Techno-Science Research Journal V14 (Special Issue) (2026) P 48-59

Content list available at ITC

Techno-Science Research Journal

Journal Homepage: http://techno-srj.itc.edu.kh/

Techno-Science Research Journal

A Computational Approach to Labor Market Analytics: Analyzing Skill Trends in Cambodia Using an LLM-Based Pipeline

Sela Songeam*, Dona Valy, Sothea Has

Research and Innovation Center, Institute of Technology of Cambodia, Russian Federation Blvd., P.O. Box 86, Phnom Penh, Cambodia

Received: 12 September 2025; Revised: 15 October 2025; Accepted: 28 October 2025; Available online: 31 July 2026

Abstract: *The rapid evolution of labor markets, particularly in emerging economies like Cambodia, necessitates a dynamic and real-time skill demand analysis. Traditional methods, such as employer surveys, are often static and infrequent, failing to capture the fluidity of an evolving workforce. This research addresses this data gap by presenting a novel computational pipeline that leverages recent advancements in Large Language Models (LLMs) and Natural Language Processing (NLP). The proposed framework integrates multilingual text processing, LLM-driven information extraction, and taxonomy-based skill alignment using internationally recognized standards such as the European Skills, Competences, Qualifications and Occupations (ESCO) and the World Economic Forum (WEF) job taxonomy. Evaluation results demonstrate strong extraction performance for key job fields, while skill extraction shows high conceptual alignment despite lower strict-match accuracy. The findings confirm the potential of LLM-based methods to enhance labor market intelligence and provide actionable insights for policy, education, and workforce development in Cambodia.*

Keywords: Job Market, Web Scraping, Large Language Model (LLM), Natural Language Processing (NLP)

1. INTRODUCTION

Skills is the foundation of job performance and employability. In today's rapidly evolving of digital landscape, new skills are continually emerging and existing ones are being updated to keep pace with technological advancements and the introduction of new tools. This has become a challenge for job seekers, especially freshly graduated students, who constantly need to learn and upgrade their knowledge and skills for their career goal. Without seeking proper skill development guidelines from teacher or experienced professionals, job applicants most likely ended up in the skills mismatch between them and their employer's needs.

The challenge has consistently persisted in Cambodia's job market, according to the survey by EUROCHAM which conducted a Skill Gap Assessment across 106 companies in over 11 sectors in Cambodia in 2024; approximately 74% of companies reported difficulties in finding qualified staff. A significant portion of these issues, about 60% were attributed to area-specific labor shortages and general skill gaps [1]. While such survey-based measurements are useful, they often lack accessibility, remain static in format, and are not updated frequently. Conventional skill gap analyses rely heavily on employer surveys, which are limited by scope, timeliness, and potential response biases.

Conversely, online recruitment postings provide a real-time, large-scale, and directly observable record of labor demand. They capture evolving trends by industry, offering a more dynamic and accurate view of required competencies. However, extracting meaningful insights from such diverse and unstructured data is technically challenging, especially in Cambodia's context, where multilingual job descriptions, non-standard classifications, and inconsistent terminology are common.

To address these limitations, our research proposes a data-driven approach to skill trend analysis by leveraging web scraping techniques to collect real-time job posting data from major online job platforms in Cambodia. Unlike traditional surveys that are limited in scope and frequency, job postings provide continuously updated insights into the labor market and reflect the actual demand for skills across various industries. By extracting and analyzing job descriptions using Large Language Model (LLM) and skill taxonomies, we aim to uncover emerging skill trends and a scalable, transparent, and automated way to support growing skill insight for inform policy decisions, curriculum

* *Correspondence: sela_songeam@gsc.itc.edu.kh*

DOI: https://doi.org/10.64702/Techno-SRJ.2026.V14.SI.06

Published by Institute of Technology of Cambodia.
© 2026 INSTITUTE OF TECHNOLOGY OF CAMBODIA. All rights reserved.
This article is an open access article under CC BY-NC-ND 4.0 license.

development, and empower job seekers with actionable information.

Contribution: In this work, we seek to contribute to, **(1)** Utilizing web scraping framework of the public available job posting data across sectors in Cambodia. **(2)** Applying Large Language Model (LLM) with a few-shot prompting to extract job information from scraped HTML page, and skill extraction from job description. **(3)** Adapted standardized terminology of job category and skills. **(4)** Building a full automated pipeline for collecting job posting and visualize skill trends.

2. RELATED WORK

In this section, we review related works, in the area of occupational-related datasets, general occupational data mining and analysis, and skill extraction.

Online job market data sources such as LinkedIn, Indeed, and Glassdoor; along with proprietary platforms like Burning Glass Technologies (now Lightcast) are widely used for analyzing labor market trends, evolving qualification requirements, and skills demand [2,3,4,5]. While these datasets provide structured access, their commercial and proprietary nature limits transparency and reproducibility. In the Cambodian context, there is no centralized, open-access repository of job postings. Instead, labor market insights rely primarily on surveys conducted by the National Employment Agency or EuroCham Cambodia [6,1], which, though informative, are infrequent and depend on self-reported data, introducing potential biases. This gap highlights the need for automated approaches capable of collecting more representative and timely data from multiple recruitment platforms.

Data collection via web scraping offers a scalable and efficient solution to this challenge. The exponential growth of online information has made the web a critical source for real-time labor market data acquisition [7,8]. Scraping techniques are broadly divided into static parsing of HTML and dynamic rendering via headless browsers such as Selenium to process JavaScript-generated content. Although API-based access is preferable for its structured format, hybrid approaches are often required in practice. Tools such as BeautifulSoup, Scrapy, and Selenium are commonly employed. Scraping pipelines are frequently integrated with NLP methods to transform raw postings into structured datasets through preprocessing steps (tokenization, lemmatization, stopword removal) and advanced techniques like Named Entity Recognition (NER) and topic modeling [8]. However, challenges persist, particularly in adapting to frequent website layout changes and heterogeneous formatting.

