

Demand-Driven Competency Extraction from Online Job Postings Using Precision-Constrained Context-Aware Gated Fusion

Shi Fengyuan¹, Megat F. ZUHAIIRI^{2*}, Huang Yonghua³, Aabid A. MIR⁴, Ali TUFAIL⁵, MH Zafar⁶

School of Artificial Intelligence and Electronic Information, Nantong Vocational University, Nantong, China^{1,3}

Universiti Kuala Lumpur, Malaysian Institute of Information Technology, Kuala Lumpur, Malaysia^{1,2,4}

School of Digital Science, Universiti Brunei Darussalam, Gadong, Brunei⁵

Cybersecurity and Information Networks Centre (CINC), Cardiff School of Technologies,

Cardiff Metropolitan University, Cardiff, UK⁶

Abstract—Online job postings are rich sources of competency data, yet extraction is challenging because explicit terms, semantic paraphrases, and contextual cues overlap. To address this, we introduce a demand-driven extraction framework centered on Precision-Constrained Context-Aware Gated Fusion (PC-CAGF) to balance recall and precision. We collected 14,030 Chinese IT job postings from Boss Zhipin and Zhaopin. After removing invalid, duplicate, near-duplicate, and low-information entries, 12,513 postings remained. Recruitment-demand themes were identified using TF-IDF and non-negative matrix factorization (NMF). NMF was compared with Latent Dirichlet Allocation (LDA) and BERTopic, and topic stability was examined through repeated random-seed runs and 80% subsampling. We selected a 22-topic solution and consolidated it through expert review into five primary dimensions and 25 secondary competency indicators. A manually annotated reference set of 500 postings yielded 12,500 posting-indicator decisions, with 90.44% inter-coder agreement and Cohen's kappa of 0.7728. The rule-only baseline achieved precision of 0.9402, recall of 0.6577, and F1 score of 0.7740. PC-CAGF increased recall to 0.7378 and F1 to 0.8215 while maintaining a precision of 0.9267. Compared with rule-only extraction, F1 score improved by 0.0475 (95% bootstrap confidence interval: 0.0400–0.0549), with significant paired-test results ($p < 0.01$). Ablation analysis showed that contextual features contributed most of the performance gain, while specialist feature sets further improved context-dependent indicators. Applied to the full corpus, the frozen model increased estimated positive indicator assignments from 61,532 to 69,858. The results suggest that precision-constrained semantic augmentation improves competency extraction while largely preserving lexical precision, supporting interpretable recruitment-text analysis and demand-side competency profiling.

Keywords—Competency extraction; online job postings; natural language processing; semantic similarity; context-aware gated fusion; topic modelling; multilabel classification; PC-CAGF

I. INTRODUCTION

Online job ads offer a timely, large-scale lens on evolving labor-market demands. These texts include technical skills, qualifications, experience, and broader professional competencies. But they rarely come in clean, consistent forms.

*Corresponding author.

This work was supported by the Sixth Jiangsu Provincial Vocational Education Teaching Reform Research Project under Grant No. ZDZC110 and the 2024 Special Research Project for Higher Education Institutions in Jiangsu Province under Grant No. 2024JCSZ20.

Instead, the language is heterogeneous, noisy, and highly context-dependent. This messiness poses a concrete obstacle for employability analytics. To support downstream tasks like labor-market tracking or graduate assessment, competency requirements must be accurately extracted from raw recruitment text first. The difficulty is reinforced by two strands of prior work. Existing conceptual frameworks provide divergent competency definitions [1]. Empirical machine learning studies repeatedly identify skill gaps between computing graduates and employer expectations [2], [3]. Together, these findings push us toward an employer-grounded representation, one derived not from theoretical constructs but from direct recruitment evidence.

Standardized taxonomies such as ESCO offer consistent descriptions of skills, competences, occupations, and qualifications [4]. Such taxonomies improve comparability, but they may not immediately reflect emerging technologies, tools, and rapidly changing work practices, particularly in information technology occupations [5]. Online job postings can complement standardized taxonomies because they provide more current evidence of employer demand and have been used to investigate AI skills, digital occupations, and workforce transformation [6]. However, transforming large volumes of recruitment text into structured competency information remains challenging because postings contain both applicant requirements and contextual material such as company descriptions, benefits, project background, and generic recruitment language.

Research on job-posting data has largely developed along three lines. Some studies use topic models to uncover recurring demand patterns and groups of related skills [7], [13]. Others focus on extracting technical and non-technical requirements directly from vacancy descriptions [8], [9], while taxonomy-based approaches link the extracted terms to standardized occupational or skill vocabularies [10], [11]. What is missing from current automated approaches is a higher-level organizational framework. Most studies stop at producing broad thematic clusters or standalone skill terms. They do not take the additional step of structuring these outputs into a hierarchical indicator system. Nor do they systematically assess whether the extracted indicators are conceptually coherent or empirically accurate. These two dimensions of evaluation, structural and

operational, are rarely addressed together in the literature. This gap is precisely what we aim to fill.

Balancing precision and recall poses another challenge. Rule-based approaches are accurate and interpretable for explicit requirements. Yet their coverage is limited. They often overlook paraphrased or contextually conditioned expressions. Semantic encoders address this limitation by capturing linguistic variation. Even so, they are not immune to errors. Similarity-based matching may inadvertently treat employer-related information as applicant qualifications. For example, a statement that a company serves financial clients makes this point clear. Such a statement does not equate to a requirement for financial business knowledge. A robust employability index, therefore, depends on two conditions. One is a well-specified competency construct. The other is an extraction procedure that respects contextual boundaries.

To fill the gaps identified above, we propose PC-CAGF, a model designed for demand-driven competency extraction from online job postings. We begin by applying TF-IDF-based non-negative matrix factorization to our cleaned corpus of 12,513 Chinese IT job ads. This initial step yields a preliminary topic structure. We then compare this structure against two widely used alternatives, Latent Dirichlet Allocation and BERTopic. To ensure that our chosen topic solution is not an artifact of random variation, we run multiple tests. These include multi-criterion topic selection, repeated fitting with different random seeds, and repeated subsampling at 80 percent of the data. Only after passing these stability checks do we proceed to expert consolidation. This final step reduces the extracted topics into five primary dimensions and twenty-five secondary competency indicators. PC-CAGF then integrates rule-based evidence, semantic similarity, contextual features, and indicator identity through a pooled logistic gate. Unlike direct rule-semantic combination, the framework preserves rule-positive predictions and introduces additional semantic positives only when indicator-specific thresholds, contextual evidence, precision constraints, and No-Harm conditions are satisfied. Two sentence-level specialist feature sets are further incorporated for indicators that are particularly difficult to identify from general semantic evidence.

The main contributions of this study are threefold. First, it connects topic-based competency discovery with an operational posting-level extraction task instead of treating latent topics as final competency labels. Second, it introduces a precision-constrained context-aware fusion strategy that selectively augments high-precision lexical evidence rather than replacing it with unrestricted semantic predictions. Third, the model is evaluated using a manually annotated reference set under nested out-of-fold validation, baseline comparison, ablation analysis, bootstrap uncertainty estimation, and paired statistical testing. The resulting approach provides an interpretable and auditable method for extracting structured competency demand from large-scale recruitment text.

II. RELATED WORK

A. Employability and Labor-Market Analytics

Employability research has traditionally examined the competencies and characteristics associated with graduates'

transition into employment. Existing frameworks include academic knowledge, technical capabilities, interpersonal skills, behavioral attributes, and career-related competencies, although their definitions and measurement dimensions vary across educational and occupational settings [1]. Machine learning approaches have extended this work by analyzing graduate characteristics and skill profiles to identify employability-related factors and mismatches between education and labor-market expectations [2], [3]. These studies demonstrate the value of data-driven employability analysis, but most operate primarily from the supply side by describing or predicting the capabilities of students and graduates.

Labor market analytics provides a complementary demand-side perspective. Standardized resources such as the European Skills, Competences, Qualifications and Occupations taxonomy support consistent descriptions of skills and occupations [4]. However, rapid technological change can create discrepancies between standardized skill classifications and terminology currently used by employers [5]. Online job postings have consequently become an important source for examining changing AI-related and emerging technical skill demand [6]. Their scale and temporal sensitivity make them useful for detecting changes in employer requirements that may not yet be represented in formal competency frameworks. The proposed hierarchy of five primary dimensions and 25 secondary indicators is not intended to replace ESCO or other established taxonomies; rather, it provides a task-specific representation of competency requirements derived from Chinese IT recruitment texts and may complement standardized classifications.

