into a competency structure.

Building on the foregoing methodological discussion, a further challenge concerns how the outcomes of skill and competency induction should be evaluated. This issue has received relatively limited attention, largely because competency extraction from recruitment text is often conducted in the absence of standardized labels or benchmark datasets [26]. Soare [27] and McClelland [28] relied on expert-curated competency frameworks and skill taxonomies as references for workforce planning and training design, which provided guidance on how competency structures could be organized but did not yield direct evaluation criteria. Hoyle et al. [29] adopted topic coherence as an intrinsic indicator of semantic consistency for topic word sets, which offered a practical way to judge whether a topic summary formed a coherent group of terms. Kim et al. [30] further suggested that representativeness and distinctiveness were essential for reusable and computationally meaningful competency structures, which implied that coherence alone was insufficient because it did not assess document-level representativeness or cross-topic redundancy. This line of work motivated evaluation protocols that complement coherence with structure-oriented indicators, yet a unified evaluation approach for determining whether coherent topic summaries translate into verifiable and non-redundant competency units remains absent.

The above literature highlighted two gaps that motivated this study. First, in the robotics domain, existing studies on talent development and skill requirements have largely proceeded from the supply side and have not yet established a demand-side competency structure grounded in large-scale recruitment evidence across the major positions of the robotics industry. Second, although existing recruitment text studies have identified skill signals and uncovered latent topics, these outputs have not formed a competency taxonomy. More

importantly, approaches for translating topic evidence into an interpretable competency structure have remained underdeveloped. This study addressed these gaps by introducing a translation approach and evaluation metrics to enable interpretable competency induction in robotics recruitment analysis.

The structure of this study is organized as follows. Section II describes the method, including data collection, preprocessing, topic discovery using LDA, and translation from topic summaries to a competency taxonomy. Section III introduces the experimental setup, baseline, and evaluation metrics. Section IV reports results, including both quantitative findings and qualitative case analyses. Section V concludes with limitations and future directions.

II. Methodology

This study proposed a structured pipeline that transforms unstructured job advertisements into an interpretable competency taxonomy. The process consists of five core stages: 1) Data Collection; 2) Data Preparation; 3) Topic Discovery; 4) Topic-to-Competency Translation; 5) Quantitative Evaluation. This pipeline takes raw job recruitment texts as input and produces a competency taxonomy as output. Fig. 1 illustrates the overall framework.

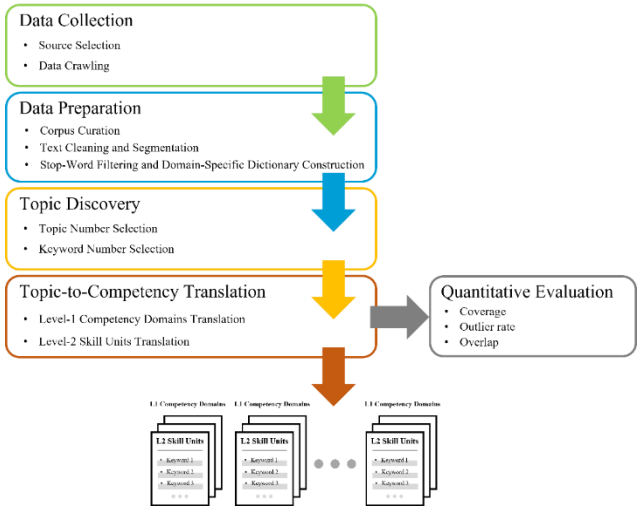

Fig. 1. The proposed study framework.

A. Data Collection

1) Source selection: Robotics job advertisements are distributed across multiple online platforms and often exhibit inconsistent formatting and varying levels of detail. To balance market representativeness with textual suitability for systematic analysis, the study selects Zhaopin (www.zhaopin.com) as the data source for three considerations. First, Zhaopin is one of the leading online recruitment platforms in China, with broad geographic coverage and substantial posting volume. Its market presence facilitates the collection of a diverse sample of robotics-related positions and helps reduce potential selection bias. Second, job advertisements on Zhaopin tend to follow a relatively standardized structure and are subject to platform-level quality control, which makes them well suited for natural language processing (NLP). In particular, the platform encourages employers to provide detailed job description texts that explicitly separate "position responsibilities" and "job requirements", offering a high-quality corpus for extracting skill-related signals. Third, Zhaopin data have been used in prior peer-reviewed studies to analyze labor market demand patterns, providing an external validity signal for leveraging its job description texts in skill demand mining [31][32]. To reduce temporal drift and ensure that the induced topics reflect a consistent market snapshot, the study restricts data collection to a fixed sampling window spanning December 2025 to January 2026.

2) Data collection: Selenium was adopted as the primary tool for crawling and parsing job advertisement pages in this phase. It is a widely used automation framework that programmatically controls real web browsers, enabling reliable interaction with web pages [33][34]. To target robotics-related vacancies while maintaining regional diversity, the study queried the platform using the keywords "Robotics" and "Robot", which aim to capture openings across the robotics value chain, including R&D, system integration, and maintenance. Search locations were restricted to six major regions in China: Beijing, Jiangsu, Shanghai, Zhejiang, Guangdong, and Shandong Province. This location set was chosen because these regions collectively represent a large share of the national robotics industrial chain and host a high concentration of industrial robot manufacturing enterprises. For each location, the crawler iterated over search result pages, collected posting URLs, and parsed the corresponding detail pages to extract the job title, full job description text, and other data such as company name and salary. The initial crawl yielded 6,802 advertisements, reflecting the current labor demand for robotics-related positions in the aforementioned cities.

B. Data Preparation

1) Corpus curation: At this stage, the study curated a corpus of valid robotics-related positions to reduce noise that might interfere with downstream topic discovery. The advertisements were curated through three steps: removing invalid or incomplete records, excluding irrelevant advertisements such as finance jobs, administrative jobs, and promotional or advertisement entries, and eliminating exact duplicate advertisements. After cleaning and deduplication, the resulting corpus contains 3,290 job descriptions and serves as the input to subsequent preprocessing.

2) Text cleaning and segmentation: Recruitment texts often contain punctuation noise, irregular symbols, and platform-specific artifacts that are not informative for skill demand mining. A cleaning procedure is performed before segmentation. Specifically, Python regular expressions are applied via re.sub() to remove punctuation marks and abnormal characters and to normalize whitespace. Additionally, unlike English processing pipelines, Chinese

text does not require uppercase or lowercase conversion, lemmatization, or stemming; therefore, emphasis is placed on accurate word segmentation and term normalization as the primary preprocessing steps.

The corpus is then segmented using Jieba, a widely used Chinese word segmentation toolkit that supports dictionary-based tokenization through prefix-dictionary matching and statistical disambiguation. Jieba is selected for its robustness on large-scale web text, its ability to incorporate a lexicon to preserve robotics technical terms and vendor names, and its lightweight, reproducible integration in standard Python environments, which facilitates transparent replication of the preprocessing pipeline [35].

3) Stop-word filtering and domain-specific dictionary construction: Recruitment corpora contain a substantial amount of platform-specific and HR-related expressions. To mitigate this effect, stop-word filtering is performed using an expanded stop-word set based on the widely used Harbin Institute of Technology stop-word dictionary (hit_stopwords), which is commonly adopted in Chinese NLP studies due to its broad coverage of high-frequency non-content terms [36]. Building on this foundation, the stop-word set is further extended with recruitment-related terms that are prevalent in job advertisements but provide limited value for skill demand analysis. Typical examples include template headers and HR-related expressions such as “compensation and benefits”, “job responsibilities”, “job requirements/qualifications”, and “bonus points/plus”. Removing such terms reduces the risk that topic discovery is dominated by generic recruitment language rather than domain skills.

In addition to stop-word filtering, a domain-specific user dictionary is constructed to preserve robotics-related terms that might otherwise be incorrectly split or discarded. The dictionary covers two major categories. Robotics brands and company names are included, such as Siemens, KUKA, ABB, and FANUC. Technical terms are added to capture core skills, including key algorithms and platforms such as SLAM, ROS, OpenCV, and multimodal learning, as well as industrial automation and engineering terms such as CAD, AGV, Robot Studio, and robotic arm.

By integrating this user dictionary into the segmentation process, domain terms are more consistently tokenized as single units, reducing vocabulary sparsity and improving the stability and interpretability of the downstream topics and competency taxonomy.