Skill extraction from job postings is central to computational labor market analytics, enabling the identification and structuring of skill mentions for applications such as workforce analytics, curriculum design, and skill gap detection. Early methods relied on rule-based matching and curated dictionaries, such as Bastian et al. (2014) [9], who conducted topic extraction from LinkedIn profiles. While effective for frequent skills, these approaches struggled with synonyms, multi-word expressions, and contextual variations. More recently, embedding-based and deep learning approaches have advanced the field. A major milestone was the release of the SkillSpan benchmark [10], which includes 14,500 sentences annotated with over 12,500 skill mentions. Evaluations with BERT-based models showed that domain-adaptive pre-training significantly improved accuracy, and single-task models outperformed multi-task frameworks. Beyond extraction, normalization and taxonomy alignment are crucial. The JAMES framework [11] demonstrated how graph embeddings can normalize job titles and improve skill–taxonomy mapping. Synthetic data generation frameworks such as JOBSKAPE [12] have been introduced to address training data scarcity, while Senger and Zhang (2024) [13] highlight two recent trends: aligning extracted skills with standardized taxonomies such as ESCO to ensure interoperability, and adopting LLMs to overcome limited annotated data. Despite progress, extracting implicit skills remains an open research challenge.

Recent advances in Large Language Models (LLMs) have transformed information extraction from unstructured text, including job postings. Early neural approaches such as BERT and RoBERTa introduced contextualized embeddings that significantly improved entity recognition and semantic matching for skills [10, 13]. However, their limited context window and domain-specific pretraining requirements restricted scalability across diverse labor datasets. The emergence of instruction-tuned LLMs (e.g., GPT-3, LLaMA, Mistral, and Qwen families) enables zero-shot and few-shot skill extraction without extensive manual annotation. Recent studies (e.g., Chen et al., 2024) highlight how LLM-based frameworks can effectively automate complex entity alignment (EA) tasks across domains, demonstrating broader contextual reasoning and adaptability that can also benefit labor market text analysis [14]. These models leverage richer contextual understanding and multilingual capability, making them particularly valuable for analyzing heterogeneous job postings from multiple sources.

Occupational and skill taxonomies provide the structured frameworks necessary for harmonizing extracted data. The International Standard Classification of Occupations (ISCO-08), developed by the International Lab-

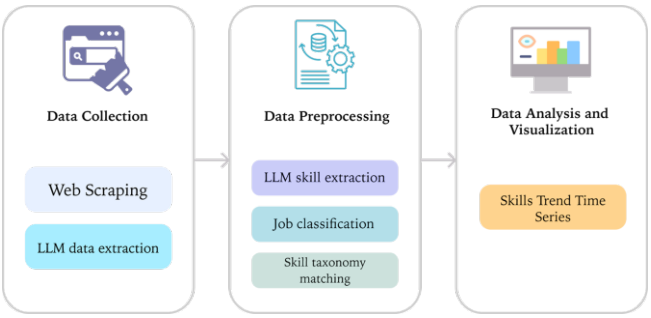

Fig. 1. Pipeline Overview

our Organization [15], remains a cornerstone for statistical labor reporting and cross-country comparability. However, ISCO's broad aggregation limits its granularity for analyzing dynamic or emerging roles. The U.S. Occupational Information Network (O*NET) offers detailed mappings of occupations to tasks, skills, and knowledge domains [16], but its reliance on subjective rating scales has raised concerns about interpretability and behavioral specificity [17]. The European Skills, Competences, Qualifications and Occupations (ESCO) framework extends ISCO by linking occupations directly with standardized skills and qualifications, supporting multilingual, semantic- cally structured labor intelligence [18]. However, since it reflects primarily European labor structures, contextual adaptation is often needed when applied in other regions such as Cambodia, where informal employment and non-standardized job titles are common [19]. In parallel, the World Economic Forum (WEF) job taxonomy from The Future of Jobs Report 2025, highlights emerging domains such as Technology and Digital, Green Jobs, and the Care Economy [20], offering a future-oriented perspective on global skill trends. While its emphasis is on global megatrends rather than local contexts, it complements ESCO by providing insights into occupations and skills projected to transform in the coming decade.

While existing work has advanced methods for labor market analysis, applications in developing economies remain limited. This study addresses that gap by applying automated data collection and skill extraction to Cambodian job postings, aligning results with both ESCO and the WEF Jobs taxonomy to provide context-specific and forward-looking insights.

3. METHODOLOGY

The study employs an automated job posting analysis approach to examine labor market skill trends in Cambodia. We implement an LLM-based analytical pipeline that integrates web scraping, natural language processing, and taxonomy alignment. As shown in **Fig. 1**, the methodology is organized into three modules: **(1)** a data acquisition layer that collects job postings from multiple Cambodian recruitment platforms, **(2)** a preprocessing layer that structures and cleans job descriptions, a skill extraction layer that applies Large Language Models (LLMs) for identifying skills, and a taxonomy alignment layer that standardizes extracted skills using ESCO and maps occupations to the WEF jobs taxonomy, **(3)** build skill timeline trend.

3.1 Data Collection Pipeline

We now introduce our job data collection pipeline, shown in **Fig. 2**. Our method consists of 2 steps: **1)** the job collection process in the job listing sites; **2)** collecting job detail of the job and save the final data to the database storage. Each job record in the dataset follows a structured JSON schema that defines the fields extracted by the LLM models. These fields include *job_title*, *company_name*, *posted_date*, *job_url*, *job_category*, *job_location, employ-ment_type*, and *company_industry*. **Table 1** summarizes the meaning of each key used in the prompts.

Step1: Job Listing Collection: In the first step, we targeted job posting websites in Cambodia, including LinkedIn, BongThom, and CamHR. We accessed these sites using Python Selenium, a powerful tool for web scraping, which allowed us to simulate user interactions such as scrolling and clicking buttons to load additional pages and collect more data in a single run. To extract only the job listing elements, we used Python BeautifulSoup to remove irrelevant content, such as JavaScript and CSS, and filter elements containing the job listings. The cleaned job elements were then passed to a large language model (LLM) along with a few-shot prompting strategy, instructing the model to return the data in a structured JSON format. Specifically, we used a structured prompt that included:

- **Input Text** – the HTML content of a job card.
- **Prompt Instruction:** Given the HTML content below containing a job posting, extract the following details include *job_title*, *company_name*, *job_url*, and *posted_date*.
- **Rule:** Output must be a valid JSON object, containing the keys "*job_title*", "*company_name*", "*posted_time*", and "*job_url*". Only response in a Json Object with No any additional text.