The growing sophistication of NLP has made recruitment data more accessible to automated systems. Ifakir et al. [12] exemplified this trend by combining named entity recognition, transformer-based semantic matching, and large language models into a single processing pipeline. Their approach advances the state of the art in converting unstructured ad text into structured fields. But for employability analytics, structure alone is insufficient. We do not just need structured data. We need a competency representation that speaks to educators and employers alike. It must be interpretable without specialized NLP training. It must support tracking across time. And it must allow for direct comparison with educational or individual performance metrics. These additional requirements are not addressed by the current generation of recruitment automation systems.

B. Topic Modeling of Online Job Postings

Topic modeling provides an unsupervised approach for discovering recurring structures in large collections of job advertisements. Previous studies have applied topic models to identify clusters of technical skills, professional competencies, and emerging labor-market themes from recruitment text [7], [13]. Such methods are particularly useful when a predefined competency inventory is unavailable because latent patterns can be discovered directly from the corpus.

The effectiveness of topic modeling depends strongly on the representation method, model family, number of topics, and evaluation criteria. Ao et al. [13] compared dictionary-based methods with several unsupervised topic-modeling approaches for identifying skill requirements in large-scale online job

advertisements. Their results illustrate that alternative extraction and topic-modeling techniques may produce different representations of labor-market skills and that the usefulness of a topic solution should be considered in relation to the downstream analytical objective. Other job-posting studies have similarly demonstrated that neural topic representations can capture emerging technical and professional demand patterns that may be difficult to specify in advance [7].

A limitation of many topic-based approaches is that latent topics are treated as the principal analytical output. A topic may contain several related competencies, overlap with other topics, or include organizational and recruitment-template language. Consequently, a latent topic is not necessarily equivalent to an operational competency indicator. Moreover, a topic solution that appears interpretable in one fitted model may change across different initializations or samples. These limitations motivate the use of multiple evaluation criteria and stability analysis when topic models are used as evidence for competency construction. In the present study, topic modeling therefore serves as a discovery stage rather than as the final competency representation. The identified themes are subjected to stability assessment, expert interpretation, and semantic consolidation before operational indicators are defined.

C. Skill and Competency Extraction

A second research stream focuses on extracting individual skills or competencies directly from recruitment text. Early lexical and dictionary-based approaches offer high interpretability and can reliably identify explicitly stated technical requirements, but their coverage is limited when employers use paraphrases, indirect expressions, or domain-specific terminology. The development of contextual language models and sentence embeddings has therefore shifted skill extraction toward semantic representations.

Zhang et al. introduced SkillSpan, an expert-annotated benchmark dataset for hard and soft skill extraction from English job postings [14]. The study demonstrated the importance of domain adaptation and contextual language representation for identifying skill expressions in recruitment text. Weakly supervised approaches have also been proposed to reduce dependence on large manually annotated datasets. Zhang et al. [11], for example, used weak supervision and latent representations to identify skills from job advertisements, showing that external skill resources and semantic representations can extend extraction beyond direct lexical matching. Taxonomy-oriented approaches further map extracted text to standardized skill structures. ESCOX [10], for instance, applies semantic representations and large language models to extract and align skills and occupations from unstructured text.

Recent studies have paid increasing attention to nontechnical and transversal competencies. Leon et al. [9] used sentence embeddings for hierarchical classification of transversal skills in job advertisements. Akkasi [8] similarly applied explainable transformer-based ensemble models to extract technical and nontechnical skills from job descriptions. Collectively, these studies show that semantic representations improve the ability to recognize implicit and variable competency expressions.

Nevertheless, semantic coverage introduces another problem. Recruitment documents contain not only applicant requirements but also company descriptions, project backgrounds, benefits, customer information, and generic organizational statements. A high semantic similarity between a sentence and a competency definition therefore does not necessarily imply that the competency is required from an applicant. Existing recruitment NLP research has demonstrated the usefulness of contextual representations, semantic matching, weak supervision, and taxonomy-informed extraction [8]–[12], [14], but controlling false-positive semantic augmentation remains important when the purpose is to produce reliable posting-level competency indicators.

D. Research Gap

The literature reveals three related but largely separate research directions. Employability analytics identifies competencies relevant to graduate employment but often relies on supply-side student or graduate data [1]–[3]. Topic-modeling studies use job postings to discover labor-market themes but generally stop at latent topics or skill clusters [7], [13]. Skill-extraction research can identify individual technical or transversal skills with increasingly sophisticated semantic models [8]–[12], [14], but commonly begins with an existing skill inventory, taxonomy, or annotation schema. The present study brings these streams together for a demand-derived, operational competency hierarchy and evaluates extraction with shared in-domain baselines. Its contribution is the controlled decision policy and integrated workflow, not a claim that its encoder or topic model is universally superior to all existing systems.

Two gaps therefore remain. First, relatively little work connects unsupervised demand discovery with the construction of a concise, operational competency hierarchy and then evaluates whether the resulting indicators can be automatically extracted from the same type of recruitment evidence. Treating topic discovery, competency construction, and extraction as separate tasks makes it difficult to establish a consistent path from changing labor-market language to an interpretable demand-side competency representation. Second, although semantic models improve coverage, a practical extraction system must balance this gain against the false positives introduced by contextual or organizational language. Simple lexical rules provide precision, whereas unrestricted semantic augmentation may reduce reliability.

The present study addresses these gaps by combining demand-driven competency discovery with Precision-Constrained Context-Aware Gated Fusion (PC-CAGF). Topic modeling is first used to identify latent recruitment-demand structures, which are evaluated for quality and stability before expert consolidation produces five primary dimensions and 25 operational indicators. PC-CAGF subsequently performs posting-level multilabel extraction using rule-based, semantic, and contextual evidence. Rather than replacing lexical evidence, its add-only design preserves rule-positive predictions and selectively introduces additional semantic positives through an indicator-conditioned pooled gate, indicator-specific thresholds, precision constraints, and No-Harm activation.

The framework therefore links competency discovery, structured representation, and controlled extraction within a unified and auditable recruitment-text analysis pipeline.

III. MATERIALS AND METHODS

A. Research Design

This study adopts a sequential mixed-method design that integrates unsupervised topic modeling, expert-guided competency construction, and supervised hybrid extraction. As

shown in Fig. 1, the workflow comprises corpus construction, topic discovery, competency-index construction, extraction-model evaluation, and frozen full-corpus application. The 25-indicator hierarchy was created through expert interpretation and consolidation rather than by equating one topic with one indicator. A manually annotated reference standard was then used for nested out-of-fold evaluation. Only after evaluation were the fitted gate, indicator thresholds, and No-Harm switches frozen and applied to the complete corpus, ensuring that full-corpus demand profiles did not influence model selection.