C. Topic Discovery

LDA was employed to uncover latent competency topics from the robotics recruitment texts. Its selection was primarily guided by two considerations. First, the purpose of the study was to explore latent competency structures embedded in recruitment texts rather than to assign documents to predefined categories. Under this objective, an unsupervised topic modeling method was methodologically appropriate [37]. Second, given the scale of the corpus and the practical constraints on computational resources, LDA offered advantages in terms of computational efficiency and implementation feasibility. For these reasons, LDA was adopted in the present study. LDA represents each document as a mixture of latent topics, where each topic is modeled as a probability distribution over the vocabulary, thereby facilitating the extraction of interpretable themes representing skill domains and recurring requirements. Given K topics, the model estimates a topic-word distribution $\emptyset_K$ for each topic t_k ($k = 1, \dots, K$) and a document-topic distribution θ_i for each document d_i ($i = 1, \dots, N$) . These distributions provide a ranked list of words that characterize each topic.

Model training is implemented using the Scikit-learn (sklearn) module. Scikit-learn is a widely used open-source machine learning library in Python that provides consistent APIs for data preprocessing, model training, and evaluation across a broad range of algorithms, including topic models. Its LDA implementation offers practical advantages for applied studies, such as integration with standard pipelines and built-in metrics.

1) Topic number selection: In practice, the topic number K is determined by jointly considering topic coherence and perplexity to balance interpretability and model fit. This criterion helps mitigate the risk of underfitting (overly coarse topics) and overfitting (overly fragmented topics), yielding topics that remain interpretable while retaining reasonable explanatory power on data.

2) Keyword number selection: For each topic t_k, the Top-M keywords with the highest probabilities are extracted as an interpretable topic summary, which serves as a compact and easily understandable representation of the topic and is also used as the basis for downstream topic-to-competency translation and structural evaluation.

M is set to 30 as a balanced choice that trades off information coverage against noise introduction. When M is too small, secondary but still important skill terms may be omitted, causing topics to be under-expressed and reducing representational adequacy for recruitment competencies. When M is too large, long-tail and weakly related words are more likely to enter the summary, which can increase overlap between topics and weaken discriminability between topics.

D. Topic-to-Competency Translation

Topic keywords provide concise summaries, but do not directly constitute a competency taxonomy. To bridge this gap, each discovered topic is translated into a two-level structure consisting of task domains and competencies. The translation aims to preserve the semantic content of each topic while producing a structured representation that remains interpretable, traceable, and suitable for subsequent evaluation.

The translation is conducted independently for each topic, based on its ranked Top-M keyword list. No terms outside the current topic are introduced during translation. This constraint preserves topic discriminability and avoids artificial gains in coverage caused by borrowing terms from other topics. In addition, the procedure prioritizes representativeness, meaning that the resulting competency structure is expected to retain the dominant semantic content of the topic rather than rely on a

small number of isolated keywords. Redundancy control is also enforced, such that different competencies under the same task domain capture distinct functional elements rather than repeated variants of the same meaning. The procedure consists of five steps.

Step 1: Assign the task domain. The ranked Top-M keyword list of each topic is first examined as a whole in order to identify its dominant capability orientation. The task domain is then assigned as a task domain-level descriptor that summarizes the central functional area implied by the topic keywords. This naming process follows established robotics and engineering terminology whenever possible, so that the resulting label remains professionally interpretable and consistent with domain usage. The task domain label is required to reflect the main semantic center of the topic rather than a peripheral or overly generic aspect of the keyword set.

Step 2: Partition keywords into candidate competency subsets. After the task domain is determined, the same Top-M keyword list is partitioned into several candidate competency subsets representing fine-grained functional elements. Each subset is intended to capture one operationally meaningful competency, such as design, programming, debugging, testing, deployment, or maintenance-related capability. The grouping process remains restricted to the current topic and organizes keywords according to semantic coherence and task domain relevance. Composite or overly broad groupings are avoided when the underlying competencies can be separated into more precise units.

Step 3: Refine competency boundaries. The candidate competency subsets are then refined by prioritizing higher-ranked keywords, as these terms carry the strongest evidence of the semantic core of the task domain. Particular attention is paid to preserving the dominant task domain-level signals, rather than allowing the translation to be driven by lower-ranked or weakly informative terms. During this step, the boundaries between competencies are adjusted so that each competency retains a clear functional focus within the task domain. When two competencies exhibit substantial semantic overlap, they are merged or revised. When a competency combines multiple distinct functional elements, it is further divided into more fine-grained competencies.

Step 4: Check evidence retention and reduce redundancy. The provisional competency-level structure is then evaluated against the full Top-M keyword list of the topic. This audit examines whether the main topic evidence is retained, whether important high-ranking keywords remain unassigned to the induced structure, and whether different competencies reuse the same semantic content to an excessive degree. If substantial unassigned keywords are identified, especially among the higher-ranked terms, the partition is revised to improve representativeness. If strong overlap is observed between competencies, the structure is adjusted to improve non-redundancy. This step ensures that the translation does not become a purely naming process detached from the original topic evidence.

Step 5: Finalize Competency labels. Once the keyword partition reaches an acceptable balance between representativeness and redundancy, each subset is formalized as a competency. The corresponding competency label is expressed as a concise operational capability statement grounded in the assigned keyword subset. The final output for each topic therefore consists of one task domain, several competencies, and an explicit mapping from each competency to its supporting topic-specific keywords. This mapping preserves traceability between the induced competency structure and the original topic summary, thereby enabling subsequent audit of evidence retention and structural consistency.

Through this procedure, topic summaries are transformed into a two-level competency taxonomy. The resulting structure does not aim to establish a complete ontology of robotics competencies. Instead, it organizes the underlying topic evidence into a two-level representation consisting of task domains and competencies under unlabeled conditions.

E. Quantitative Evaluation

Quantitative evaluation is conducted to assess the interpretability and structural validity of topic summaries in an unsupervised setting without standardized labels. The evaluation protocol is based on the Top-M keyword summary of each topic. Specifically, three metrics are considered: coverage, outlier rate, and overlap.

1) Coverage: Coverage evaluates whether the Top-M keyword summary is representative of the documents assigned to a topic. Let C_k denote the set of documents associated with topic k, and let T_k denote its Top-M keyword set. Let $W(d)$ be the set of tokens appearing in document d. Document-level coverage for topic k is defined as:

$$\mathrm{Cov}(k) = \frac{1}{|C_k|}\sum_{d\in C_k} I\left(W(d)\cap T_k \neq \emptyset\right) \quad (1)$$

where, $I(\cdot)$ is the indicator function. The corpus-level coverage is computed as a weighted average:

$$\mathrm{Cov} = \sum_{k=1}^{K}\frac{|C_k|}{N}\mathrm{Cov}(k) \quad (2)$$

Cov(k) quantifies the proportion of job descriptions in cluster k that contain at least one representative keyword from T_k. Higher coverage indicates that the keyword summary better reflects the cluster content.

As matching on a single keyword can be an insufficiently strict criterion, a stricter variant is also considered, specifically the thresholded coverage. For a threshold $\tau \geq 2$, thresholded coverage requires that a document match at least τ keywords from T_k:

$$\mathrm{Cov}_\tau(k) = \frac{1}{|C_k|}\sum_{d\in C_k} I\left(|W(d)\cap T_k| \geq \tau\right) \quad (3)$$

$$\mathrm{Cov}_\tau = \sum_{k=1}^{K}\frac{|C_k|}{N}\mathrm{Cov}_\tau(k) \quad (4)$$

This variant strengthens the representativeness requirement and reduces the possibility of inflated coverage caused by occasional generic term matches.

2) Outlier rate: The outlier rate quantifies the proportion of documents not explained by the topic summary under the same criterion:

$$\mathrm{Out}(k) = 1 - \mathrm{Cov}(k) \quad (5)$$

The corpus-level outlier rate is computed as:

$$\mathrm{Out} = 1 - \mathrm{Cov} \quad (6)$$

3) Overlap: Overlap measures discriminability by quantifying redundancy among topic summaries, using the Jaccard similarity between keyword sets T_i and T_j:

$$\mathrm{Jacc}(i,j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|} \quad (7)$$

The overall overlap is defined as the mean pairwise similarity:

$$\mathrm{Overlap} = \frac{2}{K(K-1)} \sum_{1 \le i < j \le k} \mathrm{Jacc}(i,j) \quad (8)$$

The Jaccard similarity is a widely used metric in the analysis of set-based data [38]. It measures the proportion of shared elements relative to the union of sets, providing an interpretable quantification of similarity or redundancy. Its extensive application in computational linguistics and topic analysis establishes it as a reliable metric for evaluating the distinctiveness of topic keyword sets. Lower overlap indicates that different topics are characterized by more distinct keyword summaries, improving interpretability and reducing ambiguity in topic-to-competency translation. Because recruitment texts may share unavoidable common terms, overlap is interpreted jointly with coherence and coverage.

