Jurnal Teknik Industri, Vol. 28, No. 1, June 2026
ISSN 1411-2485 print / ISSN 2087-7439 online
DOI: https://doi.org/10.9744/jti.28.1.47-58

Big Data Analysis of Skill Requirements in the Indonesian Manufacturing Sector: A Semantic Approach Using Large Language Models

Rubanto Sidi Hambaly*, Muhammad Akbar, Ni Luh Saddhwi Saraswati Adnyani
Faculty of Industrial Technology, Bandung Institute of Technology
Jl. Ganesha No. 10, Bandung 40132, Indonesia
Email: rubanto.sidi@gmail.com*, muhammad@itb.ac.id, saddhwi@itb.ac.id
*Corresponding author

Abstract: The rapid acceleration of Industry 4.0 has fundamentally reshaped industrial competency demands, resulting in the "skill mismatch" phenomenon and contributing to structural unemployment in Indonesia. Effective labor market analysis is required, but traditional analyses often rely on rigid, retrospective survey methodologies that fail to capture these fast-paced dynamics in real time. This study addresses this gap by introducing a novel data-driven pipeline that validates 2,688 web-scraped job advertisements against official national manufacturing registries: Statistics Indonesia (BPS) and the Mandatory Labor Report (WLKP). This registry-based validation ensures data integrity by filtering out 51.7% of unverified postings, guaranteeing that the analysis is derived exclusively from legitimate firms within the verified manufacturing sector. A semantic approach using the Gemini-based Large Language Model (LLM) was implemented to extract, normalize unstructured data into the ESCO taxonomy, and categorize it. Unlike traditional NLP metrics that often fail to maintain functional relevance, the LLM-based approach successfully preserves professional context. While automated exact matching with the rigid ESCO framework yielded low accuracy (24.3% for titles; 9.8% for skills), expert validation confirmed high semantic accuracy of 81.5% and 85%, respectively. Strategic insights reveal a dual-track workforce structure: vocational graduates require technical dexterity for operational roles, while higher education graduates are sought for strategic oversight. Analysis reveals a dominant focus on operational excellence, with specialized digital demand varying by sector, such as CATIA for high-precision engineering in the automotive sector and Optitex for 3D-digital workflows in the apparel industry. This framework serves as an industrial demand blueprint for curriculum-industry alignment, while offering a synthesized scientific interpretation of the underlying labor market patterns.

Keywords: Data-driven decision-making, skill mismatch, Large Language Models, ESCO taxonomy, manufacturing industry.

Introduction

Industry 4.0 reshapes manufacturing through automation [1], intensifying Indonesia's "skill mismatch" as education outputs struggle to meet specialized demands [2]. This misalignment drives structural unemployment in hubs like West Java, where high vacancies paradoxically coexist with graduate unemployment [2], further compounded by the need for integrated "green" digital competencies [3].

Traditional surveys lack the real-time granularity needed to track these fast-paced changes [4], while raw web-scraped data often contains unverified "noise" [5], [6], [7]. This study addresses these gaps by validating job advertisements against official BPS and WLKP registries, utilizing text-mining to optimize industrial decision-making [8]

Furthermore, standard keyword-matching fails to capture the semantic nuances of required competencies [9]. Recent LLM advancements enable dynamic taxonomy generation and superior contextual accuracy [10]. By employing the Gemini architecture, this study bridges localized job terminology with the standardized ESCO taxonomy, overcoming the limitations of traditional models [9], [11].

The primary objective of this research is to develop an automated computational framework for analyzing industrial competency demands within the manufacturing sector by leveraging Large Language Models (LLMs) to process unstructured job advertisements. This study specifically aims to automate data acquisition through web scraping and to implement a validation pipeline against official national registries, namely Statistics

Submitted: Dec 31, 2025; Revised: Mar 09, 2026; Accepted: Apr 06, 2026

Indonesia (BPS) and the Mandatory Labor Report (WLKP), to ensure data integrity. Furthermore, the framework focuses on the automated standardization of job titles and the extraction of technical and soft skills, mapping them into three distinct domains: Hard Skills, Soft Skills, and Tools/Software [12]. The effectiveness of this methodology is measured by a primary success criterion: achieving semantic accuracy>80% as determined by expert validation, which prioritizes functional and professional relevance over rigid string-matching metrics.

Methods

As illustrated in Figure 1, this study utilizes a six-phase computational framework to extract and synthesize manufacturing labor intelligence. The methodology builds on established text-mining approaches used to evaluate industry performance and workforce dynamics [6], [13], [14].