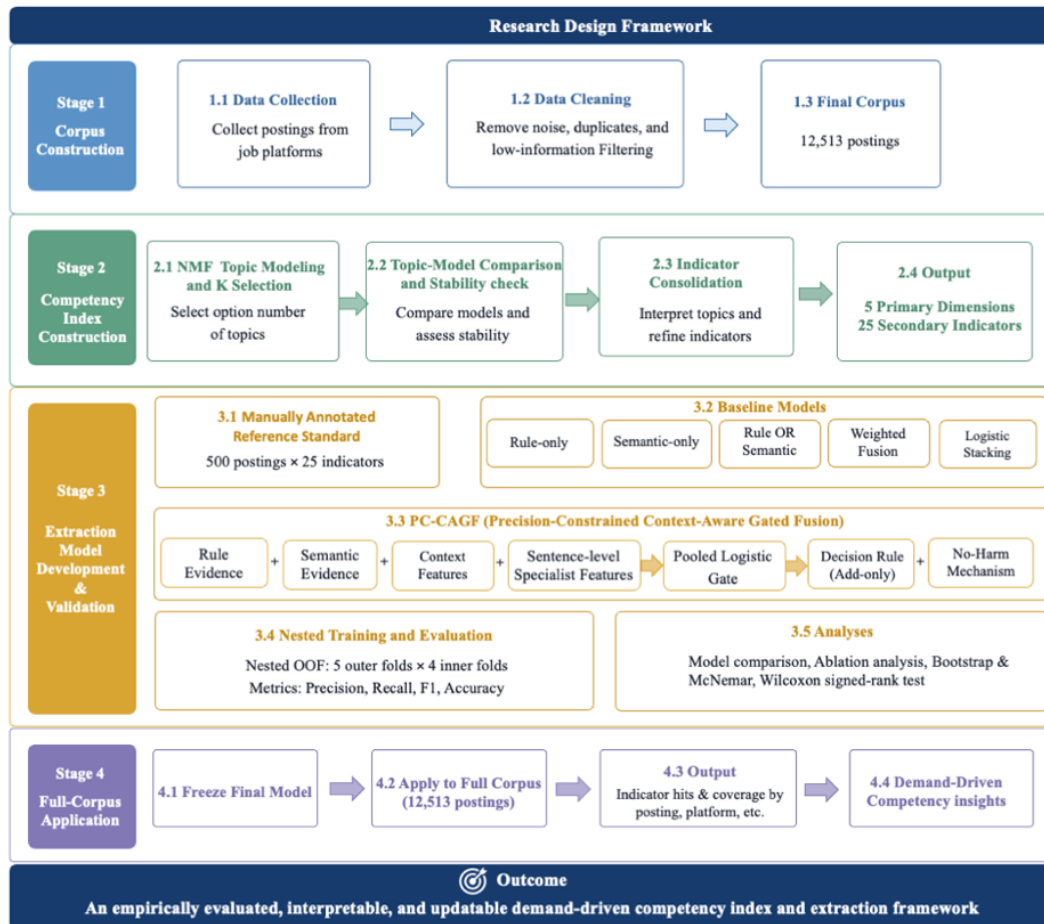


Fig. 1. Research design framework.

B. Data Source and Preprocessing

The source corpus consisted of 14,030 postings collected from Boss Zhipin (11,011) and Zhaopin (3,019), two prominent Chinese recruitment platforms, during the period from October 17 to October 21, 2025. The search encompassed areas such as software development, data and artificial intelligence, software testing, cloud computing and technical support, electronics and communications, industrial automation, product and project management, design, and digital operations. Records from both platforms were standardized into a unified schema that included job titles, search categories, job descriptions, work experience requirements, platform details, locations, and other available structured fields. The dataset and associated reproducibility materials have been deposited in Zenodo under <https://doi.org/10.5281/zenodo.21638124> [15].

A study-specific corpus was selected because the research required current Chinese information technology recruitment text containing mixed Chinese–English technical terminology, complete vacancy descriptions, structured experience fields, and platform-specific search categories. Available benchmark datasets address a range of extraction tasks. Common foci include English-language skill spans, predefined skill classification, and mapping to occupational taxonomies. These resources are valuable for the tasks they target. But our study introduces requirements beyond their scope. We need Chinese-language topic discovery. We need expert involvement in constructing the competency index. And we need multilabel extraction at the posting level for all twenty-five indicators. Current benchmarks do not jointly provide these three capabilities. A purpose-built corpus was therefore necessary to

meet the research objectives, although it limits direct numerical comparison with studies based on established benchmark datasets.

Data cleaning was completed prior to text representation and topic modeling. A total of 554 invalid, incomplete, or out-of-scope records were excluded. Job descriptions were subsequently normalized, following which 839 exact duplicates identified within the normalized text and 90 near-duplicates were removed based on a predefined text-similarity criterion. An additional 34 records characterized by being very short or low in information content were also excluded. During normalization, mixed Chinese–English technical tokens, including Java, MySQL, TCP/IP, C++, C#, and Node.js, were preserved. The final analytical corpus comprised 12,513 postings, accounting for 89.19% of the original collection, as detailed in Table I.

TABLE I. CORPUS CONSTRUCTION AND CLEANING RESULTS

Stage	N	Description
Collected postings	14,030	Boss Zhipin: 11,011; Zhaopin: 3,019
Invalid/out-of-scope	554	Removed before deduplication
Exact duplicates	839	Normalized-text matches
Near duplicates	90	Similarity-based filtering
Low-information records	34	Insufficient usable tokens
Final analytical corpus	12,513	89.19% retained

Postings were collected using a custom Python crawler implemented with DrissionPage and Chromium browser automation. The crawler operated through an authenticated browser session and captured job-detail JSON responses generated while listings were viewed through the platform interface; no licensed official API was used. The crawler retrieved up to 100 postings per occupational category, inserted approximately 0.3 s between job-card selections and 1 s after each detail record, and paused for 3 min between occupational categories, with additional short delays during page scrolling. Only listing-level fields visible through the authenticated interface were retained, including job title, salary range, displayed skills, job description, location, work-experience requirement, company name and description, and listing URL. Based on the stored data schema, no applicant-level data, private-account data, or sensitive personal identifiers were collected.

C. Text Representation, Topic Modeling, and Stability

We converted each cleaned job posting into a unified recruitment text and represented it with TF-IDF unigrams and bigrams. The preprocessing pipeline included Chinese segmentation, stop-word removal, boilerplate filtering, and document-frequency filtering. Technical terms and domain-specific abbreviations were retained throughout.

NMF [16] was applied to the TF-IDF matrix to identify recruitment-demand topics. We selected the number of topics from a candidate range of $K=6$ to $K=26$, guided by reconstruction error, topic diversity, and normalized pointwise mutual information (NPMI) [17], which we combined through

an average-ranking procedure. For baseline comparison, we fitted LDA [18] to the corresponding count matrix and implemented BERTopic [19] with paraphrase-multilingual-MiniLM-L12-v2 embeddings, UMAP reduction, HDBSCAN clustering, and class-based TF-IDF. We calculated NPMI and topic diversity from the same preprocessed token corpus using each topic's top 12 terms, and evaluated all models on coherence, diversity, effective topic count, outlier rate, and runtime efficiency.

The stability of the topics was evaluated through a ten-run multi-seed analysis coupled with ten repeated 80% subsamples drawn without replacement. Topics from each refitted model were aligned with the reference solution via cosine similarity, whereas the overlap of top terms was assessed using Jaccard similarity.

D. Competency Index Construction

The competency index was developed through a systematic interpretation and consolidation of the $K=22$ NMF topics. The consolidation involved the same two reviewers described in Section III-G, with backgrounds in computer science and human-resource management. For each topic, the highest-weight terms, representative job postings, and predominant recruitment contexts were collectively analyzed. Topics were not used as direct signs of competency. Related topics were grouped, while broader topics were split up if they showed different competency needs. Recruitment templates and content that were not linked to competencies were removed. As a result, several topics might support a single indicator, and a single topic could support more than one indicator.

We retained candidate indicators that met three conditions. Each indicator had to reflect an identifiable employer requirement. It had to be conceptually distinct from other indicators. It had to be operationalizable for later annotation and extraction. We also included structured criteria such as education, professional qualifications, and work experience. These were added even when they did not emerge as independent topics. Topic prevalence and stability supported our decisions. We interpreted low-frequency or locally unstable topics with caution.

The resulting two-level competency hierarchy is presented in Section IV-B. It comprises five primary dimensions and 25 secondary indicators. We arrived at this structure through semantic consolidation and expert review of the retained indicators. For every secondary indicator, we defined four components: an operational description, inclusion and exclusion criteria, representative expressions, and relevant keyword or phrase patterns. This codebook underpinned all subsequent tasks, including manual annotation, rule-based extraction, and PC-CAGF development.

E. Rule, Semantic, and Context Evidence

Rule-based evidence was derived from indicator-specific lexical patterns and regular expressions developed from the operational codebook. For each job posting, the title, skill tags, structured work-experience field, and full description were examined against the rule sets corresponding to all 25 competency indicators. For each posting–indicator pair, a binary rule match and a rule-confidence score were generated based on

pattern specificity and evidence location. Direct references to technical skills, qualifications, certificates, or experience requirements were assigned higher confidence. In contrast, matches found within negated statements, company profiles, employee-benefit sections, or other standard recruitment templates were down-weighted or excluded if they did not indicate an actual requirement for applicants.