The proposed metrics jointly capture representativeness and discriminability aspects of topic outputs. It is important to clarify that these metrics do not aim to judge the linguistic aesthetics of competency labels, nor do they evaluate clustering quality in a supervised sense. Instead, the evaluation focuses on the structural validity of the evidence used for translation. At the task domain-level, the evaluation examines whether the Top-M topic summaries provide a stable, representative, and discriminative basis for inducing competencies. Coverage and outlier rate quantify how well the summary aligns with documents assigned to a topic in terms of representativeness and noise, and overlap indicates how distinguishable different topic summaries are by capturing redundancy across topics. At the competency level, the same validation logic is extended to the internal audit of the induced competencies. In this stage, the evaluation examines whether the translated competencies retain the dominant topic evidence and whether the decomposition remains internally coherent with limited redundancy. Collectively, these measures assess not only whether the topic-derived evidence is reliable enough to support competency induction, but also whether the resulting topic-to-competency translation remains traceable and interpretable under unlabeled conditions.

Topic summaries, metrics, and translation details are discussed in the Appendix.

III. Experimental Setup

This section describes the experimental setup used to evaluate the proposed translation framework. A baseline method is introduced to provide a conventional reference point, enabling a quantitative comparison that highlights the effectiveness and interpretability of the proposed approach. The baseline serves to contextualize the results and to demonstrate that observed improvements are attributable to the design of the proposed method rather than to trivial differences in data or partitioning.

A. Experimental Protocol

This section describes the evaluation protocol used to assess both the quality of topic evidence and the structural validity of the downstream competency translation. The protocol operates at two complementary levels. The first level evaluates topic summaries as the evidence basis for task domain induction, while the second level audits the resulting competencies derived from the proposed topic-to-competency translation procedure. This design is necessary because coherent topic summaries do not inherently guarantee that the resulting competency decomposition remains evidence-grounded, non-redundant, and traceable.

At the task domain-level, the protocol compares LDA topic discovery with a classical document clustering baseline under matched granularity and a unified interpretability interface. Comparability is enforced through four controls. First, all approaches operate on the same cleaned corpus and follow the same preprocessing and vocabulary construction procedure. Second, the number of topics or clusters is fixed to K across methods. Third, each topic or cluster is represented by a Top-M keyword set under a unified summary length. Fourth, all approaches produce document assignments and keyword summaries that are evaluated using the same metrics. Under this design, coverage, outlier rate, and overlap are used to assess whether the resulting summaries provide a representative and discriminative evidence interface for subsequent competency translation.

At the competency level, the evaluation shifts from topic summaries to the internal structure of the induced competency taxonomy. This level focuses on the transformation from topic evidence to a two-level representation consisting of task domains and competencies. Because the TF-IDF + K-Means baseline does not produce an explicit competency decomposition, a competency-level audit does not apply to the baseline and is therefore conducted only on the proposed LDA-based taxonomy. This audit examines whether the competencies preserve the dominant topic evidence and whether the decomposition remains structurally coherent within each task domain. Specifically, within-topic coverage is used to assess whether the highest-ranked topic keywords are retained in the competency structure; within-topic outliers are used to identify evidence loss during translation; and within-topic overlap among competency keyword subsets is used to examine internal redundancy. Taken together, the task domain-level comparison and the competency-level audit form a complementary validation protocol that evaluates not only the suitability of the topic evidence but also the interpretability and structural validity of the resulting task domain-competency hierarchy.

A traditional text clustering baseline is implemented using TF-IDF document representation followed by K-Means clustering. Based on the same preprocessed recruitment texts, each advertisement is represented as a TF-IDF vector and

partitioned into K clusters, where K corresponds to the number of topics used in the LDA model for direct comparison. TF-IDF is a standard weighting scheme in information retrieval and text mining. It assigns higher weights to terms that are frequent within a document but relatively less frequent across the corpus, which helps emphasize terms that are more discriminative for a given advertisement. In this study, TF-IDF vectors are constructed under a shared vocabulary derived from the same preprocessing pipeline. K-Means is a widely used unsupervised clustering method that partitions a set of vectors into K clusters by minimizing within-cluster dispersion. It iteratively assigns documents to cluster centers and updates the centers until the objective stabilizes. For each cluster, a fixed-length set of representative keywords is extracted from the cluster centroid. The centroid is obtained by aggregating the TF-IDF vectors of documents assigned to the cluster, and the Top-M terms with the highest centroid weights are selected as the cluster keywords. This centroid-based keyword extraction provides an interpretable summary for each cluster under the same Top-M interface used for topic summaries, enabling a consistent comparison across methods. This baseline therefore serves as a conventional reference for examining whether the evidence provides a more suitable basis for downstream competency translation than a classical clustering summary, rather than representing an alternative implementation of the proposed framework.

B. Implementation Details

Experiments are conducted on the robotics job advertisements corpus. The web crawling stage retrieved 6,802 advertisements, and 3,290 advertisements were retained after cleaning and deduplication. All methods operate on the same cleaned corpus to ensure that observed differences are attributable to modeling choices rather than data handling.

Latent competency topics were induced using the LatentDirichletAllocation class in scikit-learn. Before model training, the corpus was represented in document-term form, where min_df = 5 was used to remove extremely rare terms, and max_df = 0.9 was used to exclude overly common terms. The number of topics was specified as K, and the LDA hyperparameters were determined with reference to the expected concentration of the latent distributions and the practical requirements of model implementation. Specifically, the doc_topic_prior was set to $\alpha=50/K$, and the topic_word_prior was set to $\beta=0.01$. These settings were adopted as empirical parameter settings reported in prior LDA applications [39]. The model was trained with learning_method = 'batch', under which all documents in the corpus were used in each variational EM update, and random_state = 42 was fixed to ensure reproducibility. In addition, max_iter = 800 was specified as a sufficiently large upper bound on the number of training iterations.

For the TF-IDF representations used in the clustering baseline, vocabulary construction applied the same document frequency thresholds as in the LDA analysis, namely min_df = 5 and max_df = 0.9. The resulting shared vocabulary was used to construct TF-IDF vectors for clustering and to support comparable keyword summaries across methods. The K-Means baseline was configured with n_init = 11 and max_iter = 800. Multiple initializations were used to reduce sensitivity to local optima in sparse high-dimensional text representations. These settings were aligned with those used in the LDA analysis to facilitate controlled comparison across methods.

All experiments were set up on Windows 10 with Python 3.13.5. The implementation used Scikit-learn 1.6.1, NumPy 2.1.3, Pandas 2.2.3, Selenium 4.20.0, and Jieba 0.42.1.

IV. Results

A. Topic Discovery and Translation

The LDA model yields a set of competency-related topics, each summarized by a ranked list of top words derived from the topic-word distributions. To support subsequent competency translation, the number of topics is selected by jointly considering quantitative diagnostics and interpretability requirements. Following standard LDA practice, topic quality is evaluated using coherence and perplexity. Coherence quantifies the semantic consistency of the top words within each topic, based on word co-occurrence statistics computed from the tokenized corpus. Perplexity measures the model's probabilistic fit on holdout documents. Fig. 2 reports both criteria across a range of topic numbers.

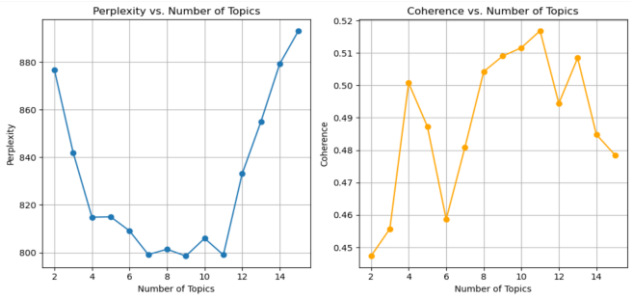

Fig. 2. Coherence and perplexity scores across topic numbers.