Step2: Job Detail Extraction: In this step, we used the job data obtained from the previous step, which included the *job_url* for accessing each job detail page. Using the same tools (Python Selenium and BeautifulSoup), we extracted

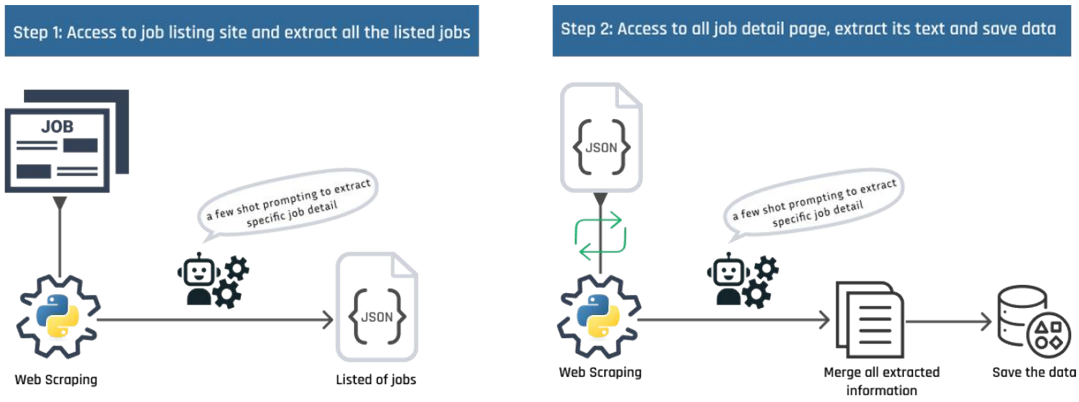

Fig. 2. 2 steps pipeline of job data collection. We illustrate our pipeline of web scraper with the use of LLM for dynamic data extraction by using a few shots prompt engineer.

Table 1 Description of extracted fields

JSON Key	Description
job_title	The title of the advertised position.
company_name	The name of the employer or hiring organization.
posted_date	The date or relative time when the listing was published.
job_url	The hyperlink to the detailed job description page.
job_category	The occupational group or functional area of the job.
job_location	The geographic or remote location of the job.
employment_type	The type of employment (e.g., full-time, part-time, contract).

only the visible text from each page. Next, we applied a large language model (LLM) to extract specific details from the job description, such as *job_category*, *company_industry*, and *job_location*. The prompt for this step was designed to be precise, ensuring the model focuses only on the necessary information:

- **Input Text** – the full text content of the job page.
- **Prompt Instruction:** Given the job page content below, extract the following details include *job_category*, *job_location*, and *employment_type*.
- **Rule:** Output must be a valid JSON. Output exactly a JSON element, containing the keys "*job_category*", "*job_location*", "*employent_type*", and "*company_industry*". Only response in A Json Object with No any additional text.

3.2 Skill Extraction Approach

In this study, skills were extracted from unstructured job postings using a Large Language Model (LLM) approach, as illustrated in **Fig. 3**. After preprocessing, each job description was submitted to the model with a structured prompting strategy that requires the model to return all identified skills as a comma-separated list. This design ensures that extracted skills can be directly passed into the subsequent taxonomy alignment stage. The choice of an LLM-based approach reflects the limited availability of annotated corpora and the model's ability to capture both explicit and implicit skills, which traditional dictionary or rule-based methods often fail to detect. Specifically, each job description was submitted to the model with a structured prompt that included:

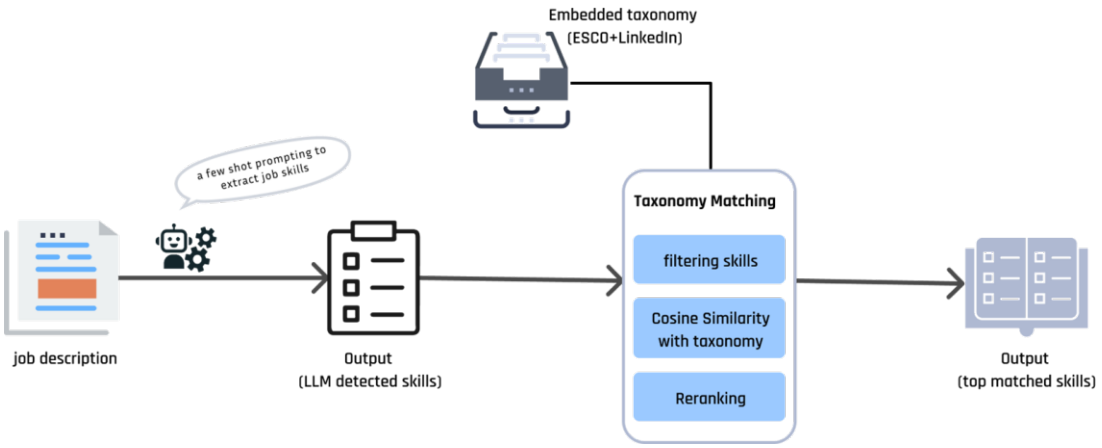

Fig. 3. Skill extraction from job description. In this pipeline we use LLM as a skill detector from the job description by using a few shots prompting. The output from LLM, then need to go through 3 steps process of taxonomy matching: skill filtering, compute match cosine similarity with stored embedded vector taxonomy, use reranker method to reranked the matched skills, and we choose to save the skills with the top score.