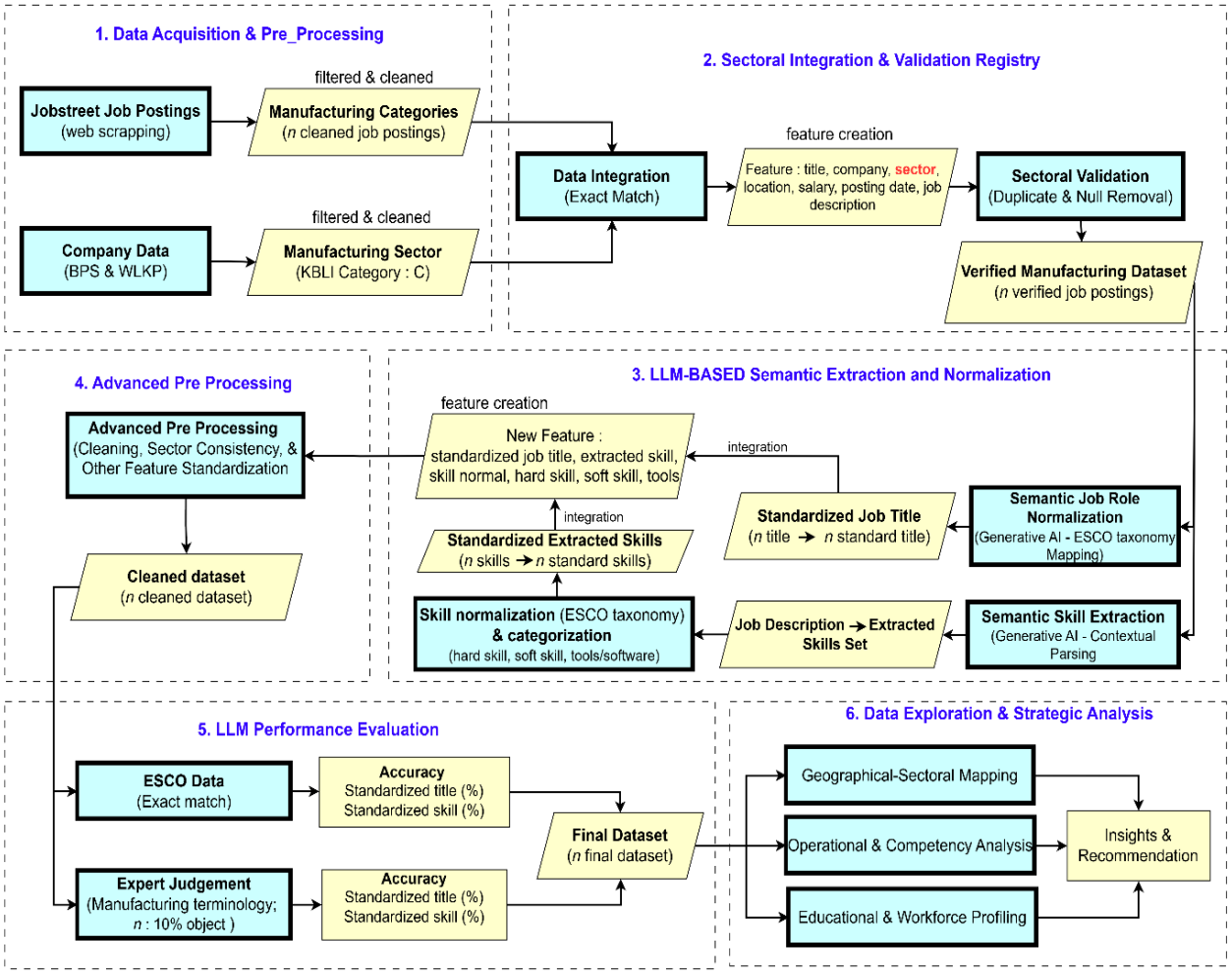

Figure 1. Proposed research pipeline

Data Acquisition and Initial Pre-Processing

Data was collected from JobStreet on December 1, 2025, using a targeted web-scraping approach. To ensure source precision, practitioners pre-filtered "Manufacturing, Transportation & Logistics," and "Engineering" categories to generate entry-point URLs. A Python script (Selenium/BeautifulSoup) then extracted job descriptions and metadata [6], which were integrated with the official BPS and WLKP registries, specifically isolating firms within the Manufacturing Industry Category C (KBLI).

The raw dataset underwent multi-stage pre-processing using Pandas and Regex to eliminate HTML noise and metadata. Textual normalization was achieved through case-folding to ensure consistency [13], while feature selection isolated company names and sectors from secondary registries. These steps establish a structured foundation, mitigating noise and misinformation risks during subsequent LLM-based skill extraction [5], [9].

Dataset Limitation

While certain constraints exist, several justifications ensure the data's representativeness. First, although the scraping was conducted on a single day (December 1, 2025), the dataset successfully captured active postings from the preceding 60 days, providing a year-end snapshot of the 2025 labor market. This methodology is designed as a repeatable framework that allows for future longitudinal updates to mitigate potential seasonal bias. Second, JobStreet was selected due to its high volume of advertisements and coverage of operational, technical, and vocational roles, which are often underrepresented on managerial-focused platforms like LinkedIn.

The representativeness of this source is further evidenced by a high correlation with official BPS statistics in terms of sectoral distribution (e.g., Food Industry: 15% in dataset vs. 19% BPS; Textiles: 5% vs. 5%), confirming JobStreet as a statistically valid proxy for the industrial population. Compared to retrospective manual surveys like Sakernas [1], this methodology provides real-time granularity.

Sectoral Integration and Validation Registry

The second phase establishes ground truth by cross-referencing advertisements with the BPS (2025) and WLKP (until Oct 2025) registries. Using an exact-matching algorithm for company names, the process targets firms registered under KBLI Category C (Manufacturing), spanning 24 sub-sectors (codes 10–33) to ensure domain relevance. Verified profiles were subsequently merged with job attributes and refined to eliminate null values or redundancies. This systematic validation guarantees that the analysis is derived exclusively from verified industrial data rather than unverified digital noise [13].

LLM-based Semantic Extraction and Normalization

Unstructured job descriptions were processed using the Gemini Large Language Model (LLM) architecture. The model was prompted to map skill to the European Skills, Competences, Qualifications, and Occupations (ESCO) taxonomy. This phase consisted of (1) Job Title Standardization and (2) Skill Extraction, Normalization, and Categorization.

Job Title Standardization

Post-validation, job titles were standardized to remove noise—such as "Urgent Hiring" or department codes—that obscures fundamental role identities [5]. As illustrated in Figure 2, this utilized a Gemini 2.5-Pro LLM with zero-shot prompting, leveraging the model's pre-trained reasoning to map titles into recognized occupation labels without requiring prior examples [4].