From a quantitative perspective, K=11 corresponded to the highest coherence score among the tested values, while perplexity remained within a relatively stable low range for $K \in [7,11]$. Taken together, these results indicate that the 11-topic model offered a suitable trade-off between topic interpretability and model fit among the candidate solutions. This suggests that the selected model not only fits the corpus adequately but also produces topic structures that are more coherent and more readily interpretable. From a modeling and application perspective, this level of granularity aligns with major robotics task domains and supports a two-level representation consisting of task domains and competencies. Fewer topics tend to merge heterogeneous skill signals, reducing discriminability at the task domain-level and complicating competency naming. More topics can fragment coherent domains into overly specific themes, increasing redundancy at the task domain-level and making the downstream competency decomposition less stable. The baseline comparison is carried out with the same partition count to ensure a fair evaluation across methods.

Based on this configuration, the topics are summarized by their top keywords and then translated into a task domain-competency taxonomy. Table I presents the resulting task

domains and competencies, including the top 30 keywords for each topic.

TABLE I. TOPIC-TO-COMPETENCY TRANSLATION RESULT

Task Domains	Competency	Top-30 Keywords
#1 Industrial On-site Equipment Maintenance and Malfunction Diagnosis	1. Troubleshooting and Repair of Automated Equipment Malfunction	Equipment, Fault, Automation, Team, Issue, Mechanical, Technology, System, Collaboration, Fundamentals, Electrical, Production, AGV, Logging, Content, Data, Troubleshooting, Task, Problem Solving, Awareness, Professional Spirit, Skill, Optimization, Learning Ability, Testing, Teamwork, On-Site, Industrial, Sensor, Principle
	2. On-site Process Data Logging and Analysis	
	3. Teamwork and Problem Solving	
	4. Mechatronic System Principles Understanding and Technical Optimization	
	5. Industrial Sensor Testing and Equipment Maintenance	
#2 Mechanical Structure Design and Engineering Implementation	1. Mechanical Structure Design and Analysis	Design, Mechanical, Software, Process, Structure, Product, Mechanical Design, Analysis, Structural Design, Material, Selection, Automation, Machining, Technology, Equipment, CAD, Drive System, Issue, Production, Mechanism, Drawings, Principle, Work Experience, Project, Solution, Mechatronic, Optimization, Motor, Collaboration, Engineering
	2. Automation and Electromechanical Drive Systems Optimization	
	3. Material Selection and Manufacturing Process Planning	
	4. CAD Drafting and Design Documentation Output	
	5. Project Collaboration and Resolution of Product Technical Issues	
#3 Full Lifecycle Project Management and Operational Execution	1. Project Planning and Documentation Management	Product, Project, Technology, Team, Management, Industry, Customer, Quality, Analysis, Solution, Company, Process Flow, Issue, User, Market, Plan, Deployment, Project Management, Intelligent, Collaboration, Department, Documentation, Resources, Planning, Objective, Business, Design, Process, Solution, Automation
	2. Customer Requirements Analysis and Solution Design	
	3. Cross-Departmental Collaboration and Process Improvement Execution	
	4. Product Implementation Delivery and Technical Issue Resolution	
	5. Industry Resource Integration and Requirement Fulfillment	
#4 Robotics Core Module Selection Design and Testing	1. Circuit Design and Electronic Component Selection	Hardware, Design, Motor, Circuit, Electronics, Structure, Joint, Product, Testing, Module, Wiring Harness, Production, Team, Mechanical, Professional Spirit, Issue, Laboratory, Hands-On, Component, Attitude, Quality, Mindset, Power Supply, System, Circuit Design, Collaboration, Selection, Fundamentals, Management, Tooling
	2. Motor and Joint Module Integration with Wiring Harnesses	
	3. Hardware Prototyping, Testing, and Verification	
	4. Practical Engineering Skills and Team Collaboration	
#5 Electrical Design and System Integration for customized Automation	1. PLC Programming and System Software Debugging	Electrical, Automation, PLC, Project, Equipment, Programming, Design, Program, Customized, Software, Selection, Industrial, On-Site, Work Experience, Technology, System, Mechatronic, Electrical Design, Drawings, Vision, Robot Manipulator, Schematic Diagram, Communication, Principle, Wiring, Industry, Mitsubishi, Drafting, CAD, Signal
	2. CAD Drafting and Electrical Schematic Interpretation	
	3. On-site Wiring, Commissioning, and System Integration	
	4. Industrial Communication Configuration and Electromechanical Joint Debugging	
	5. Vision and Robot Manipulator Application in Custom Automation Projects	
#6 Embedded Hardware and Driver Software Development for Robotics	1. Embedded Linux Driver Programming and Development	Testing, System, Software, Driver, Automation, Design, Documentation, Programming, Sensor, Hardware, Embedded, Linux, Analysis, Technology, Issue, Computer, Project, Platform, Product, Motion, Functionality, Module, Principle, ROS, Python, Code, Electronics, Work Experience, Tooling, Motor
	2. Sensor and Motion Control Module Development	
	3. Test Analysis and Fault Localization	
	4. ROS System Design and Application Development	
	5. Project Documentation and Product Deliverable Handover	
#7 Robotics Algorithm Development and Deployment	1. Trajectory Planning and Control Algorithm Development	Algorithm, Planning, Model, Motion, Optimization, Data, Project, Technology, Python, Robotic Arm, Deep Learning, Framework, Robot, ROS, Kinematics, Deployment, Path, Domain, Real-World, Vision, Fundamentals, Control Algorithm, Slam, Environment, Scenario, AI, System, Mathematics, Training, Dynamics
	2. Deep Learning Model Training and Optimization	
	3. SLAM and Environmental Perception Algorithm Implementation	
	4. Algorithm Integration and Deployment in ROS Framework	
#8 Robotic System Performance Optimization and Reliability Testing	1. System Performance and Parameter Testing and Optimization	Optimization, Testing, Design, System, Technology, Performance, Issue, Product, Analysis, Solution, Data, Functionality, Documentation, Hardware, Stability, Reliability, Motion, Evaluation, Coordination, Module, Efficiency, Scenario, Team, Sensor, Planning, Process, Parameter, Review, Environment, Core
	2. Reliability-oriented Design and Stability Validation	
	3. Scenario-based Testing for Motion Modules and Sensors	
	4. Product Function Analysis and Test Documentation Preparation	
	5. Team Coordination and Efficiency-oriented Planning	
#9 Industrial Robot Process Programming and	1. Mainstream Industrial Robot Controllers Programming	Programming, Process, ABB, Industrial, Project, KUKA, Automation, FANUC, Work Experience, On-Site, System, Program, Automotive,
	2. Specific Process Application and Parameter Configuration	

Commissioning	3. Teamwork and Responsibility Awareness	YASKAWA, Brand, Offline, Sense Of Responsibility, Arc Welding, Mechatronic, Parameter, Configuration, Optimization, Spot Welding, Teamwork, Laser, Body-In-White, Responsibility, Mechanical, Trajectory, Dispensing
#10 On-site Technical Support and After-sales Service	1. On-site Emergency Fault Diagnosis and Handling	On-Site, Customer, Project, Equipment, Technical Support, Issue, Product, Process, Production, Electrical, Company, After-Sales, Mechanical, Personnel, Solution, Technology, Process, Materials, Fault, Automation, Service, Organization, Guidance, Management, Optimization, Aspects, Tooling, File, Client, Department
	2. Customer Communication and Technical Service Provision	
	3. Materials Organization and After-sales Documentation Management	
	4. Project Service Delivery and Resource Coordination	
#11 Machine Vision Inspection System Solution Implementation	1. Image Processing Algorithm Programming	Vision, System, Machine Vision, Software, Project, Image Processing, Inspection, Hardware, Automation, Industrial, Design, Technology, Team, Algorithm, Selection, Customer, Light Source, Computer, Lens, Equipment, Analysis, Programming, Solution, OpenCV, Evaluation, Localization, Content, On-Site, VisionPro, Work Experience
	2. Vision Hardware Selection and Integration	
	3. Technical Solution Design and Evaluation	

B. Quantitative Comparison

Table II summarizes the task domain-level structural metrics for the LDA result and the TF-IDF plus K-Means clustering under the same protocol, including the same partition count $K = 11$ and the same summary length $M = 30$. The LDA topics achieve a coverage of 0.9739 with an outlier rate of 0.0261, indicating that the topic summaries remain broadly representative of the documents assigned to each topic. The baseline method yields a near-saturated weighted coverage of 0.9936 and an outlier rate of 0.0064, reflecting strong within-cluster lexical alignment between centroid-derived keywords and clustered documents. Although the LDA method attains a lower coverage than the baseline, the two coverage scores remain relatively close in magnitude, suggesting that both approaches provide broadly representative task domain-level evidence summaries under the same protocol.