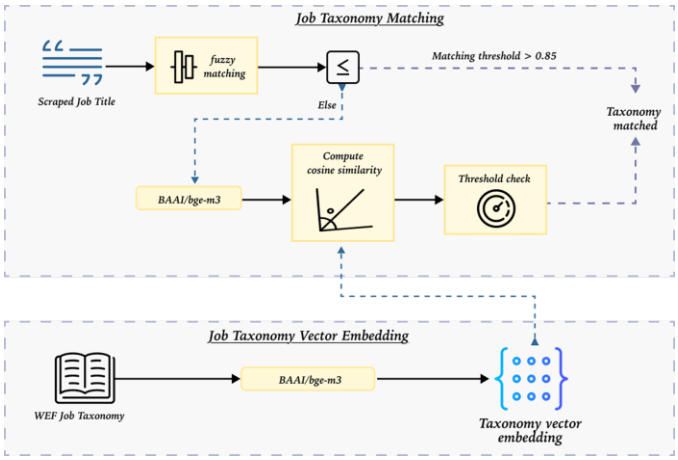

Fig. 4. WEF Taxonomy classification

- **Input Text –** Job description
- **Prompt Instruction:** Analyze the following job description and extract ALL skills mentioned. Include technical skills, tools, programming languages, frameworks, and skills requirements.
- **Rule:** Return only a simple list of skills, one skill per line, no numbering, no bullets, no extra text. Just the skill names, separate by comma.

To mitigate the risk of LLM hallucinations, a post-processing validation step was applied to the extracted skills. Specifically, each predicted skill was cross-checked against the original job description, and only those explicitly present in the source text were retained. To ensure that the extracted skills align with industry-standard terminology, we use the European Union's European Skills, Competences, Qualifications and Occupations (ESCO) framework, combined with LinkedIn's Skill Taxonomy, one of the leading hiring platforms globally. This combination expanded our skills vocabulary to approximately 13,890 skills from ESCO and 1,800 skills from LinkedIn, enabling our pipeline to cover a broader range of skills.

3.3 Job Classification Framework

While the collected data contained job categories, these classifications were largely inconsistent. To standardize the occupational categorization, we adopt the job taxonomy from the World Economic Forum's Future of Jobs Report 2025, a modified and extended version of the O*NET SOC classification. Our data classification follows the hierarchical structure where each occupation is categorized into:

(1) **Job Family:** Broad occupational categories (e.g., "Architecture and Engineering").
(2) **Occupation:** Specific roles within each family (e.g., "Civil Engineers", "Environmental Engineers")

As mentioned, job titles in scraped postings were often inconsistent, abbreviated, or context-specific; therefore, a two-stage hybrid classification approach was implemented. **Fig. 4** illustrates the algorithm for our job classification. In the first stage, fuzzy string matching was applied to map job titles to taxonomy occupations using Levenshtein distance [21], with matches above a 0.85 similarity threshold directly assigned to the corresponding category. The similarity threshold was set at 0.85 to balance precision and recall in the fuzzy matching stage. This choice aligns with prior studies

Table 2. Statistic of text input to LLM for job information extraction

Average per input	**Job information card**	**Job detail page**
Characters	1103.5	5997.1
Words	64.8	814.1

on fuzzy string matching, which commonly employ thresholds between 0.8 and 0.9 to capture near-identical variants while minimizing false matches [22]. For cases where fuzzy string matching did not meet the selected threshold, a second stage performed semantic similarity matching using vector embeddings. **BGE-M3** model was selected because it generates multi-lingual, context-sensitive embeddings [23], suitable for English–Khmer code-mixed job titles. Each job title and taxonomy label was encoded into vectors, and cosine similarity was computed to identify the most semantically aligned occupation. A separate similarity threshold was then applied to validate the match, where the top-ranked occupation was selected only if its cosine similarity score exceeded 0.4, as lower scores often indicated weak or irrelevant semantic matches. This two-stage process improved classification accuracy across heterogeneous job titles.

4. EXPERIMENT SETUP

4.1 Data Collection Evaluation

Experimental Design: The data collection pipeline evaluation focuses on BongThom.com as the primary testing platform utilizing three LLM models: *mistral:7b*, *Llama-SEA-LION-v3-8B-IT*, and *qwen3:8b*. In this study, these three models were selected based on their complementary strengths. *Mistral:7B* is a compact, open-weight model optimized for efficiency and high inference speed, making it suitable for large-scale extraction tasks [24]. *Llama SEA-LION-v3-8B-IT* is fine-tuned for Southeast Asian languages, providing improved handling of regional multilingual content, including English–Khmer code-mixing common in Cambodian job postings [25]. *Qwen3:8B*, developed by Alibaba, integrates strong multilingual and instruction-following capabilities, enabling robust skill phrase interpretation and semantic normalization [26]. Collectively, these models were chosen to balance efficiency, multilingual coverage, and generalization, ensuring reliable performance in Cambodia's diverse digital labor landscape. **Table 2** presents statistical insights into the average text input sizes processed by the LLM models, providing context for the scope and complexity of the extraction task.

Ground Truth: To evaluate LLM-based extraction accuracy, we established a ground truth dataset through manual extraction using Python scripts that directly parse HTML elements via developer tools and tag inspection. This This manual extraction serves as the baseline for comparing LLM-generated outputs from the same job postings' cleaned text.

Evaluation Metrics: The evaluation employs multiple metrics including field-level precision, recall, and F1-score to assess the reliability of LLM-based structured information extraction in real-world web scraping applications.

As presented in **Table 3**, the three evaluated LLMs demonstrate clear performance differences in extraction performance across different field types. Overall, *qwen3:8b* achieved the highest and most consistent results, attaining near-perfect F1-scores in well-structured fields such as *job_url* (F1 = 1.00), *employment_type* (F1 = 0.99), and *job_location* (F1 = 0.95). *mistral:7b* demonstrated similarly strong performance, particularly for *job_title* (F1 = 0.93) and *employment_type* (F1 = 0.98), indicating reliable extraction of explicit attributes.

In contrast, *Llama-SEA-LION-v3-8B-IT* showed more inconsistent performance, relatively strong on *company_name* (F1 = 0.95) but markedly lower on context-dependent fields such as *job_location* (F1 = 0.43) and *job_category* (F1 = 0.43). Across all models, fields characterized by high lexical and structural regularity namely *job_title*, *company_name*, and *job_url*; achieved uniformly high precision and recall, suggesting that explicit identifiers are readily captured by LLM-based extractors. Conversely, *job_category* and *job_location* showed moderate to low scores for some models, reflecting the semantic diversity and inconsistent categorization commonly found in real-world postings (e.g., "IT & Software" vs. "Software Engineering").