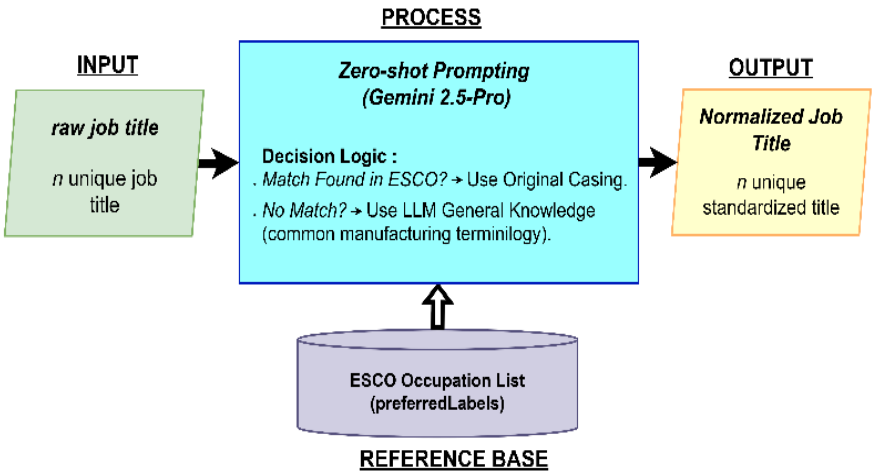

Figure 2. Job title standardization workflow

The technical implementation involved a structured instruction that prioritized the ESCO taxonomy (preferredLabel) to ensure semantic consistency [11]. The model was executed with the following prompt:

> "*Standardize the following job title into a recognized occupation label, ideally from the ESCO classification or a commonly understood, general manufacturing job title that accurately describes the role.*"

This architecture incorporates a hybrid matching logic: if the LLM output matched an ESCO label, it was mapped back to the official ESCO casing; otherwise, the LLM's descriptive output was accepted as a flexible fallback.

Skill Extraction, Normalization, and Categorization

Following title standardization, the methodology proceeded to extract and refine competencies from unstructured job descriptions. As illustrated in the multi-tiered pipeline in Figure 3, the Gemini LLM was utilized to transform raw text into a structured competency registry, emphasizing the retrieval of both explicit and implicit skills [9].

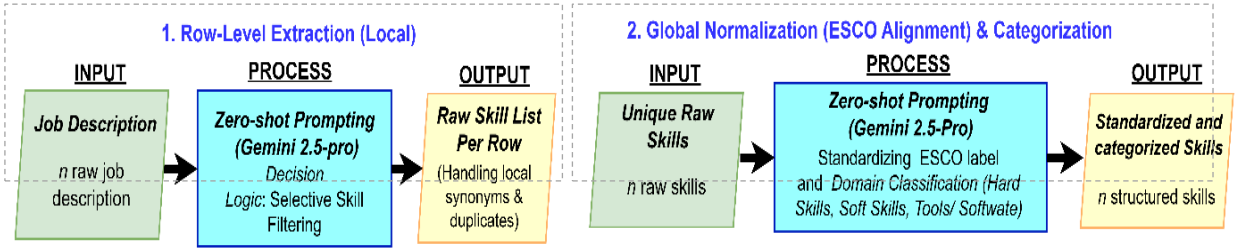

Figure 3. Multi-tiered skill extraction, normalization, and categorization workflow

The first phase, Row-level Extraction, utilized a zero-shot strategy to identify technical and soft skills using the following instruction:

> *"Extract a list of relevant technical and soft skills from the following job description. List the skills as a comma-separated string. If no skills are found, return 'No skills found'."*

To ensure stability, a retry mechanism with exponential backoff was implemented to handle API rate limits while filtering out administrative noise. The second phase involved Global Normalization and Categorization, where the LLM served as a semantic bridge to align disparate terminologies into standardized ESCO preferred labels. Then it categorized them into three domains (Hard Skills, Soft Skills, or Tools/Software) using this prompt:

> *"Normalize the following skill to an ESCO Preferred Label (if applicable, otherwise use a common, concise term). Prioritize merging synonyms and closely related terms into a single, most representative terminology. Categorize the normalized skill as 'Hard Skill', 'Soft Skill', or 'Tools/Software'."*

This approach ensures semantic consistency [11] and allows for a multidimensional analysis of manufacturing requirements. Finally, a data scrubbing layer eliminated linguistic artifacts and redundant whitespace, resulting in a verified foundation for analyzing industry-workforce alignment [13].

LLM Performance Evaluation

To validate the semantic pipeline, a dual-layered evaluation framework was applied to a representative sample of 120 records, accounting for 10% of the validated dataset. The first layer involved a computational similarity assessment—including Cosine, Jaccard, Jaro-Winkler, Levenshtein, and exact string matching—against the ESCO taxonomy. These quantitative benchmarks serve as a baseline for evaluating the model's adherence to rigorous international terminological standards and its semantic reasoning capabilities.

Simultaneously, a human-in-the-loop (HITL) validation was conducted on the same sample to assess practical accuracy. This qualitative oversight by domain experts ensures that the analysis remains industry-relevant, especially when localized manufacturing terminologies do not align syntactically with the ESCO framework. By comparing these multi-metric automated scores with expert judgment, the study illustrates the "semantic gap" that traditional NLP methods often fail to bridge [4], [13].

Advanced Data Refinement and Feature Engineering

The final stage of data preparation focused on categorical consistency and feature engineering to transform raw metadata into granular analytical dimensions. Geographical data were normalized into standardized Province entities, while minimum educational requirements were consolidated into structured categories (e.g., Vocational and Undergraduate). Additionally, unstructured salary information was parsed into numerical metrics—including Minimum, Maximum, and Average Salary—to enable economic demand modeling.

A final global synchronization phase was performed to eliminate malformed strings and null values, resulting in a feature set ready for statistical frequency analysis. This systematic refinement ensures that the dataset is mathematically consistent and optimized for identifying complex labor market patterns within the manufacturing sector [6], [13].

Data Exploration and Analysis

Strategic insights were extracted from the validated dataset using a multidimensional framework focusing on three analytical pillars: (1) Geographical-Sectoral Mapping to identify industrial clusters across Indonesian provinces; (2) Operational and Competency Demand analysis to examine the distribution of standardized roles, skills, and tools; and (3) Educational Alignment to correlate academic requirements with industry needs.