A different pattern is observed in inter-cluster redundancy. The average Jaccard overlap of the keyword summaries is 0.1343 for LDA and 0.1906 for the baseline method. Lower overlap indicates that topic summaries are more distinguishable at the keyword level, which is beneficial for competency induction because redundancy across topics increases ambiguity in domain naming and reduces the separability of the induced competency domains. The higher overlap of the baseline method suggests that multiple clusters share a larger portion of their top keywords, which is consistent with centroid-based summaries favoring broadly applicable recruitment terms that recur across job families. In contrast to the gap in coverage, the difference in overlap is noticeably larger, indicating a more pronounced separation in the distinctiveness of the resulting task domain-level representations.

Taken together, the task domain-level results suggest that the baseline's high representativeness is accompanied by increased redundancy across clusters. The higher coverage achieved by the baseline is largely attributable to broadly used recruitment terms that appear frequently within clusters but also recur across multiple clusters, which inflates inter-cluster overlap and weakens topic distinctiveness. In contrast, LDA yields more distinctive topic summaries with lower overlap, resulting in a more balanced trade-off between representativeness and separability. Such a balance is particularly relevant for constructing a task domain-competency hierarchy that supports consistent competency induction across different robotics role categories and avoids redundancy.

TABLE II. COMPARATIVE TASK DOMAIN-LEVEL STRUCTURAL METRICS

Structural Metrics	Proposed Method (LDA)	Baseline (TF-IDF + K-Means)
Coverage ($\tau = 2$)	0.9739	0.9936
Outlier rate	0.0261	0.0064
Overlap	0.1343	0.1906

Beyond the task domain-level comparison, the proposed taxonomy is further examined through competency-level metrics that operate on the internal decomposition within each task domain. Unlike the task domain-level metrics, which evaluate whether the keyword summaries provide a representative and discriminative evidence basis for competency induction, the competency-level metrics evaluate whether the resulting competencies remain grounded in the original topic evidence after translation. Because the TF-IDF plus K-Means baseline does not generate an explicit competency structure, this audit is applied only to the proposed LDA-based taxonomy.

Table III reports the competency-level audit results. The results indicate that the proposed translation procedure retains a substantial proportion of the dominant task domain evidence at the competency level. The average within-topic coverage reaches 0.9667, suggesting that the highest-ranked topic signals are largely preserved in the translated competencies rather than being lost during decomposition. At the same time, the within-topic outlier remains 0.0333 on average, indicating that most evidence is incorporated into the competency structure rather than remaining outside the induced competencies. In addition, the average within-topic overlap among competency keyword subsets is 0.0246, showing that the decomposition does not merely restate the same semantic content under multiple labels but instead maintains relatively clear internal boundaries among competencies.

TABLE III. COMPETENCY-LEVEL AUDIT METRICS

Task Domain Number	Within-topic Coverage	Within-topic Outlier	Within-topic Overlap
# 1	1.0000	0.0000	0.0000
# 2	0.9667	0.0333	0.0100
# 3	0.9333	0.0667	0.0000
# 4	0.9333	0.0667	0.0414
# 5	0.9333	0.0667	0.0268
# 6	0.9667	0.0333	0.0798
# 7	1.0000	0.0000	0.0781
# 8	0.9667	0.0333	0.0232
# 9	0.9667	0.0333	0.0000
# 10	1.0000	0.0000	0.0111
# 11	0.9667	0.0333	0.0000
Mean	**0.9667**	**0.0333**	**0.0246**

When interpreted jointly, the two levels of quantitative evidence support different but complementary claims. The task domain-level metrics show that the adopted LDA summaries provide a comparatively suitable evidence basis for competency induction in terms of representativeness and inter-cluster distinctiveness. The competency-level result further shows that, once translated, the competency-level structure preserves the dominant task domain evidence while limiting excessive within-topic repetition. Accordingly, the quantitative results do not establish an absolute notion of taxonomy correctness, but they do support the structural reliability, traceability, and interpretability of the proposed topic-to-competency translation under unlabeled conditions.

C. Qualitative Interpretation

Overall, the discovered topics cover major task domains frequently appearing in robotics recruitment texts, including industrial on-site operation and maintenance, mechanical design and engineering implementation, system integration and commissioning, embedded hardware and driver development, algorithm development and deployment, and performance validation and reliability testing. The keyword patterns provide a suitable evidence base for subsequent competency induction because they expose interpretable, task domain-level signals rather than only word frequency effects.

Based on the task domains, the associated competencies, and the Top-30 keyword evidence, the topics are organized into four groups. This organization provides a structured interpretation of the topic set and clarifies how topic evidence supports competency induction across development, deployment, and lifecycle operation.

The first group covers on-site operation and service and includes Domain #1 Industrial On-site Equipment Maintenance and Malfunction Diagnosis, and Domain #10 On-site Technical Support and After-sales Service. These topics align with roles such as field operation and maintenance engineers, equipment maintenance engineers, and after-sales technical support engineers. The skill pattern is strongly contextual and workflow-oriented. High-frequency terms emphasize fault diagnosis, troubleshooting, and the maintenance of deployed equipment such as automated guided vehicle (AGV) systems. The topics also contain repeated signals related to after-sales service and data recording. Together, these signals indicate a closed loop of detection, restoration, documentation, and handover. This group is practically significant because it directly affects system availability in real deployments and influences recovery efficiency and downtime costs. The keyword evidence consistently highlights field constraints and operational accountability, indicating persistent demand for personnel who can restore deployed systems under real-world conditions while maintaining traceability of actions and outcomes.

The second group corresponds to engineering design and integration and spans a major engineering chain from design to delivery. It includes Domain #2 Mechanical Structure Design and Engineering Implementation, Domain #3 Full Lifecycle Project Management and Operational Execution, Domain #4 Robotics Core Module Selection Design and Testing, Domain #5 Electrical Design and System Integration for Customized Automation, Domain #9 Industrial Robot Process Programming and Commissioning, and Domain #11 Machine Vision Inspection System Solution Implementation. This group covers multiple role categories, including mechanical and hardware engineers, electrical and automation engineers, integration and commissioning engineers, project managers, solution engineers, and vision engineers. The skill characteristics reveal three connected capability streams. The first stream is mechanical and hardware engineering and is represented by Domain #2 and #4. It follows an engineering workflow centered on design, analysis, selection, prototyping, integration, and validation. The keyword evidence emphasizes mechanical design and analysis, materials and manufacturing planning, engineering drafting, circuit and component selection, and prototype testing. These capabilities define manufacturability, maintainability, and performance boundaries and constitute the engineering foundation of robotics products. The second stream is electrical design and system integration, and is represented by Domain #5 and #9. It emphasizes controller programming, on-site wiring and commissioning, industrial communication configuration, and electromechanical joint debugging, as well as process programming and parameter configuration for industrial robots. These competencies support the delivery of deployable automation solutions under industrial constraints. The third stream is solution delivery and project execution, and is represented by Domain #3 and #11. It highlights requirements analysis, solution design and evaluation, cross-department collaboration, process coordination, delivery execution, and documentation. These capabilities connect technical development to application deployment by transforming technical resources into deliverable solutions. Domain #11 further indicates combined demand for image processing programming, vision hardware selection, and engineering integration, reflecting role-specific requirements for machine vision solution implementation in industrial robotics.

The third group covers software and algorithm development and includes Domain #6 Embedded Hardware and Driver Software Development for Robotics and Domain #7 Robotics Algorithm Development and Deployment. These

topics align with robotics software engineers, embedded engineers, and algorithm engineers. Domain #6 emphasizes engineering-oriented development across embedded platforms, driver programming, sensor and motion module integration, testing and fault localization, and documentation-oriented delivery. Domain #7 focuses on motion planning and control, deep learning training and optimization, environmental perception, and deployment within robotics software frameworks. The combined signal indicates that industry demand extends beyond algorithm development and includes engineering integration and deployment. This group governs both the upper bound of system intelligence and the feasibility of transferring algorithms from development environments to production deployments with maintainable interfaces and stable performance.