Semantic evidence was generated using the Sentence-BERT framework [20] and the pretrained multilingual model paraphrase-multilingual-MiniLM-L12-v2, which was developed through multilingual knowledge distillation [21]. Posts were segmented into sentence-level units and encoded together with the operational definitions of the 25 competency indicators. For each post and indicator pair, the maximum cosine similarity between the sentence embeddings and the corresponding indicator prototype was recorded as the semantic score. The pretrained encoder was kept in a frozen state, while probability calibration and indicator-specific threshold selection were conducted within the training folds.

Context evidence was constructed from the local and structural information surrounding each candidate indicator occurrence. For each posting-indicator pair, the features captured title-indicator compatibility, section relevance, action-object co-occurrence, negation, recruitment-template noise, and normalized text length. Matches appearing in responsibility or qualification sections and near action expressions such as “responsible for,” “develop,” “coordinate,” and “familiar with” were treated as stronger evidence, while occurrences limited to company introductions, welfare descriptions, or negated statements were treated as weaker evidence. These contextual features, together with rule-confidence and calibrated semantic

features, were included in a pooled logistic gate conditioned on one-hot indicator identity. Indicator-specific deployment was subsequently implemented through separate thresholds and No-Harm switches.

For baseline comparison, three rule-semantic fusion strategies were evaluated. Rule OR Semantic classified a posting-indicator pair as positive when either the rule-based extractor was positive, or the semantic score exceeded an indicator-specific threshold. Weighted Fusion combined rule confidence and calibrated semantic probability using fixed weights of 0.65 and 0.35, respectively. Logistic Stacking used a pooled L2-regularized logistic regression with rule output, calibrated semantic probability, and one-hot indicator identity as predictors. Indicator-specific thresholds for the fusion baselines were selected using inner-fold predictions to maximize F1 subject to the predefined precision constraint. All reported baseline results were obtained from held-out posting-level predictions.

F. PC-CAGF and Sentence-Level Specialists

The proposed PC-CAGF model operated at the posting-indicator level and combined rule-based, semantic, contextual, and specialist evidence, as illustrated in Fig. 2. For indicator j in posting d , a feature vector x_{dj} was constructed from the binary rule output, rule-confidence features, calibrated semantic probability, title and section relevance, action-object co-occurrence, negation and template-noise indicators, and normalized text-length features. For the two specialist indicators, additional sentence-level features were appended to the same vector. Indicator identity was represented using one-hot encoding.

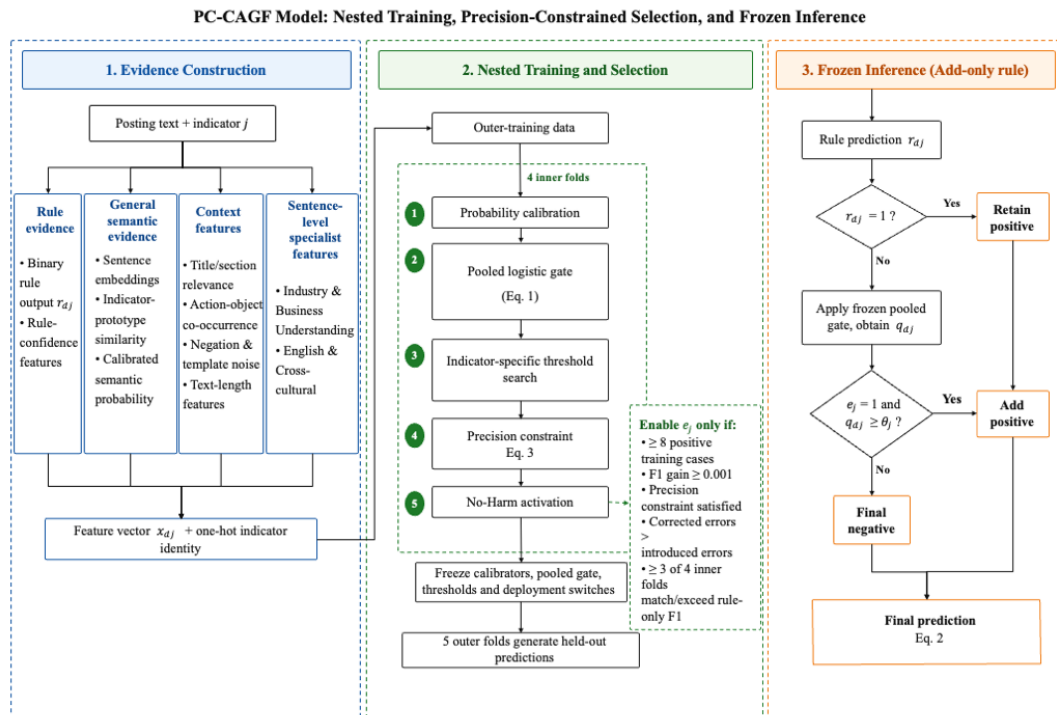


Fig. 2. Nested training, precision-constrained selection, and frozen inference of the PC-CAGF model.

Rather than fitting an independent gate for each indicator, PC-CAGF used a pooled logistic gate conditioned on indicator identity, which was defined as Eq. (1):

$$q_{dj} = P(y_{dj} = 1 \mid x_{dj}, j) = \sigma(\beta_0 + \beta^T x_{dj}) \quad (1)$$

where, q_{dj} is the gated probability for posting d and indicator j , and $\sigma(\cdot)$ denotes the logistic function. Because the indicator identity was included in x_{dj} , the pooled model could learn shared evidence patterns while accounting for indicator-specific differences.

PC-CAGF followed an add-only decision rule. Rule-positive cases were retained directly to preserve the high precision of explicit lexical evidence. For rule-negative cases, the pooled gate could add a positive prediction only when the corresponding indicator-level deployment switch was enabled and the gated probability exceeded the selected indicator-specific threshold. The decision was defined as Eq. (2):

$$\hat{y}_{dj} = r_{dj} \vee [e_j \mathbb{I}(q_{dj} \geq \theta_j)] \quad (2)$$

where, r_{dj} is the rule-only prediction, θ_j is the selected threshold for indicator j , $e_j \in \{0,1\}$ is the indicator-level No-Harm deployment switch, and $\mathbb{I}(\cdot)$ is the indicator function. Thus, semantic, contextual, and specialist evidence supplemented rather than replaced rule-based evidence.

Model selection was performed within a nested cross-validation design [22]. Five outer folds generated held-out predictions, whereas four inner folds were used for semantic-probability calibration, pooled-gate fitting, threshold selection, precision-constrained screening, and No-Harm activation. Candidate thresholds were drawn from a fixed grid between 0.05 and 0.95 in increments of 0.025 and supplemented with empirical probability quantiles. For indicator j , a threshold θ was considered feasible in Eq. (3), only when

$$P_j(\theta) \geq \max(0.60, P_{j,rule} - 0.02) \quad (3)$$

where, $P_j(\theta)$ is the inner-fold precision under the candidate threshold and $P_{j,rule}$ is the corresponding rule-only precision. Among feasible candidates, the threshold maximizing F1 was selected; ties were resolved sequentially using higher recall, higher precision, and a higher threshold.

The No-Harm mechanism determined whether gated augmentation was deployed for each indicator. The switch e_j was enabled only when the indicator contained at least eight positive training cases, the selected candidate improved F1 over the rule baseline by at least 0.001, satisfied the precision constraint in Eq. (3), corrected at least one more error than it introduced, and matched or exceeded the rule-only F1 in at least three of the four inner folds. Otherwise, e_j was set to zero and the indicator reverted to rule-only prediction. All calibration, gate fitting, threshold selection, and deployment decisions were therefore completed without using the corresponding outer held-out labels.

Two sentence-level specialist feature sets were developed for Business/Domain Understanding and English and Cross-cultural Competence, for which the general semantic representation produced comparatively low recall. Each posting

was segmented into sentences, and specialist features were constructed from indicator-specific lexical rules, positive and negative prototype similarities, prototype contrast, sentence-section relevance, negative evidence, and context-interaction features. The resulting raw specialist scores, calibrated specialist probabilities, and contextual features were appended to x_{dj} and evaluated by the same pooled logistic gate. The specialists were consequently incorporated as specialized feature sets into the pooled gate rather than functioning as independent classifiers or separate logical-OR prediction branches.