The analysis utilized Python's data science ecosystem (Pandas and Seaborn) to identify co-occurrence patterns between job roles and competency requirements. This approach establishes a blueprint for industrial demand, enabling a synthesized interpretation of the underlying labor market landscape [9], [13].

Results and Discussions

Data Acquisition and Pre-processing Results

The automated pipeline retrieved 2,688 raw job postings, capturing essential attributes including job titles, company names, salaries, locations, and descriptions. Utilizing online portals proved effective for capturing dynamic competency requirements that traditional, retrospective surveys often fail to represent on time [15].

Technical cleaning—primarily removing duplicates and handling missing values—refined the dataset to 2,456 records, an 8.6% reduction. This decrease highlights the presence of redundant or incomplete information typical of unstructured online advertisements [6]. Simultaneously, the secondary dataset processing managed a substantial volume of verified industrial records: 32,661 from the BPS directory and 525,843 from the WLKP registry. This refined pool of 2,456 records provides a robust empirical foundation for identifying technical skill demands in the Indonesian manufacturing sector.

Sectoral Validation and Registry Integration Results

The validation process yielded 1,185 verified records (48.3%), representing a 51.7% filtration rate. As illustrated in Figure 4, this was achieved through a two-stage sequential validation where Stage 1 cross-referenced records against the BPS registry, and Stage 2 validated the remaining unmatched data against the WLKP database. This registry-based approach isolates firms strictly (exact match) within KBLI Category C (Manufacturing) and ensures high data integrity by cross-referencing against official national records.

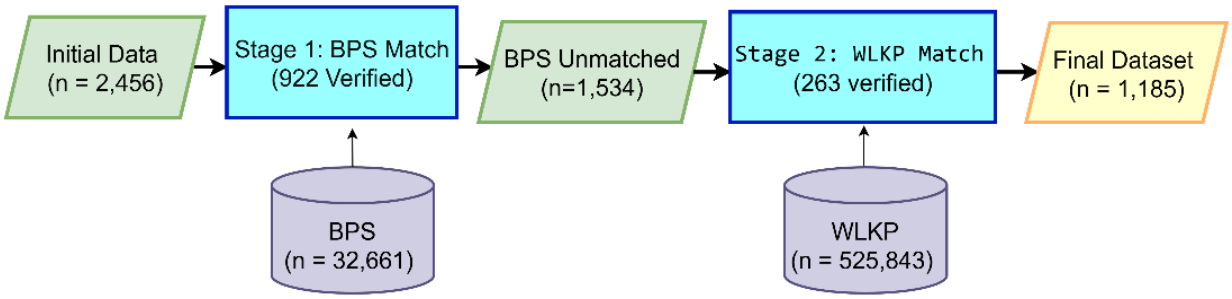

Figure 4. Data integration and sectoral validation framework

This high exclusion rate is a deliberate "Data Purity" outcome that filters out recruitment agencies and unverified entities that mislabeled themselves on the portal. Although a one-month temporal gap exists between the scraping (December 2025) and registries (October 2025), it is considered statistically insignificant as formal industrial licenses for legal entities do not fluctuate drastically within a 30-day window. By prioritizing exact matching over fuzzy logic, this study ensures that the final n=1,185 dataset provides a verified, integrity-based foundation for skill analysis.

Job Title Standardization Results

The LLM-based pipeline effectively consolidated the high terminological variance found in raw job postings. As summarized in Table 1, the diversity of job titles was reduced from 926 unique raw titles to 482 standardized labels, representing a 47.9% reduction. This consolidation is vital to prevent data fragmentation and ensure that identical roles with different naming conventions are grouped accurately for aggregate analysis [9].

Table 1. Metrics of job title standardization using Gemini 2.5-Pro

Metric	Value
Verified dataset size (n)	1,185
Unique raw job titles	926
Unique standardized job titles	482
Diversity reduction percentage	47.9%

The qualitative samples in Table 2 demonstrate the system's ability to map diverse inputs to standardized labels. By mapping specific terms like "Operator Milling" to the ESCO-aligned "Machine Operator," the system achieved high semantic alignment. This process resulted in an analytically robust dataset that accurately reflects the occupational structure of the Indonesian manufacturing sector [11], [13].

Table 2. Samples of raw job title normalization results

Raw Job Title (Input)	Standardized Job Title (Output)	Logic / Reason
Production supervisor (Dumai-Based)	production supervisor	Removed noise & metadata
Junior operator slitting	slitting operator	Removed seniority levels
Operator milling	machine operator	Mapped to ESCO label
Warehouse & plant logistics administrator	warehouse administrator	Consolidated synonyms

Skill Normalization and Domain Classification Results

The semantic pipeline successfully transformed unstructured job descriptions into a standardized competency registry. As summarized in Table 3, the normalization process reduced terminological variance from 4,887 raw skills to 3,321 unique normalized skills, resulting in a 32.04% reduction in diversity. This consolidation is critical for preventing data fragmentation, ensuring that identical competencies—previously obscured by minor linguistic differences—are correctly aggregated for frequency analysis [9].

Table 3. Metrics of skill normalization using Gemini 2.5-Pro

Metric	Value
Total raw skills extracted	4,887
Unique normalized skills	3,321
Diversity reduction	32.04%

Table 4 provides illustrative examples of the transformation from raw Indonesian job descriptions to standardized labels. This mapping process involves aligning localized terms, such as "lathe machine operation," into broader categories, such as the ESCO-aligned "machine operation." These entities are subsequently classified into three domains—Hard Skills, Soft Skills, and Tools—to provide a structured representation of the competency requirements identified within the manufacturing sector [9].

Table 4. Samples of skill normalization and categorization results

Raw skill (from text)	Normalized skill	Category
Excellent communication skills	communication	Soft skill
Lathe machine operation	machine operation	Hard skill
Proficiency in AutoCAD 2D/3D	autocad	Tools/software

This structured registry establishes a high-fidelity foundation for identifying core industrial demand blueprints. By resolving linguistic noise, the resulting insights accurately reflect the authentic requirements of the Indonesian industrial landscape [2].