The most challenging field was *posted_date*, where all models recorded low F1-scores ranging from 0.27 to 0.36. This weakness can be attributed to the prevalence of non-standardized and relative temporal expressions in job listings (e.g., "3 hours ago," "yesterday," or varying date formats), which complicate normalization and hinder consistent structured extraction.

4.2 Skill Extraction Experiment

To validate the effectiveness of our skill extraction methodology, we used the SkillSpan Dataset, a benchmark annotated corpus specifically designed for skill extraction evaluation. SkillSpan contains human-annotated skill spans across a variety of job descriptions. We compared the output of our LLM-based method against the SkillSpan annotations using standard evaluation metrics, including precision, recall, F1-score, and exact-match accuracy. Because skill spans often consist of multi-word phrases (e.g. "integration testing", "cloud native applications"), we report both a strict

Table 3. Overall LLM Extraction Performance with Precision(P), Recall(R), and F1-Score(F1). Type JC - is the information from Job Card, and JP- is the information extracted from the job details page.

LLM Model→		**mistral:7b**			**Llama-SEA-LION-v3-8B-IT**			**qwen3:8b**		
Field	*Type*	P	R	F1	P	R	F1	P	R	F1
job_title	*JC*	1	0.87	0.93	1	0.82	0.90	1	0.92	0.96
company_name	*JC*	1	0.83	0.91	0.91	0.91	0.95	1.0	0.90	0.95
job_location	JP	0.87	0.86	0.86	0.47	0.41	0.43	0.95	0.95	0.95
job_category	JP	0.76	0.75	0.75	0.44	0.42	0.43	0.69	0.68	0.69
posted_date	JC	0.41	0.33	0.36	0.28	0.24	0.27	0.38	0.33	0.34
job_url	JC	1.0	0.93	0.96	1.0	0.99	0.99	1.0	1.0	1.0
employment_type	JP	0.99	0.98	0.98	0.88	0.85	0.87	0.99	0.98	0.99

Table 4. Span-Level Skill Extraction Performance.

model	**qwen3:8b**			**mistral:7b**			**Llama-SEA-LION-v3-8B-IT**		
	P	R	F1	P	R	F1	P	R	F1
EXACT-MATCH	0.289	0.279	0.284	0.104	0.205	0.138	0.093	0.227	0.132
PARTIAL-MATCH	0.693	0.671	0.682	0.501	0.988	0.665	0.4	0.975	0.567

exact-match score and a partial-match score. Under the partial-match criterion, we count a predicted span as correct if its normalized text overlaps in any way with a gold span (i.e. one string is a substring of the other). This lets us credit the model even when it splits or trims a long phrase and provide a comprehensive evaluation, demonstrating both the accuracy and practical applicability of our skill extraction framework.

Evaluation metrics: We compared the spans produced by our LLM pipeline against these annotations using standard information-retrieval measures. We retained 603 job-description records that contain at least one skill annotation. Each description was processed by the LLM with zero-shot prompts that instruct it to return lists of skills. Predicted spans were token-normalized (lowercased, punctuation removed) and aligned to the gold spans using 2 mentioned matric: exact-match, where a predicted span must match the gold span character-for-character, and partial-match, where a predicted span is considered correct if either string is a substring of the other, allowing minor boundary mismatches such as “integration testing” vs. “testing integration”.

As **Table 4** shows, all three models performed poorly on the strict exact-match evaluation, with F1-scores below 0.30. This indicates that a zero-shot LLM approach struggles to precisely identify the exact boundaries of skill spans as defined by the gold-standard dataset. However, the results for the partial-match evaluation reveal a significantly stronger performance, demonstrating the models' effectiveness at a conceptual level. The *qwen3:8b* model achieved the highest F1-score in both exact-match (0.284) and partial-match (0.682), which indicates it has the most balanced precision and recall across both evaluation criteria. Notably, the *mistral:7b* model resulted in moderate performance under partial-match, with a precision of 0.501, but extremely high recall (0.988), meaning it tends to extract more candidates (many correct, some incorrect). The model *Llama-SEA-LION-v3-8B-IT*, on the other hand, produced the lowest overall results; although its recall was high (0.975), its precision (0.400) was substantially lower, indicating over-generation and noisy extraction.

Overall, the experimental results suggest that while zero-shot LLMs can recognize relevant skill concepts with reasonable semantic accuracy, they remain limited in producing exact span boundaries that align with human annotations. Among the evaluated models, *qwen3:8b* demonstrates the most reliable performance across both strict and partially match criteria, highlighting its effectiveness for skill identification in real-world job descriptions. These findings underscore the potential of LLM-based methods for scalable skill extraction, while also indicating the need for fine-tuning or hybrid approaches to improve exact boundary alignment.