The specialist definitions, lexical rules, contrastive prototypes, contextual criteria, and fixed feature formulations were held constant during the reported nested evaluation. Only probability calibration, pooled-gate fitting, indicator-specific threshold selection, and No-Harm activation were learned from the corresponding training folds. Nested cross-validation was used only for performance estimation. After evaluation, a final deployment model was refitted on all 500 annotated postings, with cross-fitted predictions used to determine indicator-specific thresholds and No-Harm switches. The frozen model was then applied to all 12,513 postings solely for demand profiling.

For reproducibility, the analysis was conducted in Python 3.13.5 within an Anaconda environment on macOS using standard scientific-computing, NLP, topic-modeling, and visualization libraries. Chinese text was segmented with Jieba in accurate mode and a custom technical dictionary. Near duplicates were detected using three-token shingles, 64-bit SimHash, a maximum Hamming distance of 3, a minimum length ratio of 0.72, and a Jaccard threshold of 0.84. Fixed seeds were used for topic modeling, stability analysis, validation, bootstrap resampling, and PC-CAGF training.

G. Manually Annotated Reference Standard and Evaluation

A manually annotated reference standard was constructed from 500 job postings selected through proportional stratified sampling by platform and work-experience group. The two coders had backgrounds in computer science and human-resource management, and prior experience in recruitment-text analysis and qualitative coding. For each sampled posting, the two coders independently evaluated all 25 secondary indicators using a blind annotation template that did not expose automatic rule predictions, yielding 12,500 paired posting-indicator coding decisions. Before formal annotation, the coders reviewed the operational definitions, inclusion and exclusion criteria, and representative examples for all indicators to establish a common coding protocol. Each posting-indicator pair was assigned a binary label, where 1 indicated that the competency was explicitly or substantively expressed and 0 indicated its absence. Disagreements were resolved through consensus adjudication, and the resulting consensus labels were used as the reference standard. Pre-adjudication reliability was evaluated across the 12,500 paired decisions using percentage agreement and Cohen's kappa [23].

Model development used nested posting-level cross-validation, with five outer folds for held-out prediction and four inner folds for probability calibration, gate fitting, threshold selection, and No-Harm decisions. Posting-level multilabel stratification [24] preserved indicator prevalence and prevented information leakage. Indicator-specific thresholds were selected

to maximize F1 subject to a precision constraint, and semantic augmentation was activated only when it consistently improved inner-fold performance; otherwise, the rule-only output was retained. Final performance was assessed using pooled precision, recall, F1, and mean per-indicator F1. Indicator-level precision, recall, F1, and class support are reported in Supplementary Table VIII. Uncertainty and paired differences were examined using 1,000 posting-level bootstrap resamples [25], the exact McNemar test [26], and the Wilcoxon signed-rank test.

Supplementary group differentiation was examined using rule-based primary-dimension coverage, which was available independently of fitted-model predictions. The 30 most frequent search-category groups comprised 3,491 postings and reduced sparse contingency-table cells. Associations between search category and primary-dimension coverage were summarized using Cramér's V, with larger values indicating stronger differentiation. Experience-tier and platform associations were treated as supplementary heterogeneity evidence. After nested evaluation, the final PC-CAGF configuration was frozen and applied to all 12,513 postings; these complete-corpus outputs were used for estimated demand profiling and not as additional accuracy evidence.

IV. RESULTS

A. Corpus and Topic Modeling Results

Following preprocessing, 12,513 of the 14,030 collected job postings were retained, corresponding to a retention rate of 89.19% (Table I). These postings constituted the analytical corpus used for topic discovery and competency-index construction.

TABLE II. TOPIC-NUMBER SELECTION

K	Reconstruction Error	Topic Diversity	NPMI	Average Rank
6	108.4604	0.9583	0.2491	7.83
8	107.9624	0.9271	0.2876	8.33
10	107.4815	0.9500	0.3302	7.00
12	107.1024	0.9583	0.3618	5.50
14	106.7338	0.9286	0.3641	5.67
16	106.4059	0.9167	0.3521	6.83
18	106.0956	0.9120	0.3736	6.33
20	105.8280	0.9125	0.3706	6.00
22	105.5654	0.9167	0.3936	3.50
24	105.2939	0.9028	0.3848	5.33
26	105.0600	0.9135	0.3903	3.67

The candidate topic numbers ranged from $K = 6$ to $K = 26$. As shown in Table II, reconstruction error decreased continuously as K increased and therefore could not independently determine the optimal topic number. We evaluated candidate solutions using reconstruction error, topic diversity, and NPMI. Each candidate was ranked separately on these three criteria, with lower reconstruction error and higher topic diversity and NPMI considered preferable. The arithmetic mean of the three ranks was used as the average-rank criterion,

with tied values assigned average ranks. The $K = 22$ solution achieved the highest NPMI of 0.3936, maintained high topic diversity of 0.9167, and obtained the lowest average rank of 3.50. Although $K = 26$ produced a lower reconstruction error, its NPMI and topic diversity were slightly lower, resulting in an average rank of 3.67. Accordingly, $K = 22$ was selected as the best overall balance among reconstruction quality, semantic coherence, and topic diversity.

The selected NMF solution was subsequently compared with LDA and BERTopic. As reported in Table III, NMF attained the highest NPMI and topic diversity. LDA produced substantially lower NPMI and topic-diversity values, whereas BERTopic produced only 21 effective topics and classified 24.47% of the corpus as outliers. NMF was also markedly more computationally efficient, requiring 1.11 seconds in comparison to 9.57 seconds for LDA and 68.12 seconds for BERTopic. These findings support the utilization of NMF for subsequent topic analysis.

TABLE III. COMPARISON OF TOPIC-MODELING METHODS

Model	Requested Topics	Effective Topics	NPMI	Topic Diversity	Outlier Rate (%)	Run time (s)
NMF	22	22	0.3936	0.9167	N/A	1.11
LDA	22	22	0.1923	0.7273	N/A	9.57
BERTopic	22	21	0.3539	0.9095	24.47	68.12

The robustness of the $K = 22$ solution was evaluated using a ten-run multi-seed analysis, and ten repeated 80% subsamples drawn without replacement. Values are reported as mean $\pm$ standard deviation (SD). The multi-seed analysis yielded matched cosine and top-term Jaccard similarities of 0.9341 ± 0.0231 and 0.8281 ± 0.0574 , respectively, with 90.91% of matched topics reaching cosine similarity ≥ 0.90 . Under 80% subsampling, the corresponding values were 0.8939 ± 0.0318 and 0.7016 ± 0.0520 , with 80.00% of matched topics reaching the same threshold. These results indicate that the selected NMF topic structure was relatively stable across random initialization and repeated subsampling, although several low-prevalence topics showed weaker local stability.

B. Competency Index Results

Expert review of the $K = 22$ topic solution yielded a two-level competency hierarchy with five primary dimensions and 25 secondary indicators (Table IV). We consolidated overlapping topics, separated broad themes where appropriate, and excluded non-competency or template-related content. We also added structured requirements, including education, professional qualifications, and work experience, to the final hierarchy.

Table IV summarizes the final index, which covers business understanding, technical capabilities, engineering delivery, general professional competencies, and entry requirements. The technical and implementation-oriented indicators are most concentrated in Core Technical Skills and Engineering Delivery Competencies. In comparison, General Professional Competencies capture transferable workplace requirements that are less tied to specific technical domains. Entry Requirements

and Background represent explicit screening conditions frequently reported in structured recruitment fields. Together, the five dimensions provide a comprehensive representation of employer demand at both competency and qualification levels.