LLM Validation Results and Comparative Analysis

Performance evaluation revealed a significant disparity between rigid technical accuracy and functional semantic validity. A verified audit of the Exact Taxonomy Matching yielded rates of 24.3% for job titles and 9.8% for skill entities. While higher than initial estimates, these metrics remain low due to the linguistic rigidity of the ESCO database, which struggles with bilingual (Indonesian-English) descriptions and localized manufacturing contexts.

The Semantic Gap in Similarity Metrics

Table 5 presents the failure patterns of traditional similarity metrics when processing manufacturing job titles. As shown, these mathematical approaches often yield high similarity scores for functionally incorrect labels. For example, with 'Machine Operator' as the input string, Cosine Similarity (0.87) fails to capture the hierarchical gap between 'Machine Operator' and 'Supervisor'. At the same time, Jaro-Winkler (0.84) produces literal errors by outputting 'Candy Machine Operator'.

Table 5. Comparative analysis of similarity metrics

Metric	Input string	Output label	Similarity score	Error / Logic analysis
Cosine similarity	Machine operator	Machine operator supervisor	0.87	Hierarchical error: Fails to distinguish operator from management.
Jaro-Winkler& Levenshtein		Candy machine operator	0.84	Literal error: Misled by string similarity in niche industrial terms
Jaccard similarity		Hydrogenation machine operator	0.40	Specificity error: Fails to achieve functional generalization.

While traditional metrics like Cosine Similarity, Jaro-Winkler, or Jaccard Similarity are often misled by string overlap—incorrectly suggesting "Supervisor" or niche roles like "Candy Machine Operator"—the LLM-based approach (Gemini 2.5-Pro) maintains functional integrity by correctly retaining the "Machine Operator" label. This confirms that, in the Indonesian industrial landscape, a flexible-semantic approach is superior to rigid numerical metrics for bridging the "semantic gap" [16].

Expert Consensus and Final Accuracy

To ensure objectivity, a Human-in-the-Loop validation was conducted by two independent experts: a Senior Vocational Instructor (E_1) and a Manufacturing Industry Practitioner (E_2). Expert scores (s) ranging from 1 to 3 were converted into a percentage accuracy metric (Acc) using Equation (1):

$$Acc = \left(\frac{\bar{s} - S_{min}}{S_{max} - S_{min}}\right) x\ 100\% \quad (1)$$

where $\bar{S}$ represents the combined average score from $(\bar{s}_1)$ and $(\bar{s}_2)$, $S_{min} = 1$, and $S_{max} = 3$. The results, summarized in Table 6, indicate high consensus and functional accuracy.

Table 6. Expert validation results for title, skill standardization, and categorization

Evaluation Category	Expert 1 Avg ($\bar{s}_1$)	Expert 2 Avg ($\bar{s}_2$)	Combined Average ($\bar{s}$)	% Accuracy (*Acc*)
Job title standardization score (1-3)	2.74	2.53	2.63	81.5%
Skill extraction and categorization score (1-3)	2.83	2.57	2.70	85%

The results prove that the LLM functions as an effective semantic bridge, capable of interpreting nuances that traditional algorithms overlook. By aligning real-world requirements with expert-validated standards, this registry provides a foundation of fidelity for addressing skill analysis in Indonesia [13].

Advanced Refinement and Final Dataset Integrity

The implementation of the advanced refinement pipeline successfully finalized the dataset at 1,173 records. As summarized in Table 7, the process achieved 0% missing values for skill attributes and verified the integrity of the newly engineered features.

This registry provided a comprehensive overview of the Indonesian manufacturing landscape by integrating company profiles, official industrial classifications (2-digit KBLI), and a multi-layered competency structure encompassing hard skills, soft skills, and tools/ software. The standardized salary and education metrics established an empirical basis for correlating specific technical competencies with their economic value and required academic background.

Table 7. Final dataset integrity and feature schema post-refinement

Category	*Status*	*Engineered features/columns*
Identity & sector	Verified	company, province, location, sector_2_digits, kbli_2_digits
Economic metrics	Numerical	average_salary, min_salary, max_salary, salary_per_month (rp)
Education	Categorized	education_category, min_education
Competency registry	Standardized	hard_skill, soft_skill, tools_software, skills_norm_global, skills_norm
Raw source data	Refined	extracted_skills, job_description, posting_date
Final sample (n)	Consistent	1,173 records

Data Exploration and Analysis

Exploration of the validated dataset (n=1,173) reveals a significant concentration of industrial activity within the Java region, primarily in West Java, Banten, and Central Java. As illustrated in Figure 5, West Java functions as the nation's primary multi-sectoral hub. Banten exhibits a specialized cluster in chemical processing and the rubber/plastics industry, while Central Java remains a stronghold for the wearing apparel industry.