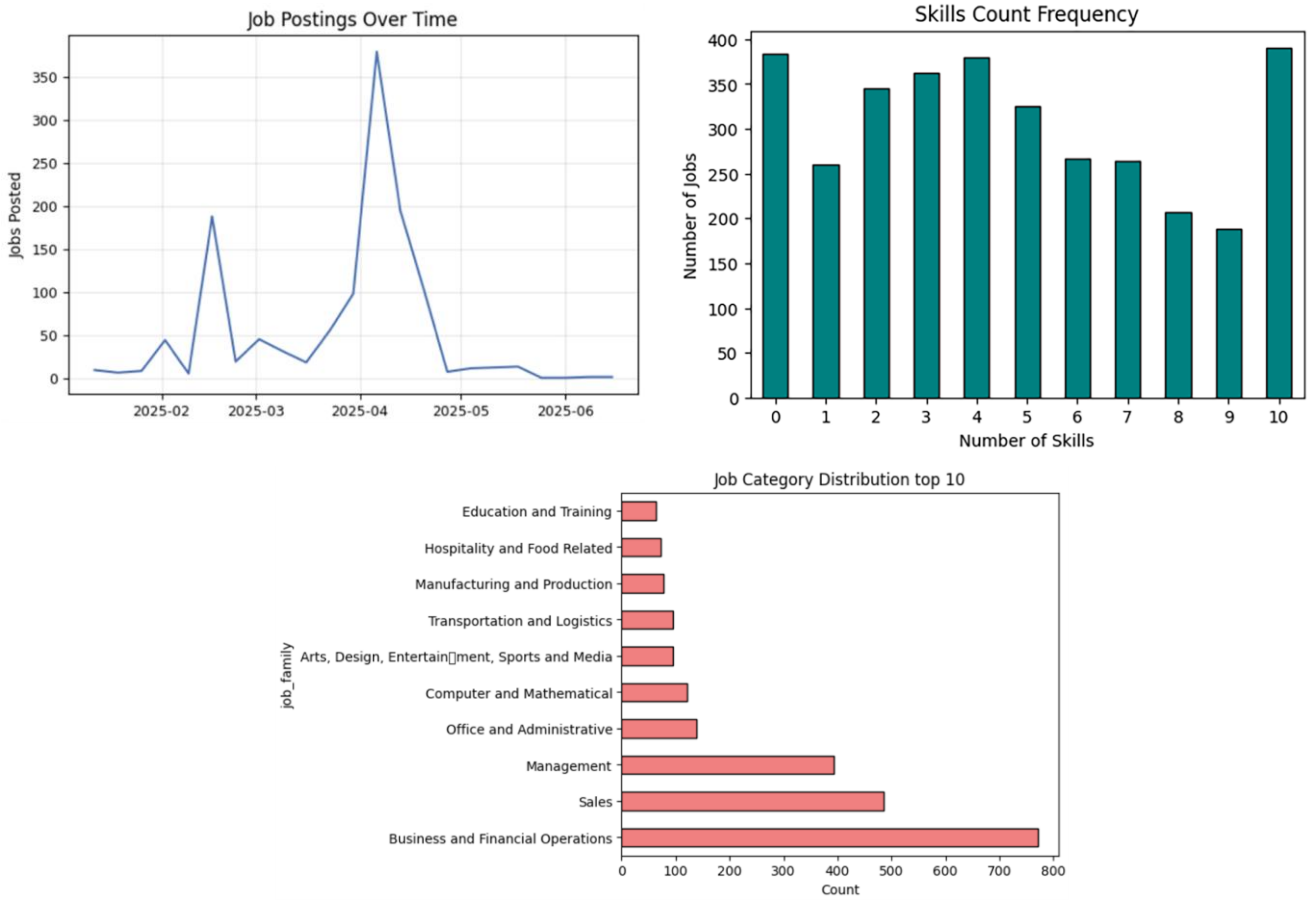

Fig. 5. Summarizes descriptive statistics of collected jobs in Cambodia

4.3 Analysis of the Cambodia's job dataset

This section presents a detailed analysis of the skill distributions and trends extracted from job postings data collected from online platforms in Cambodia. The primary objective of this analysis is to identify both established and emerging skills, prevalent competencies within specific job categories, and derive insights useful for curriculum planning and workforce strategies. After completing data collection, preprocessing, job classification, skill extraction, and standardized skill matching, the data were stored in a structured database, facilitating flexible queries and trend analyses across various job categories and sectors.

Fig. 5 summarizes our collected job dataset in Cambodia using our scraping pipeline. The dataset, gathered from early to mid-2025, includes approximately 3,500 job postings from online platforms. Posting volume fluctuated over this period, with a pronounced peak in April 2025 of over 350 postings, reflecting intensive data collection rather than an actual labor market trend. Due to the limited and methodologically biased timeframe, this dataset is unsuitable for time-series analysis. The dataset is heavily concentrated in the "Business and Financial Operations" and "Sales" categories, which together account for over 30% of the postings. Other prominent categories include "Management" and "Computer and Mathematical," indicating high demand for corporate and technological roles. Skill distribution shows three notable high-frequency points at 0, 4, and 10 skills per posting, with the highest frequency at 10 skills, suggesting that many postings list extensive competencies. Postings with 0 extracted skills likely correspond to advertisements that omit explicit skill lists, describe qualifications in vague terms (e.g., "strong communication skills," "relevant experience required"), or are written in multiple languages, which can limit the model's ability to recognize skill expressions. This pattern highlights both the inconsistency in how employers present skill information and the limitations of automated extraction for implicit or multilingual content.

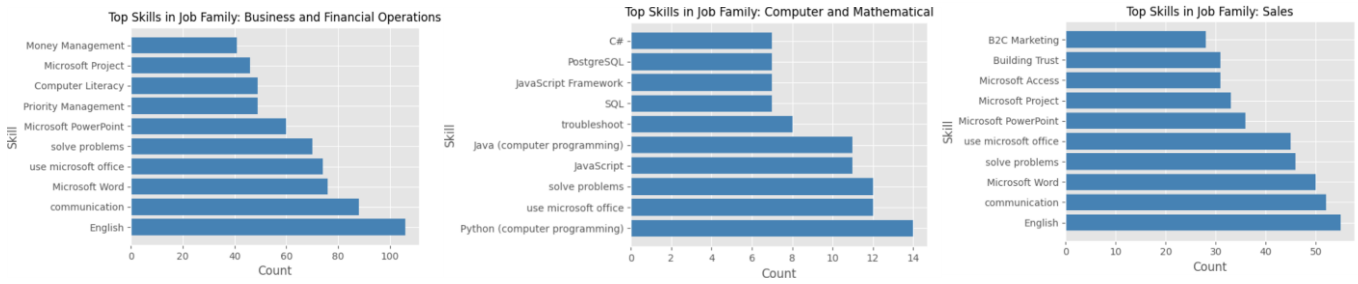

Fig. 6. Top skills among top job category

From **Fig 6**, We presents an analysis of selected high-population job categories in the dataset: Business and Financial Operations, Sales, and Computer and Mathematical, which reveals distinct skill requirements. The Business and Financial Operations, and Sales categories are dominated by soft skills, with Communication being the most in-demand skill in both. These roles also require practical skills such as Microsoft Word and proficiency in English. In contrast, the Computer and Mathematical category is defined by a strong demand for specific technical skills, including Python and JavaScript. This analysis clearly shows how skill demand varies significantly across different sectors, while highlighting the universal importance of foundational skills like communication, English and problem solving.