The fourth group concerns system verification and reliability assurance and is represented by Domain #8, Robotic System Performance Optimization and Reliability Testing. It aligns with roles such as system test engineers, validation engineers, and reliability engineers. The topic evidence concentrates on reliability testing and performance testing. It also contains documentation and coordination signals that reflect process discipline in engineering validation. This group supports sustained stability and scalable deployment by enforcing measurable validation procedures and reducing deployment risks. It also forms a quality assurance bridge between development, integration, and operation.

In summary, the four groups jointly form a competency taxonomy that covers key role categories across development, delivery, and lifecycle operation. The topic-to-competency translation enhances interpretability by converting keyword summaries into competencies under task domains. Each competency remains traceable to a task domain-specific keyword subset, enabling interpretability of competency statements and supporting qualitative interpretation. This structured view complements the quantitative results by explaining how topic distinctiveness and representativeness support a competency taxonomy for robotics recruitment mining.

V. Conclusion and Future Prospects

A. Contributions and Practical Implications

This study contributes to the literature on recruitment text mining and competency induction by proposing a data-driven approach for translating robotics recruitment texts into a structured competency taxonomy. The main outcome is a two-level competency taxonomy comprising 11 task domains and 48 competencies, grounded in topic keyword evidence and supported by a translation procedure and a traceable validation logic.

This study identifies latent competency domains and skill patterns embedded in robotics recruitment texts and organizes them into a coherent structure. The resulting taxonomy provides an empirical representation of employer competency demand across major robotics role families, including development, integration, deployment, and operation. Rather than remaining a list of isolated skills, the identified taxonomy is organized into a hierarchical structure that makes the relationships among task domains and competencies more explicit. Building on the robotics talent competency literacy framework analyzed in [13], the present study not only supports its main perspective but also further extends and enriches the skill requirements identified for the robotics domain, thereby producing a more detailed and fine-grained competency inventory. This richer and more structured competency taxonomy expands existing research on job-related competencies in the robotics field and provides a more actionable basis for occupational analysis and curriculum design.

Moreover, the study extracts latent topics from recruitment texts through LDA and then applies a translation procedure to convert unsupervised topic evidence into task domains and competencies under explicit requirements of keyword coverage and redundancy control. With explicit translation constraints and metrics, this approach further reduces the problem of naming subjectivity discussed in [25]. It also enriches the methodological pathway discussed in [21][22][23] for mining competencies through LDA and related models. The contribution of the study therefore lies not only in generating topic summaries, but also in providing an approach from topic keyword evidence to a competency taxonomy that can support both labor market analysis and educational planning.

The study also has important practical implications for labor market analysis, vocational education, and workforce development. First, the resulting taxonomy provides a traceable competency artifact that can support structured analysis of robotics labor demand, including role profiling, demand monitoring, and the identification of major competency clusters in industry recruitment. Because the taxonomy is derived from employer recruitment texts rather than predefined occupational templates, it reflects current market competency signals and can therefore complement more static competency frameworks developed through expert judgment.

Second, the study has reference value for talent cultivation in vocational education and for the development of teaching reform. Because the taxonomy is grounded in observed employer demand, it can provide an empirical reference for curriculum alignment, competency module refinement, training pathway design, and the identification of emerging skill combinations that may not yet be fully reflected in conventional educational programs. In robotics education, where the integration of mechanics, electronics, control, software, and application deployment is especially important, a competency structure grounded in observed demand can help connect educational planning more closely to industrial requirements.

Overall, the study contributes not only a robotics competency taxonomy but also a methodological perspective on how recruitment text evidence can be translated into an interpretable competency structure. Its value lies less in claiming superiority of a single topic model and more in demonstrating how unsupervised topic evidence can be organized, interpreted, and evaluated as a competency representation under realistic unlabeled conditions.

B. Limitations and Future Work

The scope of the present study is subject to the following limitations. First, the recruitment corpus is collected from a bounded set of online sources within the Chinese labor market, with job advertisements sampled from several representative major cities. This scope implies constraints on external generalizability. The resulting competency taxonomy should therefore not be directly transferred to overseas regions with different industrial structures, regulatory environments, or technological development stages. In addition, the corpus is gathered within a limited sampling period. In a rapidly evolving domain such as robotics, the observed skill demand signals primarily reflect the workforce requirements at the time of collection, and the competency profile is expected to shift as technologies mature and deployment patterns change.

Second, the evaluation emphasizes internal interpretability and structural quality using coherence and other metrics derived from topic and keyword evidence, including coverage, outlier rate, and overlap. This design matches the study setting in which labels for topic quality and competency taxonomy correctness are typically unavailable in recruitment text, and it enables comparison across methods under a unified evidence interface. The current results therefore provide internal structural support for the induced competency taxonomy, particularly in terms of representativeness, distinctiveness, and traceability. However, the present validation does not yet extend to external substantive assessment by domain experts. Additional external evidence, such as expert review, could further strengthen the interpretive claims and broaden the validation perspective.

Future research can extend the present study along two directions. First, external validity and longitudinal analysis can be strengthened through a structured cross-cultural validation framework. This includes expanding data collection to multiple countries or regions and conducting comparisons across markets with differing industrial contexts. Longitudinal designs should be prioritized to track how competency demand evolves, enabling the separation of stable competency cores from time-sensitive requirements and emerging skill trends. Second, external evidence and application validation can be incorporated as a complementary verification layer beyond internal interpretability metrics. These extensions may incorporate expert review supported by consensus analysis, which can further reduce the influence of subjective naming choices beyond the translation constraints imposed by the translation procedure, and validation through downstream tasks that require structured competency representations.

ACKNOWLEDGMENT

This research was supported by Jiangsu Province Higher Education Teaching Reform Research Project (Grant No. 2025JGYB674), Teaching Reform Project of Suzhou Vocational University (Grant No. SZDJG-24006), and Special Project of Suzhou Polytechnic University (Grant No. GJSZX-260110).

REFERENCES

[1] Alhloul A, Kiss E. Industry 4.0 as a Challenge for the Skills and Competencies of the Labor Force: A Bibliometric Review and a Survey[J]. Sci, 2022, 4(3): 34.

[2] Kipper L M, Iepsen S, Dal Forno A J, et al. Scientific mapping to identify competencies required by industry 4.0[J]. Technology in Society, 2021, 64: 101454.

[3] Sheriff N, Sevukan R. Discovering research data management trends from job advertisements using a text-mining approach[J]. Journal of Information Science, 2023: 01655515231193845.

[4] Wowczko I A. Skills and vacancy analysis with data mining techniques[C]//Informatics. MDPI, 2015, 2(4): 31-49.

[5] Hernandez-de-Menendez M, Morales-Menendez R, Escobar C A, et al. Competencies for industry 4.0[J]. International Journal on Interactive Design and Manufacturing (IJIDeM), 2020, 14(4): 1511-1524.

[6] Maisiri W, Darwish H, Van Dyk L. An investigation of industry 4.0 skills requirements[J]. South African Journal of Industrial Engineering, 2019, 30(3): 90-105.

[7] Nasir S A M, Wan Yaacob W F, Wan Aziz W A H. Analysing online vacancy and skills demand using text mining[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1496(1): 012011.

[8] Tzimas G, Zotos N, Mourelatos E, et al. From data to insight: Transforming online job postings into labor-market intelligence[J]. Information, 2024, 15(8): 496.

[9] MacKay M, Ford C, Grant L E, et al. Developing competencies in public health: a scoping review of the literature on developing competency frameworks and student and workforce development[J]. Frontiers in public health, 2024, 12: 1332412.

[10] Kiesler N. Towards a competence model for the novice programmer using bloom's revised taxonomy-an empirical approach[C]//Proceedings of the 2020 acm conference on innovation and technology in computer science education. 2020: 459-465.

[11] Irvine J. Taxonomies in education: Overview, comparison, and future directions[J]. Journal of Education and Development, 2021, 5(2): 1.

[12] Wang L, Ning Y, Li Z. Exploration and Practice of the Blended Teaching Mode of Industrial Robot Courses Based on the Background of Artificial Intelligence[C]//Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area Education Digitalization and Computer Science International Conference. 2025: 802-806.