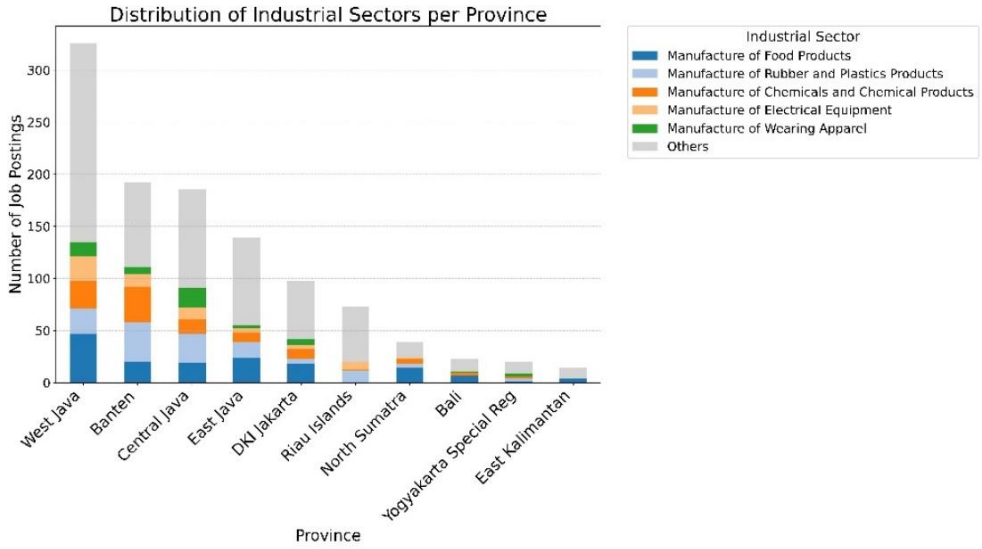

Figure 5. Distribution of the industrial sector per province

This clustering confirms Java's central role as the nation's economic engine, where high investment levels directly correlate with active labor market dynamics [2]. The observed regional specialization suggests that industrial policies and logistics infrastructure must be tailored to support specific sectoral strengths—such as wearing apparel innovation in Central Java and advanced chemical processing in Banten—to ensure efficient supply chain flow.

However, this geographical concentration fundamentally shapes a complex labor landscape. Despite the high volume of vacancies, West Java continues to face structural unemployment challenges due to a persistent mismatch between graduate profiles and industrial needs [2]. The distinct nature of each cluster necessitates a 'localized' approach to competency mapping; generic training programs are often ineffective unless they align with the specific technical demands of the local ecosystem [2]. Therefore, workforce readiness in these provinces

requires a strategic blend of adaptive and technical skills, facilitated by targeted interventions from regional institutions, such as training centers, to bridge the localized skill gap [2].

The distribution of the top 20 standardized job titles reveals a critical focus on production integrity and quality management. As illustrated in Figure 6, roles such as Production Supervisor, Quality Control Inspector, and Production Manager dominate the vacancy landscape. This concentration suggests that industrial demand in Indonesia is anchored in operational excellence and defect reduction, prioritizing the stability of production lines over purely digital roles [1], [12].

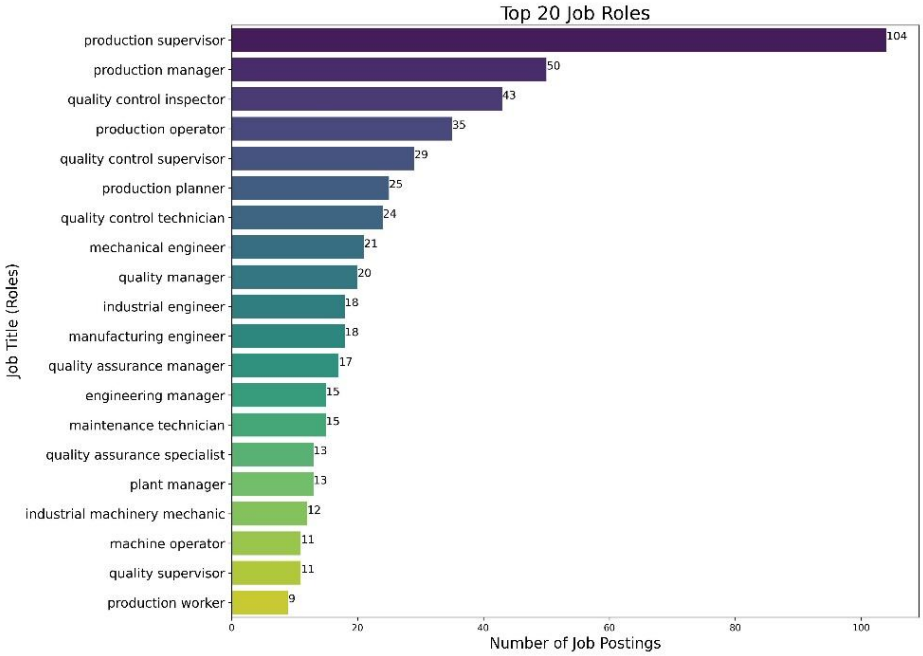

Figure 6. Top 20 job roles

Following the general distribution analysis, a granular sectoral assessment highlights how each industry's specific production paradigms dictate its technical requirements. As summarized in Table 8, while certain competencies are universal, the demand for hard skills and digital tools is highly contingent on sectoral operational logic [15].

Table 8. Top competency requirements by industrial sector (samples)

Sector	Top hard skills	Top soft skills	Top tools
Machinery & equipment	Mechanical engineering, troubleshooting, QC	Communication, attention to detail, problem-solving	AutoCAD, SAP, MS Excel
Food & beverage	GMP, quality control, inventory management	Leadership, communication, problem-solving	MS Excel, SAP, SolidWorks
Chemicals & plastics	Quality control, data analysis, production planning	Problem-solving, teamwork, leadership	SAP, SolidWorks, AutoCAD
Automotive & trailer	English language, ISO 9001, quality control	Problem-solving, communication, leadership	CATIA, CAD, AutoCAD
Textile & apparel	Business reporting, production management, QC	Leadership, communication, problem-solving	ERP, SAP, Optitex
Pharmaceuticals	GMP, ISO 9001, quality control	Problem-solving, communication, attention to detail	MS Excel, MS PowerPoint

The Machinery and Automotive sectors exhibit high demand for Mechanical Engineering and Troubleshooting, as equipment uptime and precision engineering are critical to maintaining competitive manufacturing cycles [17]. Conversely, the Food, Beverage, and Pharmaceutical sectors prioritize Good Manufacturing Practices (GMP) and ISO 9001, reflecting a mandatory focus on regulatory compliance and safety standards essential for consumer health and international certification [18], [19].

In labor-intensive sectors such as Textiles and Apparel, the emphasis shifts toward Production Management and Business Reporting. This is driven by the need to optimize high-volume operations and manage large workforces efficiently [12]. Across all 24 sub-sectors, interpersonal competencies—primarily Communication, Problem Solving, and Leadership—consistently rank at the top. This underscores the "human-centric" nature

of Indonesian manufacturing, where supervisors are expected to serve as a semantic bridge between complex technical floor operations and organizational strategic goals [20], thereby enhancing their overall employability in a competitive market [21].