TABLE IV. FINAL HIERARCHICAL COMPETENCY INDEX

Primary Dimension	Code	Secondary Indicator
Business & Industry Understanding	S01	Data Analysis & Decision Support
	S02	Business/Domain Understanding
Core Technical Skills	S03	Software Development and Database Technologies
	S04	Front-end, Mobile, and Interactive Application Development
	S05	Cloud Computing, Network Operations, and Technical Support
	S06	Artificial Intelligence and Data Engineering
	S07	Embedded Systems and Electronic Hardware
	S08	Industrial Automation and Control Technologies
	S09	Performance Optimization & High Availability
Engineering Delivery Competencies	S10	Testing & Defect Management
	S11	Project Management & Execution
	S12	Documentation & Knowledge Management
	S13	Security & Compliance
	S14	Requirements Analysis & Solution Design
	S15	Coding Standards & Quality Awareness
	S16	Accountability & Results Orientation
General Professional Competencies	S17	Teamwork
	S18	Communication & Expression
	S19	Customer/Service Orientation
	S20	Learning Agility & Self-motivation
	S21	Stress Tolerance & Adaptability
	S22	English & Cross-cultural Competence
	S23	Education Level & Major
Entry Requirements & Background	S24	Certificates & Qualifications
	S25	Years of Work Experience

C. Annotation Reliability and Extraction Performance

The manually annotated reference set comprised 500 job postings and 12,500 binary posting-indicator decisions. The two coders agreed on 11,305 decisions (90.44%), while 1,195 initially disagreed, with Cohen's $\kappa=0.7728$. The disagreements were resolved through consensus adjudication, and the resulting consensus labels served as the reference standard for model evaluation.

As shown in Table V, the rule-only baseline achieved high precision (0.9402) but moderate recall (0.6577), whereas the semantic-only model improved recall at a substantial precision cost. Conventional fusion methods yielded limited improvements, with the Rule OR Semantic approach achieving

the highest baseline F1 score of 0.7853, compared to 0.7740 for rule-only extraction. PC-CAGF without specialist features improved F1 to 0.8182, and the final PC-CAGF achieved a precision of 0.9267, recall of 0.7378, pooled F1 of 0.8215, and mean indicator F1 of 0.7655.

TABLE V. OVERALL EXTRACTION PERFORMANCE OF BASELINES AND PC-CAGF

Method	Precision	Recall	F1	Mean Indicator F1
Rule only	0.9402	0.6577	0.7740	0.7050
Semantic only	0.5337	0.7358	0.6186	0.5348
Rule OR Semantic	0.9241	0.6828	0.7853	0.7157
Weighted Fusion	0.9289	0.6774	0.7835	0.7109
Logistic Stacking	0.9257	0.6814	0.7850	0.7147
PC-CAGF without specialists	0.9268	0.7323	0.8182	0.7562
Final PC-CAGF	0.9267	0.7378	0.8215	0.7655

Note: Precision, recall, and F1 are pooled positive-class metrics across all posting-indicator decisions; Mean Indicator F1 is the unweighted mean of the 25 indicator-level F1 values. Indicator-level precision, recall, F1, positive support, and negative support are provided in Supplementary Table VIII to make class imbalance and rare-indicator performance auditable.

Relative to rule-only extraction, the final model increased recall by 8.01 percentage points and pooled F1 by 4.75 percentage points, with a precision decrease of 1.35 percentage points. Compared with Rule OR Semantic, the strongest conventional fusion baseline, PC-CAGF improved pooled F1 by 3.62 percentage points and mean indicator F1 by 4.98 percentage points.

Bootstrap analysis (Fig. 3) yielded an F1 improvement of 4.75 percentage points with a 95% CI of 4.00–5.49 percentage points. The model corrected 281 errors made by the rule-only baseline while introducing 58 new errors. Both the exact McNemar test and the posting-level Wilcoxon signed-rank test indicated statistically significant paired improvement.

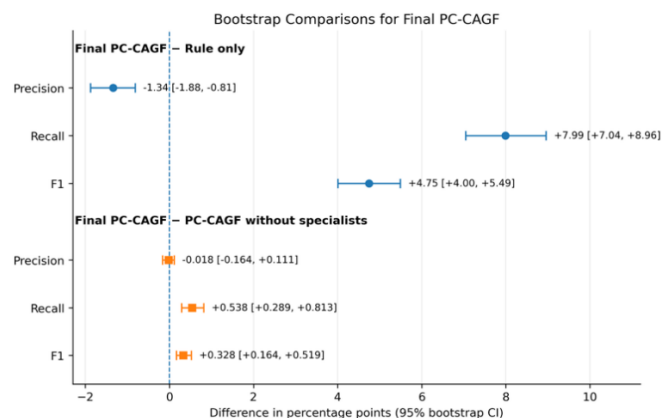


Fig. 3. Bootstrap differences between the final PC-CAGF and the rule-only and no-specialist configurations.

Indicator-level performance varied substantially (Supplementary Table VIII). Among indicators with both positive and negative reference labels, F1 ranged from 0.3037 for S02 Business/Domain Understanding (precision 0.9355, recall 0.1813) to 0.9613 for S23 Education Level and Major. S25 Years of Work Experience achieved an F1 of 1.0000, but all 500

reference labels were positive, precluding evaluation of negative-class discrimination. These results complement the F1 of 0.8215 and the unweighted mean indicator F1 of 0.7655.

D. Ablation and Specialist Analysis

Table VI reports the ablation results. Removing contextual features reduced recall from 0.7378 to 0.6788 and pooled F1 from 0.8215 to 0.7850, representing the largest degradation among the evaluated components. Removing the specialist feature sets reduced F1 to 0.8182, whereas disabling No-Harm reduced F1 slightly to 0.8207.

TABLE VI. ABLATION AND SENTENCE-LEVEL SPECIALIST RESULTS

Configuration	Precision	Recall	F1	Mean F1	F1 Difference from Final
PC-CAGF without context	0.9305	0.6788	0.7850	0.7155	-0.0365
PC-CAGF without specialists	0.9268	0.7323	0.8182	0.7562	-0.0033
PC-CAGF without No-Harm	0.9247	0.7378	0.8207	0.7647	-0.0008
Final PC-CAGF	0.9267	0.7378	0.8215	0.7655	—

The specialist feature sets produced targeted improvements for S02 Business/Domain Understanding and S22 English and Cross-cultural Competence (Fig. 4). For S02, recall increased from 0.1313 to 0.1813 and F1 from 0.2295 to 0.3037. For S22, recall increased from 0.3725 to 0.5490 and F1 from 0.5205 to 0.6747. Across the two indicators, the specialist features corrected 20 errors and introduced three. The exact McNemar test ($p=0.000488$) and posting-level Wilcoxon signed-rank test ($p=0.000674$) both indicated significant improvement. Because the two specialist indicators were selected during model development, their incremental gains are interpreted as exploratory; the broader contribution of PC-CAGF is supported independently by the no-specialist configuration, which increased pooled F1 from 0.7740 for rule-only extraction to 0.8182.

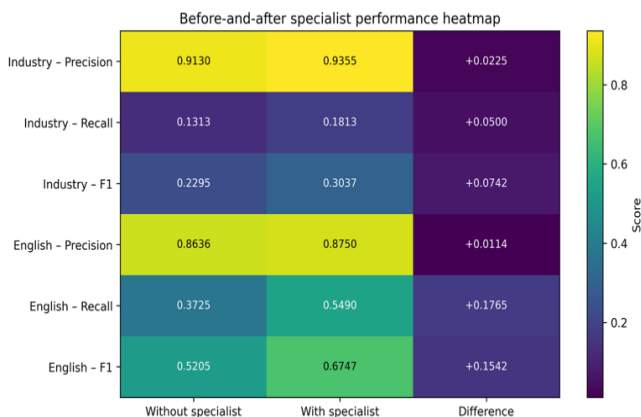


Fig. 4. Before-and-after performance of the sentence-level specialists.

E. Group Differentiation and Full-Corpus Application

Supplementary differentiation was examined across the 30 most frequent search-category groups, comprising 3,491 postings. Based on rule-derived primary-dimension coverage, the strongest differentiation was observed for Core Technical Skills ($V=0.623$), Business and Industry Understanding ($V=0.551$), and Engineering Delivery Competencies ($V=0.454$). The weaker associations for General Professional Competencies ($V=0.137$) and Entry Requirements and Background ($V=0.134$) indicate that these requirements were more broadly distributed across the selected search categories, as shown in Table VII. These findings provide supporting group-differentiation evidence rather than independent external validation of the fitted PC-CAGF predictions.