Our final analysis examines the timeline trends of skills in Cambodia's job market. **Fig. 7** illustrates the preliminary weekly trends of the top five most frequently mentioned skills. The graph shows notable fluctuations in skill demand over the analyzed period, with a pronounced spike in mentions of all top skills, such as English, Communication, Microsoft Word, Problem Solving, and Microsoft Office, around weeks 15 and 16. This temporary surge likely reflects the peak of active, non-uniform data collection rather than actual labor market changes. Although the limited timeframe and inconsistent collection frequency prevent a robust time-series analysis, this visualization provides initial insights into the temporal volatility of skill demand captured through online job postings. It also highlights the potential of this methodology to track real-time shifts in the labor market given a more sustained and consistent data collection strategy.

5. CONCLUSIONS

This study developed an AI-based pipeline for analyzing labor market trends in Cambodia, integrating web scraping, LLM-driven skill extraction, and hybrid taxonomy alignment using ESCO and the World Economic Forum's Future of Jobs taxonomy. The approach transforms unstructured job postings into structured, taxonomy-aligned datasets, providing a foundation for large-scale labor market intelligence in developing economies. Analysis of the real-world dataset highlights both cross-cutting competencies, such as communication and English, and sector-specific skills, including Python and SQL for technical roles or customer service for administrative roles. These findings demonstrate the potential of automated methods to deliver timely, context-relevant insights for curriculum development, workforce training, and policy design.

At the same time, several limitations should be acknowledged: The dataset covers only a five-month period and primarily reflects the formal sector; the pipeline currently processes only English-language postings, omitting Khmer and code-switched texts; LLM-based extraction remains vulnerable to hallucination and bias; and occupational classification lacked expert-annotated validation. Nevertheless, these constraints provide a clear roadmap for future work. Expanding data coverage across longer periods and additional portals, incorporating multilingual processing adapted to Cambodia's labor market, and introducing human-in-the-loop validation alongside domain-specific model adaptation will enhance robustness and representativeness. Despite these limitations, this study offers a novel foundation for labor market intelligence in Cambodia and demonstrates the feasibility and promise of AI-driven approaches in similar developing contexts.

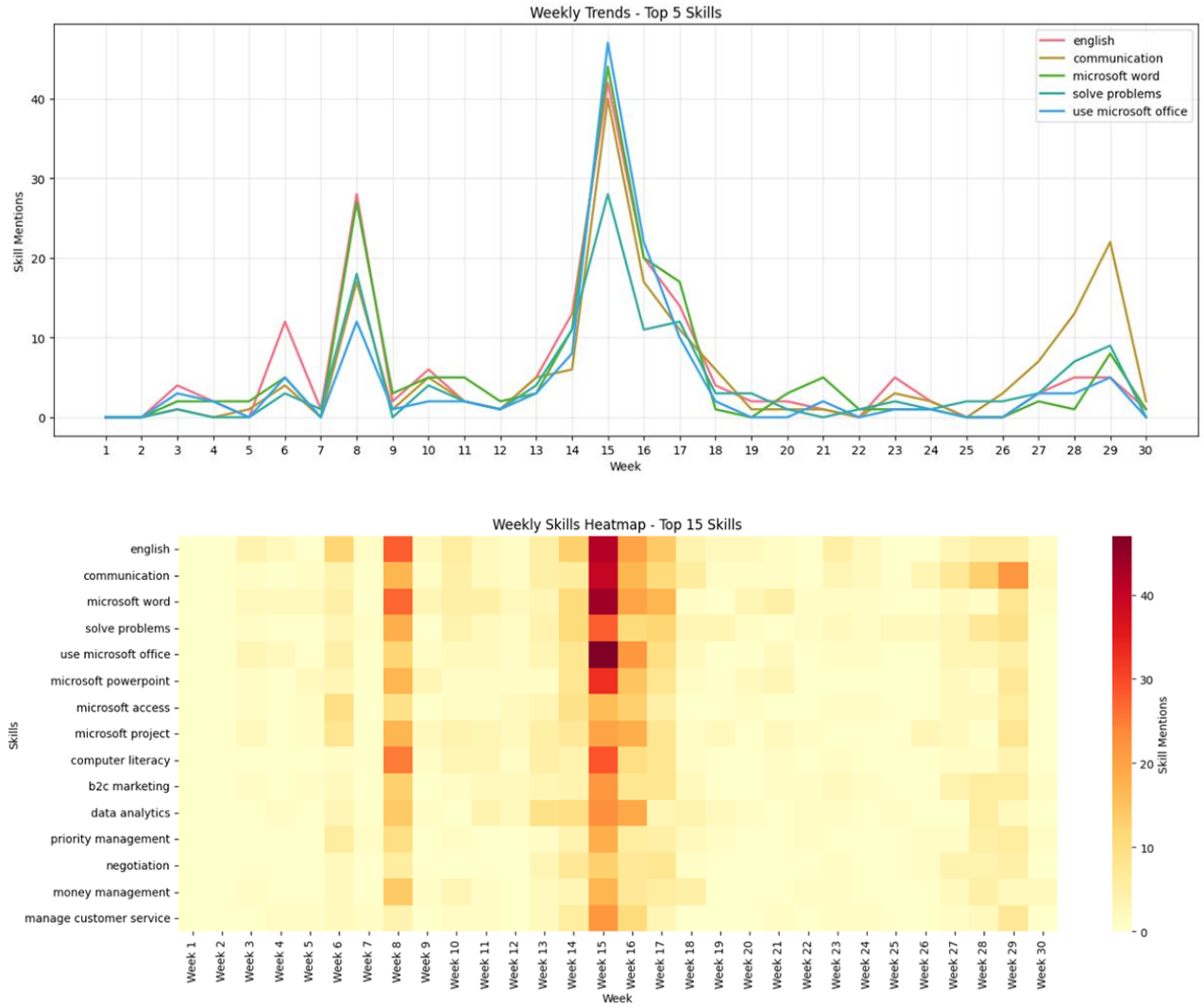

Fig.7. Time series trend and heatmap of skills

REFERENCES

[1] EuroCham Cambodia, & Swisscontact. (2024, June 28). EuroCham Skills Gap Assessment Report 2024. Phnom Penh, Cambodia: EuroCham Cambodia & Swisscontact (Skills Development Programme, SDC). https://www.eurocham-cambodia.org/uploads/782f9-final-report-skill-gap-assessment-report-digital-3.pdf.