Regarding technological adoption, Microsoft Excel remains a ubiquitous baseline for data recording. However, specialized tools are strictly sector-dependent. The high demand for SAP in the Food and Chemical industries is driven by the necessity for integrated supply chain transparency and resource planning [22]. Meanwhile, the dominance of AutoCAD, SolidWorks, and CATIA in the Machinery and Automotive sectors reflects a clear transition toward computer-aided engineering and the adoption of "Smart Factory" design principles [23]. This sectoral divergence confirms that a "one-size-fits-all" vocational curriculum is insufficient; rather, localized and sector-specific training modules are required to bridge the existing skill mismatch [2], [21].

The correlation in Figure 7 identifies a distinct, layered workforce structure within the manufacturing sector. A strong positive correlation exists between High School (SMK) and Diploma requirements, as both educational tracks frequently overlap in technical and operational roles, creating a shared "technical pool" for production-line tasks [20].

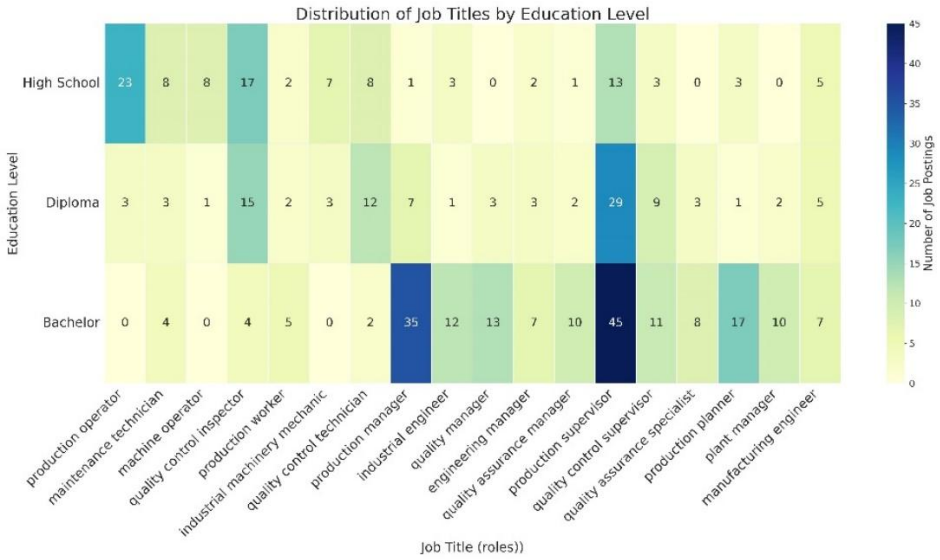

Figure 7. Distribution of job titles by education level

In contrast, Bachelor requirements display a clear demarcation, being exclusively mapped to strategic management and specialized engineering. This divergence occurs because the industry distinguishes between operational stability provided by vocational tracks and the strategic oversight required for Industry 4.0 implementation and digital transformation [20], [24]. This structural insight is vital for addressing the skill mismatch in regions like West Java; it suggests that while vocational curricula must focus on operational excellence, higher education must be synchronized with advanced technological leadership to meet actual industrial demand [25].

Conclusions

This study developed a data-driven pipeline to analyze industrial competency demands by validating 2,688 web-scraped advertisements against official BPS and WLKP registries. The results revealed that 51.7% of online postings did not align with verified manufacturing entities, underscoring the need for cross-referencing to prevent data bias. While automated exact matching with the ESCO taxonomy yielded low accuracy (24.3% for titles; 9.8% for skills) due to localized terminology, the Gemini-based LLM achieved high semantic accuracy of 81.5% for titles and 85% for skills, as confirmed by expert validation.

The resulting competency blueprint identifies a dual-track workforce structure: high school graduates (SMA/SMK) are prioritized for technical dexterity in frontline operations. In contrast, higher education graduates (Diploma/Bachelor's) are required for strategic oversight and system analysis. Furthermore, industrial competencies are increasingly shifting toward digital technologies, with requirements varying by sector, such as AutoCAD for machinery, GMP for food, and ISO 9001 for automotive. Universal soft skills, including Communication, Problem Solving, and Leadership, remain essential across all 24 sectors. Ultimately,

this high-fidelity registry provides a data-driven foundation for educational institutions to synchronize curricula with actual technological adoption, offering a strategic solution to mitigate structural unemployment in Indonesia.

References

[1] G. Li, C. Yuan, S. Kamarthi, M. Moghaddam, and X. Jin, "Data science skills and domain knowledge requirements in the manufacturing industry : A gap analysis," *J. Manuf. Syst.*, vol. 60, no. July, pp. 692–706, 2021, doi: https://doi.org/10.1016/j.jmsy.2021.07.007.

[2] S. R. Mubaroq, I. Gustiana, F. Alamsari, M. Artarina, and H. Nurohmah, "Proactive socio-technical system as an unemployment solution in West Java," *J. Phys. Conf. Ser.*, vol. 1402, no. 2, 2019, doi: https://doi.org/10.1088/1742-6596/1402/2/022072.

[3] T. Akyazi, P. del Val, A. Goti, and A. Oyarbide, "Identifying Future skill requirements of the job profiles for a sustainable European manufacturing Industry 4.0," *MDPI Recycl. J.*, 2022, doi: https://doi.org/10.3390/recycling7030032.

[4] K. C. Nguyen and A. Bosselut, "Rethinking skill extraction in the job market domain using large language models," *Assoc. Comput. Linguist*, no. Nlp4hr, pp. 27–42, 2024, doi: https://doi.org/10.18653/v1/2024.nlp4hr-1.3.

[5] P. Hoang, T. Mahoney, F. Javed, and M. McNair, "Large-scale occupational online recruitment," *AI Mag.*, pp. 5–14, 2018, doi: https://doi.org/10.1609%2Faimag.v39i1.2775.