TABLE VII. SUPPLEMENTARY GROUP-DIFFERENTIATION ASSOCIATIONS AT THE PRIMARY-DIMENSION LEVEL

Primary Dimension	Experience Tiers	Search-Category Groups	Platform
Business & Industry Understanding	0.043	0.551	0.058
Core Technical Skills	0.193	0.623	0.205
Engineering Delivery Competencies	0.236	0.454	0.175
General Professional Competencies	0.047	0.137	0.081
Entry Requirements & Background	—	0.134	0.051

Note: Values are Cramér's V. Entry Requirements and Background were excluded from the experience-tier analysis to avoid circularity. Platform effects are reported as supplementary evidence of distributional heterogeneity.

After nested evaluation, the frozen PC-CAGF model was applied to all 12,513 postings. PC-CAGF increased the estimated number of positive indicator assignments from 61,532 to 69,858. This represented 8,326 additional assignments. As illustrated in Fig. 5, the largest increases involved Stress Tolerance and Adaptability (+2,674), Cloud Computing, Network Operations, and Technical Support (+1,271), and Requirements Analysis and Solution Design (+987). These additional assignments reflect model-generated estimates associated with contextual or paraphrased expressions, rather than independently verified competency occurrences. For S21 (Stress Tolerance and Adaptability), precision, recall, and F1 on the 500-posting reference set were 0.7895, 0.7317, and 0.7595, respectively (Supplementary Table VIII).

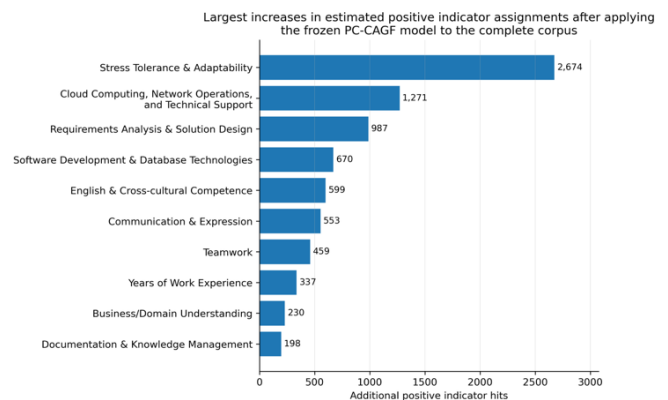


Fig. 5. Largest increase in estimated positive indicator assignments.

TABLE VIII. INDICATOR-LEVEL PERFORMANCE AND CLASS SUPPORT OF THE FINAL PC-CAGF MODEL

Code	Secondary Indicator	Positive Support	Negative Support	Precision	Recall	F1
S01	Data Analysis & Decision Support	57	443	0.8140	0.6140	0.7000
S02	Business/Domain Understanding	160	340	0.9355	0.1813	0.3037
S03	Software Development and Database Technologies	186	314	0.8862	0.7957	0.8385
S04	Front-end, Mobile, and Interactive Application Development	98	402	0.9048	0.7755	0.8352
S05	Cloud Computing, Network Operations, and Technical Support	144	356	0.7095	0.7292	0.7192
S06	Artificial Intelligence and Data Engineering	112	388	0.9722	0.6250	0.7609
S07	Embedded Systems and Electronic Hardware	77	423	0.8267	0.8052	0.8158
S08	Industrial Automation and Control Technologies	42	458	0.9200	0.5476	0.6866
S09	Performance Optimization & High Availability	109	391	0.9747	0.7064	0.8191
S10	Testing & Defect Management	126	374	0.9831	0.4603	0.6270
S11	Project Management & Execution	99	401	0.8679	0.4646	0.6053
S12	Documentation & Knowledge Management	119	381	0.9483	0.4622	0.6215
S13	Security & Compliance	59	441	0.9000	0.9153	0.9076
S14	Requirements Analysis & Solution Design	229	271	0.9444	0.5939	0.7292
S15	Coding Standards & Quality Awareness	28	472	0.9565	0.7857	0.8627
S16	Accountability & Results Orientation	154	346	0.9459	0.9091	0.9272
S17	Teamwork	181	319	0.9222	0.8508	0.8851
S18	Communication & Expression	239	261	0.9545	0.7908	0.8650
S19	Customer/Service Orientation	107	393	0.9744	0.7103	0.8216
S20	Learning Agility & Self-motivation	126	374	0.9318	0.6508	0.7664
S21	Stress Tolerance & Adaptability	164	336	0.7895	0.7317	0.7595
S22	English & Cross-cultural Competence	51	449	0.8750	0.5490	0.6747
S23	Education Level & Major	306	194	0.9896	0.9346	0.9613
S24	Certificates & Qualifications	39	461	0.8696	0.5128	0.6452
S25	Years of Work Experience	500	0	1.0000	1.0000	1.0000

Note: Precision, recall, and F1 were calculated from outer-fold out-of-fold predictions of the final PC-CAGF model. Positive and negative support denote the numbers of positive and negative consensus labels in the 500-posting reference set. S25 contained no negative cases in the reference set; therefore, its apparent perfect performance should be interpreted cautiously, and its negative-class discrimination could not be evaluated.

V. DISCUSSION

A. Interpretation of Main Findings

Our experiments show a clear tension between lexical precision and semantic coverage. A rule-based strategy delivered the highest precision, but it missed many implicit or paraphrased competency requirements. A semantic-only approach recovered more of these cases, yet its precision was notably lower. This tells us that semantic similarity alone is not enough. It struggles to separate applicant-related requirements from organizational or contextual descriptions. We also tested combinations of the two strategies, including Rule OR Semantic, weighted fusion, and logistic stacking. These produced only modest improvements, suggesting that simply layering lexical and semantic evidence does not resolve the underlying trade-off.

PC-CAGF achieved a more balanced outcome by using semantic evidence selectively. The gain in recall was substantial,

while the precision loss remained small. The ablation study confirmed this reading. Removing contextual information caused the largest F1 drop. Removing specialist features or the No-Harm mechanism had noticeably smaller effects. This pattern points to a specific conclusion. The primary strength of the framework lies in its selective use of semantic evidence, guided by contextual information and indicator-specific decision constraints.

B. Methodological Contribution of PC-CAGF

The methodological contribution of PC-CAGF lies in the controlled integration of complementary evidence sources. Previous recruitment-text studies have demonstrated the value of topic modeling [7], [13], contextual and transformer-based skill extraction [8], [9], [12], [14], and weakly supervised or taxonomy-informed approaches [10], [11]. The present framework does not replace these representations with a new embedding architecture. Instead, it addresses a different problem: how to introduce semantic predictions without

unnecessarily sacrificing the reliability of explicit lexical evidence.

Four design choices distinguish PC-CAGF from conventional fusion. First, rule-positive decisions are preserved through add-only inference. Second, a pooled logistic gate learns shared evidence patterns while conditioning predictions on indicator identity. Third, semantic augmentation is controlled by indicator-specific thresholds selected under a precision constraint. Fourth, the No-Harm switch disables augmentation when inner-fold evidence does not demonstrate sufficient benefit. The nested evaluation design further separates calibration and deployment decisions from assessment on held-out postings. Together, these mechanisms make the framework particularly suitable for settings in which labeled data are limited and false-positive competency assignments have practical consequences.

C. Practical Implications

The hierarchy gives us a structured view of employer demand that scales to large recruitment corpora. We can aggregate the extracted indicators by search category, platform, experience tier, or collection period to examine demand variation across these dimensions. This flexibility is supported by the framework's design. The same indicator definitions carry through from annotation to rule construction, semantic prototypes, and automated extraction. This design supports a consistent mapping from raw vacancy text to interpretable competency profiles.

The framework may additionally support curriculum and employability applications by linking extracted employer requirements with evidence from courses, projects, certifications, internships, and student outcomes. Such applications should, however, be treated as downstream uses rather than as evidence of predictive validity in the present study. The current analysis establishes the demand-side representation and evaluates its extraction accuracy; direct relationships between the extracted competencies and graduate employment outcomes remain to be tested.