[2] LinkedIn. (2020). The LinkedIn Economic Graph: Mapping opportunities through data. LinkedIn Economic Graph Research. https://arxiv.org/abs/2010.13981.

[3] Indeed Hiring Lab. (2024, December 10). 2025 US jobs and hiring trends report. Indeed. https://www.hiringlab.org/2024/12/10/indeed-2025-us-jobs-and-hiring-trends-report.

[4] Glassdoor. (2023). Job market report: Historical trends and insights. Glassdoor Economic Research. https://www.glassdoor.com/blog/job-market-report-historical.

[5] Lightcast. (2023). Labour Insight data caveats. Lightcast Knowledge Base. https://kb.lightcast.io/en/articles/8735719-labour-insight-data-caveats.

[6] National Employment Agency (N.E.A). (n.d). https://www.nea.gov.kh/index.do

[7] LaboNFC, University of Quebec at Chicoutimi, 555 Boulevard de l'Université, Saguenay (QC), Canada, Lotfi, C., Srinivasan, S., Ertz, M., & Latrous, I. (2021). Web Scraping Techniques and Applications: A Literature Review. In Jaypee Institute of Information Technology, Noida, India, R. Pal, P. Kumar Shukla, & Babu Banarasi Das University, Lucknow, India, SCRS CONFERENCE PROCEEDINGS ON INTELLIGENT SYSTEMS (pp. 381–394). Soft Computing Research Society. https://doi.org/10.52458/978-93-91842-08-6-38.

[8] Pichiyan, V., Muthulingam, S., G, S., Nalajala, S., Ch, A., & Das, M. N. (2023). Web Scraping using Natural Language Processing: Exploiting Unstructured Text for Data Extraction and Analysis. Procedia Computer Science, 230, 193–202. https://doi.org/10.1016/j.procs.2023.12.074.

[9] Bastian, M., Hayes, M., Vaughan, W., Shah, S., Skomoroch, P., Kim, H., Uryasev, S., & Lloyd, C. (2014). LinkedIn skills: Large-scale topic extraction and inference. Proceedings of the 8th ACM Conference on Recommender Systems, 1–8. https://doi.org/10.1145/2645710.2645729.

[10] Zhang, M., Jensen, K. N., Sonniks, S. D., & Plank, B. (2022). SkillSpan: Hard and Soft Skill Extraction from English Job Postings (No. arXiv:2204.12811). arXiv. https://doi.org/10.48550/arXiv.2204.12811.

[11] Yamashita, M., Shen, J. T., Tran, T., Ekhtiari, H., & Lee, D. (2023). JAMES: Normalizing Job Titles with Multi-Aspect Graph Embeddings and Reasoning (No. arXiv:2202.10739). arXiv. https://doi.org/10.48550/arXiv.2202.10739.

[12] Magron, A., Dai, A., Zhang, M., Montariol, S., & Bosselut, A. (2024). JOBSKAPE: A Framework for Generating Synthetic Job Postings to Enhance Skill Matching (No. arXiv:2402.03242). arXiv. https://doi.org/10.48550/arXiv.2402.03242.

[13] Senger, E., & Zhang, M. (n.d.). Deep Learning-based Computational Job Market Analysis: A Survey on Skill Extraction and Classification from Job Postings.

[14] Chen, S., Zhang, Q., Dong, J., Hua, W., Li, Q., & Huang, X. (2024). Entity Alignment with Noisy Annotations from Large Language Models (No. arXiv:2405.16806). arXiv. https://doi.org/10.48550/arXiv.2405.16806.

[15] International Labour Organization. (n.d.). Classification of occupation. ILOSTAT. https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/.

[16] National Center for O*NET Development. (2021). O*NET OnLine. U.S. Department of Labor, Employment & Training Administration. https://www.onetonline.org.

[17] Handel, M. J. (2016). The O*NET content model: Strengths and limitations. Journal for Labour Market Research, 49(2), 157–176. https://doi.org/10.1007/s12651-016-0199-8.

[18] European Commission. (2022). ESCO: European Skills, Competences, Qualifications and Occupations. https://esco.ec.europa.eu.

[19] CDRI. (2021). Exploring insights into vocational skills development and industrial transformation in Cambodia (Working Paper No. 131). Cambodia Development Resource Institute. https://www.cdri.org.kh/storage/pdf/WP131%20Exploring_Insights_1637129221.pdf.

[20] World Economic Forum. (2025, January 7). The Future of Jobs Report 2025 (ISBN 978-2-940631-90-2). Geneva: World Economic Forum. Retrieved from https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf.

[21] Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, volume 10, pages 707–710. Soviet Union.

[22] Dos Santos, J. B., Heuser, C. A., Moreira, V. P., & Wives, L. K. (2011). Automatic threshold estimation for data matching applications. Information Sciences, 181(13), 2685–2699. https://doi.org/10.1016/j.ins.2010.05.029.

[23] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation (No. arXiv:2402.03216). arXiv. https://doi.org/10.48550/arXiv.2402.03216.

[24] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7B (No. arXiv:2310.06825). arXiv. https://doi.org/10.48550/arXiv.2310.06825.

[25] Ng, R., Nguyen, T. N., Huang, Y., Tai, N. C., Leong, W. Y., Leong, W. Q., Yong, X., Ngui, J. G., Susanto, Y., Cheng, N., Rengarajan, H., Limkonchotiwat, P., Hulagadri, A. V., Teng, K. W., Tong, Y. Y., Siow, B., Teo, W. Y., Lau, W., Tan, C. M., … Teo, L. (2025). SEA-LION: Southeast Asian Languages in One

Network (No. arXiv:2504.05747). arXiv. https://doi.org/10.48550/arXiv.2504.05747.

[26] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., … Qiu, Z. (2025). Qwen3 Technical Report (No. arXiv:2505.09388). arXiv. https://doi.org/10.48550/arXiv.2505.09388.