[6] A. Kumar, K. Chauhan, and J. K. Grewal, "Web scraping job portals," *Adv. Commun. Syst.*, pp. 291–303, 2024, doi: https://doi.org/10.56155/978-81-955020-7-3-25.

[7] I. Khaouja, "A survey on skill identification from online job ads," *IEEE Access*, vol. 9, pp. 118134–118153, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3106120.

[8] F. Clemens, H. H. Özdemir, and G. Schuh, "Identification of text mining use cases in manufacturing companies," in *Conference On Production Systems And Logistics*, 2023. doi: https://doi.org/10.15488/15241.

[9] L. Malandri and F. Mercorio, "SkiLLMo: Normalized ESCO skill extraction through transformer models," *Assoc. Comput. Mach.*, no. March, 2025, doi: https://doi.org/10.1145/3672608.3707960.

[10] L. J. Gonzalez-Gomez *et al.*, "Dynamic taxonomy generation for future skills identification using a named entity recognition and relation extraction pipeline," *Front. Artif. Intell.*, vol. 8, 2025, doi: https://doi.org/10.3389/ 2025.1579998.

[11] D. Christos, K. Georgiou, E. Papaioannou, K. Petrakis, N. Mittas, and L. Angelis, "ESCOX : A tool for skill and occupation extraction using LLMs from unstructured text," *Softw. Impacts*, vol. 25, no. June, p. 100772, 2025, doi: https://doi.org/10.1016/j.simpa.2025.100772.

[12] K. Z. Akhilla, A. Sukmawati, and B. Sartono, "Identifying the types of future skills needed in the manufacturing industry: A systematic literature review," *J. Apl. Bisnis Dan Manaj.*, vol. 11, 2025, doi: https://doi.org/10.17358/jabm.11.3.1099.

[13] K. Djunaidi, D. T. Kusuma, R. F. Ningrum, P. C. Siswipraptini, and D. F. Murad, "Big data analytics of knowledge and skill sets for web development using latent Dirichlet allocation and clustering analysis," *Adv. Comput. Sci. Appl.*, no. January, 2025, doi: https://doi.org/10.14569/IJACSA.2025.0160123.

[14] A. G. Budianto, A. T. E. Suryo, A. F. Zulkarnain, G. R. Cahyono, R. Rusilawati, and S. F. Az-Zahra, "A text mining approach to analyzing the omnichannel retail business performance of the KlikIndomaret app," *J. Tek. Ind. J. Keilmuan dan Apl. Tek. Ind.*, vol. 26, no. 2, pp. 131–144, 2024, doi: https://doi.org/10.9744/jti.26.2.131-144.

[15] J. Brasse, M. Förster, P. Hühn, J. Klier, M. Klier, and L. Moestue, Preparing for the future of work : A novel data - driven approach for the identification of future skills, *Journal of Business Economics,* vol. 94, no. 3, pp. 467-500, 2024. doi: https://doi.org/10.1007/s11573-023-01169-1.

[16] P. C. Siswipraptini, H. Leslie, H. Spits, A. Ramadhan, and W. Budiharto, "Information technology job profile using average-linkage hierarchical clustering analysis," *IEEE Access*, vol. 11, no. August, pp. 94647–94663, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3311203.

[17] G. Melo, M. Chaves, M. K. B, and J. H. Schleifenbaum, "Skills requirements of additive manufacturing - a textual analysis of job postings using natural language processing," *Proceeding of AMPA: International Conference on Additive Manufacturing in Products and Applications,* pp. 299–316, 2023, doi: https://doi.org/10.1007/978-3-031-42983-5.

[18] F. De Felice, C. Salzano, I. Baffo, A. Forcina, and A. Petrillo, "Towards a sustainable digital manufacturing: A state of art," *Procedia Comput. Sci.*, vol. 232, pp. 1918–1929, 2024, doi: https://doi.org/10.1016/j.procs.2024.02.014.

[19] F. Acerbi, M. Rossi, and S. Terzi, "Identifying and assessing the required i4.0 skills for manufacturing companies ' workforce," *Front. Manuf. Technol.*, vol. 2, no. July, pp. 1–19, 2022, doi: https://doi.org/10.3389/fmtec. 2022.921445.

[20] S. Saniuk, "Knowledge and skills of industrial employees and managerial staff for the Industry 4.0 implementation," *Mob. Networks Appl.*, vol. 28, pp. 220–230, 2023, doi: https://doi.org/10.1007/s11036-021-01788-4.

[21] A. Islam, "Industry 4.0 : Skill set for employability," *Soc. Sci. Humanit. Open*, vol. 6, no. 1, p. 100280, 2022, doi: https://doi.org/10.1016/j.ssaho.2022.100280.

[22] Y. O. Abdallah, E. Shehab, and A. Al-ashaab, "Understanding digital transformation in the manufacturing industry : A systematic literature review and future trends," *Prod. Manag. Dev.*, vol. 19, no. 1, pp. 1–12, 2021, doi: https://doi.org/10.4322/pmd.2021.001.

[23] B. Gajdzik and R. Wolniak, "Smart production workers in terms of creativity and innovation : The implication for open innovation," *J. Open Innov. Technol. Mark. Complex*, vol. 8, no. 2, p. 68, 2022, doi: https://doi.org/10.3390/joitmc8020068.

[24] E. Beke, R. Horvath, and K. Takacs-Gyorgy, "Industry 4.0 and current competencies," *Sciendo*, vol. 66, no. 4, pp. 63–70, 2020, doi: https://doi.org/10.2478/ngoe-2020-0024.

[25] R. Gázquez *et al.*, "Lack of skills, knowledge, and competences in higher education about Industry 4.0 in the manufacturing sector," *Rev. Iberoam. Educ. a Distancia*, vol. 24, 2021, doi: https://doi.org/10.5944/ried.24.1.27548.