D. Limitations and Generalizability

Several limitations should be acknowledged. The corpus was collected from two Chinese recruitment platforms within a relatively short period, which may limit temporal, regional, and cross-platform generalizability. The annotated reference set contained 500 postings, and the known-group analysis was restricted to the 30 most frequent search-category groups, covering 3,491 postings; therefore, rare competencies and low-frequency occupations may be underrepresented. In addition, search categories are not equivalent to standardized occupational classifications, and several narrow technical topics showed weaker local stability. The rule codebook, semantic prototypes, thresholds, and specialist feature sets were also developed for Chinese information technology recruitment text and may require recalibration for other industries or languages. The evaluation also involves potential dependence among observations. Each posting contributes 25 indicator decisions, and postings from the same company may share boilerplate text. Although the posting-level bootstrap preserves the grouping of indicator decisions within each resampled posting, it does not account for potential dependence across postings from the same

company. The reported paired-test p-values should therefore be interpreted cautiously.

Future research should extend the corpus across longer periods, additional platforms, regions, languages, and occupational sectors. Standardized occupational taxonomies, including ESCO or locally applicable occupational classifications, could be used to strengthen external validation and cross-study comparability [4], [5]. Larger, independently annotated datasets would improve the assessment of rare indicators and support broader development of specialist feature sets. The framework should also be linked with student-level evidence, including courses, projects, certificates, internships, and employment outcomes, to determine whether demand-aligned competency profiles can explain or predict graduate employability. Such extensions would strengthen both the generalizability of the index system and its practical value for curriculum improvement and for targeted educational intervention.

VI. CONCLUSION

Extracting structured competency requirements from job postings requires balancing lexical precision, semantic coverage, and false-positive control. We tackled this through three integrated components: topic-based demand discovery, expert consolidation, and operational competency extraction. From 14,030 Chinese IT job postings collected across two platforms, we retained 12,513 records after cleaning and deduplication. Multi-criterion topic selection and stability analysis supported a 22-topic NMF solution, which experts consolidated into five primary dimensions and 25 secondary indicators.

We evaluated the framework using 500 manually annotated postings and 12,500 posting-indicator decisions. PC-CAGF achieved a precision of 0.9267, recall of 0.7378, pooled F1 of 0.8215, and mean per-indicator F1 of 0.7655. Relative to rule-only extraction, recall rose by 8.01 percentage points and F1 by 4.75 percentage points, with a precision decrease of only 1.35 percentage points. Ablation analysis identified contextual features as the main source of improvement, while sentence-level specialist features added gains for two context-dependent indicators. Unlike topic-only or isolated extraction approaches, our workflow links demand-driven discovery with posting-level extraction evaluated against a manually annotated reference set. The main contribution is an interpretable, auditable, precision-constrained architecture that preserves rule-based evidence while using semantic predictions selectively. The resulting hierarchy may support recruitment-text analysis and labor-market monitoring, while its use in curriculum mapping and competency-gap analysis remains a potential downstream application. We acknowledge that our findings are based on internal validation within Chinese Information Technology (IT) postings and should not be extrapolated directly to other periods, platforms, industries, regions, or languages. Future work should test the framework using temporal and cross-platform holdouts, larger annotated datasets, public benchmarks, and standardized taxonomies, and examine whether these demand-side profiles can be linked with student-level evidence and employment outcomes.

ACKNOWLEDGMENT

The authors thank the coders who contributed to the independent annotation and adjudication of the manually annotated reference set. The authors also acknowledge the institutional and technical support provided during data processing, model development, and manuscript preparation.

DECLARATIONS

This study analyzed job advertisements and did not involve experiments on human participants, clinical data, or applicant-level research records. Only job-related fields visible in the source listings were retained, and no applicant personal data, private account data, or sensitive personal identifiers were collected. Analyses were performed at the posting-indicator or aggregate level.

During manuscript preparation, the authors used Grammarly for language polish. The authors reviewed and edited all generated output, verified the technical content against the study records, and took full responsibility for the final manuscript. No generative-AI output was used as research data or a substitute for the reported statistical analysis.

REFERENCES

- [1] A. Eimer and C. Bohnick, "Employability models for higher education: A systematic literature review and analysis," *Social Sciences & Humanities Open*, vol. 8, no. 1, 100588, 2023, <https://doi.org/10.1016/j.ssaho.2023.100588>.
- [2] W. G. Alheadary, "Controlling Employability Issues of Computing Graduates through Machine Learning-Based Detection and Identification," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10888–10894, Jun. 2023, doi: 10.48084/etasr.5892.
- [3] A. E. Yahya, W. M. S. Yafooz, and A. Gharbi, "Mapping Graduate Skills to Market Demands: A Holistic Examination of Curriculum Development and Employment Trends," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 14793–14800, Aug. 2024, doi: 10.48084/etasr.7454.
- [4] European Commission, "Skills & competences," *European Skills, Competences, Qualifications and Occupations (ESCO)*, version 1.2.1. [Online]. Available: https://esco.ec.europa.eu/en/classification/skill_main. [Accessed: Jul. 14, 2026].
- [5] F. Chiarello, G. Fantoni, T. Hogarth, G. Giordano, L. Baltina, and I. Spada, "Towards ESCO 4.0: Is the European classification of skills in line with Industry 4.0? A text mining approach," *Technological Forecasting and Social Change*, vol. 173, Art. no. 121177, 2021, doi: 10.1016/j.techfore.2021.121177.
- [6] L. Alekseeva, J. Azar, M. Giné, S. Samila, and B. Taska, "The demand for AI skills in the labor market," *Labour Economics*, vol. 71, Art. no. 102002, 2021, doi: 10.1016/j.labeco.2021.102002.
- [7] E. Akkol, M. Olucoglu, and O. Dogan, "Topic modeling for skill extraction from job postings," in *Knowledge Graphs and Semantic Web*, LNCS, vol. 14382, Springer, 2023, pp. 277–289, doi: 10.1007/978-3-031-47745-4_20.
- [8] A. Akkasi, "Job description parsing with explainable transformer-based ensemble models to extract the technical and non-technical skills," *Natural Language Processing Journal*, vol. 9, Art. no. 100102, 2024.
- [9] F. Leon, M. Gavrilescu, S.-A. Floria, and A. A. Minea, "Hierarchical classification of transversal skills in job advertisements based on sentence embeddings," *Information*, vol. 15, no. 3, Art. no. 151, 2024.
- [10] D. C. Kavargyris et al., "ESCOX: A tool for skill and occupation extraction using LLMs from unstructured text," *Software Impacts*, vol. 25, p. 100772, 2025.
- [11] M. Zhang, K. N. Jensen, R. van der Goot, and B. Plank, "Skill Extraction from Job Postings using Weak Supervision," in *Proc. 2nd Workshop on Recommender Systems for Human Resources (RecSys in HR 2022)*, CEUR Workshop Proceedings, vol. 3218, Seattle, WA, USA, 2022.
- [12] I. Ifakir, N. Mohtaram, E. H. Nfaoui, A. Zannou, and M. El Hassouni, "An Approach Based on Named Entity Recognition and Semantic Analysis for Recruitment Efficiency and Optimization," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 8, 2025, doi: 10.14569/IJACSA.2025.0160889.
- [13] Z. Ao, G. Horváth, C. Sheng, Y. Song, and Y. Sun, "Skill requirements in job advertisements: A comparison of skill-categorization methods based on wage regressions," *Information Processing & Management*, vol. 60, no. 2, Art. no. 103185, Mar. 2023, doi: 10.1016/j.ipm.2022.103185.
- [14] M. Zhang, K. N. Jensen, S. D. Sønniks, and B. Plank, "SkillSpan: Hard and Soft Skill Extraction from English Job Postings," in *Proc. 2022 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Seattle, WA, USA, 2022, pp. 4962–4984, doi: 10.18653/v1/2022.naacl-main.366.
- [15] Fengyuan. shi, "Job Position Dataset(Boss Zhipin and Zhaopin)". Zenodo, Jul. 28, 2026. doi: 10.5281/zenodo.21638124.
- [16] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
- [17] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. WSDM*, 2015, pp. 399–408.
- [18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [19] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022, doi: 10.48550/arXiv.2203.05794.
- [20] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992.
- [21] N. Reimers and I. Gurevych, "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4512–4525, 2020, doi: 10.18653/v1/2020.emnlp-main.365.
- [22] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.
- [23] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.
- [24] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the Stratification of Multi-label Data," in *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2011, Proceedings, Part III*, LNCS, vol. 6913, Springer, 2011, pp. 145–158, doi: 10.1007/978-3-642-23808-6_10.
- [25] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.
- [26] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, pp. 153–157, 1947.