[13] Yu-Shen F, Shi-Jie W, Run-Jiao H, et al. Analysis of Core Competency Literacy of Innovative Technology Talents in Intelligent Robot Industry[C]//2020 5th International Conference on Humanities Science and Society Development (ICHSSD 2020). Atlantis Press, 2020: 106-110.

[14] Olukanni E, Akanmu A, Jebelli H. Industry perception of competencies for human—robot collaboration in the construction industry: A Delphi study[J]. Frontiers of Engineering Management, 2025, 12(4): 854-879.

[15] Do H D, Tsai K T, Wen J M, et al. Hard skill gap between university education and the robotic industry[J]. Journal of Computer Information Systems, 2023, 63(1): 24-36.

[16] Bademosi F M, Issa R R A. Essential knowledge, skills, and abilities required for talent cultivation in construction automation and robotics[M]//Automation and robotics in the architecture, engineering, and construction industry. Cham: Springer International Publishing, 2022: 31-57.

[17] Salotti J M, Suhir E. Collaborative Robotics: Application of Delphi Method[J]. Journal of Field Robotics, 2025, 42(5): 1799-1807.

[18] Gavrilescu M, Leon F, Minea A A. Techniques for transversal skill classification and relevant keyword extraction from job advertisements[J]. Information, 2025, 16(3): 167.

[19] Chen Y, Pan R. Research on data analysis and visualization of recruitment positions based on text mining[J]. Advances in Multimedia, 2022, 2022(1): 9047202.

[20] Colombo E, Mercorio F, Mezzanzanica M. Applying machine learning tools on web vacancies for labour market and skill analysis[J]. Terminator or the Jetsons, 2018.

[21] Gurcan F, Soylu A, Khan A Q. Towards a sustainable workforce in big data analytics: Skill requirements analysis from online job postings using neural topic modeling[J]. Sustainability, 2025, 17(20): 9293.

[22] Kang P S, Enstroem R, Bhawna B, et al. A text mining study of competencies in modern supply chain management with skillset mapping[J]. Supply Chain Analytics, 2025, 10: 100117.

[23] Gurcan F, Cagiltay N E. Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling[J]. IEEE access, 2019, 7: 82541-82552.

[24] Djunaidi K, Kusuma D T, Ningrum R F, et al. Big Data Analytics of Knowledge and Skill Sets for Web Development Using Latent Dirichlet Allocation and Clustering Analysis[J]. International Journal of Advanced Computer Science & Applications, 2025, 16(1).

[25] Zhou Y, Tao L, Xue Z, et al. From Job Postings to Vocational Education Standards: Mapping Competency Requirements for NEV Sales and Livestreaming Hosts[J]. World Electric Vehicle Journal, 2026, 17(3): 162.

[26] Madsen A, Reddy S, Chandar S. Post-hoc interpretability for neural nlp: A survey[J]. ACM Computing Surveys, 2022, 55(8): 1-42.

[27] Soare E. Perspectives on designing the competence based curriculum[J]. Procedia-Social and Behavioral Sciences, 2015, 180: 972-977.

[28] McClelland D C. Identifying competencies with behavioral-event interviews[J]. Psychological science, 1998, 9(5): 331-339.

[29] Hoyle A, Goel P, Hian-Cheong A, et al. Is automated topic model evaluation broken? the incoherence of coherence[J]. Advances in neural information processing systems, 2021, 34: 2018-2033.

[30] Kim H, Park H, Song M. Developing a topic-driven method for interdisciplinarity analysis[J]. Journal of informetrics, 2022, 16(2): 101255.

[31] Fang F, Zhou Y. A study on recruitment of data analyst based on text mining and visualization technology[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1952(4): 042017.

[32] Zeng X, Chu S, Chen X. China's labor market demand in the shadow of COVID-19: Evidence from an online job board[J]. Journal of Asian Economics, 2025, 97: 101885.

[33] Gojare S, Joshi R, Gaigaware D. Analysis and design of selenium webdriver automation testing framework[J]. Procedia Computer Science, 2015, 50: 341-346.

[34] Thooriqoh H A, Annisa T N, Yuhana U L. Selenium framework for web automation testing: A systematic literature review[J]. JUTI: Jurnal Ilmiah Teknologi Informasi, 2021: 65-76.

[35] Cao S. New word detection algorithm combining correlation confidence and jieba word segmentation[J]. Comput Syst Appl, 2020, 29: 144-51.

[36] Na D, Xu C. Automatically generation and evaluation of stop words list for Chinese patents[J]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2015, 13(4): 1414-1421.

[37] Jelodar H, Wang Y, Yuan C, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey[J]. Multimedia tools and applications, 2019, 78(11): 15169-15211.

[38] Bag S, Kumar S K, Tiwari M K. An efficient recommendation generation using relevant Jaccard similarity[J]. Information Sciences, 2019, 483: 53-64.

[39] Heinrich G. Parameter estimation for text analysis[R]. Darmstadt, Germany: Technical report, 2005.

APPENDIX. TOPIC SUMMARIES, METRICS, AND TRANSLATION DETAILS

Task Domains	Competency	Top-30 Keywords	Within-topic Coverage	Within-topic Outlier	Within-topic Overlap	Assigned Keywords
#1 Industrial On-site Equipment Maintenance and Malfunction Diagnosis	1. Troubleshooting and Repair of Automated Equipment Malfunction	Equipment, Fault, Automation, Team, Issue, Mechanical, Technology, System, Collaboration, Fundamentals, Electrical, Production, AGV, Logging, Content, Data, Troubleshooting, Task, Problem Solving, Awareness, Professional Spirit, Skill, Optimization, Learning Ability, Testing, Teamwork, On-Site, Industrial, Sensor, Principle	1.0000	0.0000	0.0000	Equipment, Fault, Automation, Electrical, AGV, Troubleshooting
	2. On-site Process Data Logging and Analysis					Production, Logging, Content, Data, On-Site
	3. Teamwork and Problem Solving					Team, Collaboration, Problem Solving, Awareness, Professional Spirit, Learning Ability, Teamwork, Task
	4. Mechatronic System Principles Understanding and Technical Optimization					Issue, Mechanical, Technology, System, Skill, Optimization, Principle
	5. Industrial Sensor Testing and Equipment Maintenance					Testing, Industrial, Sensor,Maintenance
#2 Mechanical Structure Design and Engineering Implementation	1. Mechanical Structure Design and Analysis	Design, Mechanical, Software, Process, Structure, Product, Mechanical Design, Analysis, Structural Design, Material, Selection, Automation, Machining, Technology, Equipment, CAD, Drive System, Issue, Production, Mechanism, Drawings, Principle, Work Experience, Project, Solution, Mechatronic, Optimization, Motor, Collaboration, Engineering	0.9667	0.0333	0.0100	Design, Mechanical, Structure, Mechanical Design, Analysis, Structural Design, Principle
	2. Automation and Electromechanical Drive Systems Optimization					Automation, Equipment, Drive System, Mechanism, Mechatronic, Optimization, Motor
	3. Material Selection and Manufacturing Process Planning					Process, Material, Selection, Machining, Production
	4. CAD Drafting and Design Documentation Output					Software,CAD, Drawings, Solution, Engineering
	5. Project Collaboration and Resolution of Product Technical Issues					Product, Technology, Issue, Principle, Project, Collaboration
#3 Full Lifecycle	1. Project Planning and Documentation Management	Product, Project, Technology, Team,	0.9333	0.0667	0.0000	Project, Management, Plan, Project

Task Domains	Competency	Top-30 Keywords	Within-topic Coverage	Within-topic Outlier	Within-topic Overlap	Assigned Keywords
Project Management and Operational Execution		Management, Industry, Customer, Quality, Analysis, Solution, Company, Process Flow, Issue, User, Market, Plan, Deployment, Project Management, Intelligent, Collaboration, Department, Documentation, Resources, Planning, Objective, Business, Design, Process, Solution, Automation				Management, Documentation, Planning, Process
	2. Customer Requirements Analysis and Solution Design					Customer, Analysis, Solution, Objective, Design, Solution
	3. Cross-Departmental Collaboration and Process Improvement Execution					Team, Company, Process Flow, Collaboration, Department, Business
	4. Product Implementation Delivery and Technical Issue Resolution					Product, Technology, Quality, Issue, Deployment
	5. Industry Resource Integration and Requirement Fulfillment					Industry, User, Market, Resources
#4 Robotics Core Module Selection Design and Testing	1. Circuit Design and Electronic Component Selection	Hardware, Design, Motor, Circuit, Electronics, Structure, Joint, Product, Testing, Module, Wiring Harness, Production, Team, Mechanical, Professional Spirit, Issue, Laboratory, Hands-On, Component, Attitude, Quality, Mindset, Power Supply, System, Circuit Design, Collaboration, Selection, Fundamentals, Management, Tooling	0.9333	0.0667	0.0414	Design, Circuit, Electronics, Product, Component, System, Circuit Design, Selection
	2. Motor and Joint Module Integration with Wiring Harnesses					Motor, Structure, Joint, Module, Wiring Harness, Mechanical, Power Supply, System
	3. Hardware Prototyping, Testing, and Verification					Hardware, Product, Testing, Mechanical
	4. Practical Engineering Skills and Team Collaboration					Team, Professional Spirit, Issue, Laboratory, Hands-On, Attitude, Mindset, Collaboration, Fundamentals, Management, Tooling
#5 Electrical Design and System Integration for customized Automation	1. PLC Programming and System Software Debugging	Electrical, Automation, PLC, Project, Equipment, Programming, Design, Program, Customized, Software, Selection, Industrial, On-Site, Work Experience, Technology, System, Mechatronic, Electrical Design, Drawings, Vision, Robot Manipulator, Schematic Diagram, Communication, Principle, Wiring, Industry, Mitsubishi, Drafting, CAD, Signal	0.9333	0.0667	0.0268	PLC, Programming, Program, System, Mitsubishi
	2. CAD Drafting and Electrical Schematic Interpretation					Electrical, Design, Software, Electrical Design, Drawings, Schematic Diagram, Principle, Wiring, Drafting, CAD
	3. On-site Wiring, Commissioning, and System Integration					Equipment, Industrial, On-Site, System
	4. Industrial Communication Configuration and Electromechanical Joint Debugging					Mechatronic, Industrial, Communication, Signal
	5. Vision and Robot Manipulator Application in Custom Automation Projects					Automation, Project, Customized, Selection, Technology, Vision, Robot Manipulator
#6 Embedded Hardware and Driver Software Development for Robotics	1. Embedded Linux Driver Programming and Development	Testing, System, Software, Driver, Automation, Design, Documentation, Programming, Sensor, Hardware, Embedded, Linux, Analysis, Technology, Issue, Computer, Project, Platform, Product, Motion, Functionality, Module, Principle, ROS, Python, Code, Electronics, Work Experience, Tooling, Motor	0.9667	0.0333	0.0798	Software, Driver, Programming, Hardware, Embedded, Linux, Technology, Computer, Principle, Code, Tooling
	2. Sensor and Motion Control Module Development					Software, Sensor, Technology, Motion, Functionality, Module, Principle, Code, Electronics, Tooling, Motor
	3. Test Analysis and Fault Localization					Testing, Automation, Analysis, Issue
	4. ROS System Design and Application Development					System, Software, Design, Technology, Platform, Principle, ROS, Python, Code, Tooling
	5. Project Documentation and Product Deliverable Handover					Documentation, Project, Product

Task Domains	Competency	Top-30 Keywords	Within-topic Coverage	Within-topic Outlier	Within-topic Overlap	Assigned Keywords
#7 Robotics Algorithm Development and Deployment	1. Trajectory Planning and Control Algorithm Development	Algorithm, Planning, Model, Motion, Optimization, Data, Project, Technology, Python, Robotic Arm, Deep Learning, Framework, Robot, ROS, Kinematics, Deployment, Path, Domain, Real-World, Vision, Fundamentals, Control Algorithm, Slam, Environment, Scenario, AI, System, Mathematics, Training, Dynamics	1.0000	0.0000	0.0781	Algorithm, Planning, Motion, Python, Robotic Arm, Kinematics, Path, Control Algorithm, Dynamics
	2. Deep Learning Model Training and Optimization					Model, Optimization, Data, Deep Learning, Framework, Robot, AI, Mathematics, Training, Domain
	3. SLAM and Environmental Perception Algorithm Implementation					Algorithm, Project, Python, Real-World, Vision, SLAM, Environment, Scenario
	4. Algorithm Integration and Deployment in ROS Framework					Algorithm, Technology, Python, Framework, ROS, Deployment, Fundamentals, System
#8 Robotic System Performance Optimization and Reliability Testing	1. System Performance and Parameter Testing and Optimization	Optimization, Testing, Design, System, Technology, Performance, Issue, Product, Analysis, Solution, Data, Functionality, Documentation, Hardware, Stability, Reliability, Motion, Evaluation, Coordination, Module, Efficiency, Scenario, Team, Sensor, Planning, Process, Parameter, Review, Environment, Core	0.9667	0.0333	0.0232	Optimization, Testing, System, Performance, Data, Hardware, Parameter
	2. Reliability-oriented Design and Stability Validation					Design, Stability, Reliability, Technology, Issue
	3. Scenario-based Testing for Motion Modules and Sensors					Testing, Motion, Module, Scenario, Sensor, Environment
	4. Product Function Analysis and Test Documentation Preparation					Testing, Product, Analysis, Solution, Functionality, Documentation, Evaluation, Review
	5. Team Coordination and Efficiency-oriented Planning					Coordination, Efficiency, Team, Planning, Process
#9 Industrial Robot Process Programming and Commissioning	1. Mainstream Industrial Robot Controllers Programming	Programming, Process, ABB, Industrial, Project, KUKA, Automation, FANUC, Work Experience, On-Site, System, Program, Automotive, YASKAWA, Brand, Offline, Sense Of Responsibility, Arc Welding, Mechatronic, Parameter, Configuration, Optimization, Spot Welding, Teamwork, Laser, Body-In-White, Responsibility, Mechanical, Trajectory, Dispensing	0.9667	0.0333	0.0000	Programming, ABB, Industrial, KUKA, FANUC, Program, YASKAWA, Brand, Mechanical, Trajectory
	2. Specific Process Application and Parameter Configuration					Process, Project, Automation, On-Site, System, Automotive, Offline, Arc Welding, Mechatronic, Parameter, Configuration, Optimization, Spot Welding, Laser, Body-In-White, Dispensing
	3. Teamwork and Responsibility Awareness					Sense Of Responsibility, Teamwork, Responsibility
#10 On-site Technical Support and After-sales Service	1. On-site Emergency Fault Diagnosis and Handling	On-Site, Customer, Project, Equipment, Technical Support, Issue, Product, Process, Production, Electrical, Company, After-Sales, Mechanical, Personnel, Solution, Technology, Process, Materials, Fault, Automation, Service, Organization, Guidance, Management, Optimization, Aspects, Tooling, File, Client, Department	1.0000	0.0000	0.0111	On-Site, Equipment, Technical Support, Issue, Process, Electrical, Mechanical, Fault, Automation
	2. Customer Communication and Technical Service Provision					Customer, Product, Technology, Service, Guidance, Tooling, Client, Optimization
	3. Materials Organization and After-sales Documentation Management					Production, After-Sales, Materials, Organization, Management, File
	4. Project Service Delivery and Resource Coordination					Project, Company, Personnel, Solution, Process, Resource, Department

Task Domains	Competency	Top-30 Keywords	Within-topic Coverage	Within-topic Outlier	Within-topic Overlap	Assigned Keywords
#11 Machine Vision Inspection System Solution Implementation	1. Image Processing Algorithm Programming	Vision, System, Machine Vision, Software, Project, Image Processing, Inspection, Hardware, Automation, Industrial, Design, Technology, Team, Algorithm, Selection, Customer, Light Source, Computer, Lens, Equipment, Analysis, Programming, Solution, OpenCV, Evaluation, Localization, Content, On-Site, VisionPro, Work Experience	0.9667	0.0333	0.0000	Machine Vision, Software, Image Processing, Algorithm, Computer, Analysis, Programming, OpenCV, Localization, VisionPro
	2. Vision Hardware Selection and Integration					Vision, System, Inspection, Hardware, Industrial, Selection, Light Source, Lens, Equipment
	3. Technical Solution Design and Evaluation					Project, Automation, Design, Technology, Team, Customer, Solution, Evaluation, Content, On-Site