

DOI: https://doi.org/10.48009/2_iis_2026_113

Mapping AI job skill bundles: Frequency, co-occurrence, and market basket analysis of 500 AI job postings

Alan Peslak, *Penn State University*, arp14@psu.edu

Pratibha Menon, *Pennsylvania Western University*, menon@pennwest.edu

Abstract

This study analyzes 500 AI-related job postings collected from PostJobFree in Q1 2025 using a curated taxonomy of 120 employer-facing skill terms: 88 technical skills across AI/ML methods, programming languages, data tools, analytics, data engineering, and cloud/MLOps, plus 32 non-technical skills related to communication, leadership, and organizational competencies. A hybrid keyword-regular expression pipeline generated mention-level and posting-level frequency measures, identifying 11,250 total mentions and 4,106 posting-level occurrences. Frequency analyses include all 120 skills, while co-occurrence and association-rule analyses focus on the 88 technical skills to preserve analytical signal. Results show that AI hiring is organized around stable skill bundles, with strong associations linking core AI/ML skills, cloud platforms, deep-learning frameworks, data-engineering tools, and SQL. Grounded in human capital theory and skill complementarity, the findings provide empirical support for AI curriculum design, workforce planning, and program learning outcomes. Because the corpus comes from a single job-posting aggregator during Q1 2025, the findings should be interpreted as a cross-sectional snapshot of advertised AI skill bundles rather than a fully representative estimate of the broader AI labor market.

Keywords: artificial-intelligence, job-postings, market-basket-analysis, skill-co-occurrence, workforce-analytics, human-capital-theory

Introduction

Artificial intelligence technologies are transforming industries at an accelerating pace, creating acute demand for professionals capable of designing, deploying, and governing intelligent systems (Davenport & Ronanki, 2018). Yet the composition of AI skill sets is rarely described with the empirical precision required for curriculum design or workforce planning. Most existing analyses catalogue the most frequently mentioned individual skills without examining how those skills combine within real job roles, which is a limitation that matters because skills do not generate value in isolation (Siau & Yang, 2017).

Human capital theory (Becker, 1964) frames employer skill demands as signals of productive capacity. The economics of skill complementarity (Autor et al., 2003) extends this: the marginal return to any skill depends on the skills with which it is combined. A practitioner proficient in PyTorch but lacking deployment infrastructure knowledge, or versed in LLMs but without cloud experience, may be systematically underprepared for the roles employers advertise, not because any single skill is absent, but because the bundle is incomplete.

This study applies three complementary approaches to 500 AI-related job postings: dual-frequency analysis distinguishing mention-level intensity from posting-level prevalence, pairwise co-occurrence analysis of the technical skill network, and market basket analysis using the Apriori algorithm. Three research questions guide the analysis: what skills appear most frequently; which co-occur within the same postings; and what directional association rules characterize AI skill bundles and why. The study contributes a validated 120-skill ESCO/O*NET-anchored taxonomy, a reproducible dual-frequency extraction methodology, and some of the first market-basket analysis (MBA) applied specifically to AI labor market data.

Literature Review

Research at the intersection of computing education and labor market analytics has established job-posting analysis as the primary empirical method for identifying emerging skill requirements (Litecky et al., 2010; Burning Glass Technologies, 2021). Postings reflect employer language directly, can be collected at scale, and provide a real-time signal of evolving demand that surveys and curriculum audits cannot match.

Studies of AI skill demand have consistently identified a core of technical requirements: programming languages (especially Python), machine learning frameworks, and cloud infrastructure (Acemoglu et al., 2022; Frank et al., 2019; Manyika et al., 2017). More recent analyses document the rapid ascent of LLM-related skills following the emergence of generative AI systems in 2022 (Eloundou et al., 2023; Hui et al., 2024). These studies, however, report marginal frequencies rather than the co-occurrence structure that defines real roles, limiting their utility for constructing role-aligned curricula.

IS and computing education researchers have long emphasized alignment between curricula and industry demands (ACM/IEEE-CS Joint Task Force, 2020; Topi et al., 2010). Association rule mining has been applied to labor-market data to identify co-occurring skill pairs in software engineering and cybersecurity postings (Alekseeva et al., 2021; Squicciarini & Nachtigall, 2021), demonstrating that Apriori-derived rules can surface stable, interpretable bundles invisible to frequency analysis alone. Recent large-scale analyses of AI job postings confirm that employer skill signals have become increasingly specific and progressively decoupled from formal degree requirements, with AI role demand growing 21% as a proportion of all postings between 2018 and 2024 while concurrent university education requirements declined by 15% (Bone et al., 2025). The present study combines both approaches, grounded in a taxonomy explicitly crosswalked to ESCO (European Commission, 2022 ; Chiarello et al., 2021 ; Samek et al., 2021) and O*NET (Handel, 2016), applying them specifically to the AI labor market.

Methodology

Dataset

The dataset consists of 500 AI-related job postings collected from PostJobFree (postjobfree.com) during the first quarter of 2025. PostJobFree aggregates employer-submitted postings across industries and geographies. The corpus exceeds 340,000 words of job-description text. Each posting was treated as a single document for text analysis and as a single transaction for association-rule analysis. PostJobFree was selected because it provided accessible employer-facing job-description text suitable for reproducible skill extraction and transaction-level analysis. However, the platform should not be treated as statistically representative of the entire AI labor market. Job boards differ in employer mix, role seniority, geography, sector concentration, recruiter posting practices, and the level of technical specificity included in postings. Therefore, the results are best interpreted as evidence of structured skill bundles within this corpus, not as

population estimates for all AI jobs. Replication across LinkedIn, Indeed, Dice, and sector-specific technology job boards would strengthen external validity.

Skill Taxonomy

The taxonomy comprises 120 employer-facing skill terms across nine categories. Seven technical categories contain 88 terms spanning AI/ML Core, NLP/LLM/GenAI, Programming Languages, ML Data Tools and Libraries, Data/Analytics/Statistics, Data Engineering/Databases, and Cloud/MLOps/DevOps. Two non-technical categories contain 32 terms across Communication and Collaboration and Leadership and Management. Construction followed a three-stage corpus-driven procedure: candidate identification from the 500 postings, consolidation by normalizing variant expressions to canonical labels (e.g., ML → Machine Learning), and external validation by mapping each term to ESCO v1.2.1 and O*NET (Lassébie et al., 2021). The full taxonomy with regex patterns and ESCO mappings is documented in Appendix Tables A1–A3. Table 1 summarizes the nine categories.

Table 1. Skill Taxonomy Categories (N = 120 terms)

Category	Illustrative Terms	Terms	ESCO Concepts
AI / Machine Learning Core	Artificial Intelligence, Machine Learning, Deep Learning, Computer Vision	12	12
NLP / LLM / GenAI	LLMs, Generative AI, NLP, RAG, Transformers, Prompt Engineering	13	8
Programming Languages	Python, SQL, Java, R, Go, JavaScript	12	7
ML / Data Tools & Libraries	PyTorch, TensorFlow, scikit-learn, LangChain, Hugging Face	13	6
Data / Analytics / Statistics	Analytics, Predictive Modeling, Data Visualization, Statistical Analysis	12	9
Data Engineering / Databases	Data Engineering, ETL, Data Pipelines, Spark, Kafka	13	10
Cloud / MLOps / DevOps	AWS, Azure, Google Cloud, Docker, Kubernetes, MLOps, CI/CD	13	11
Communication & Collaboration*	Collaboration, Communication, Stakeholder Management, Mentoring	16	15
Leadership & Management*	Innovation, Strategic Thinking, Leadership, Organizational Skills	16	15

* Non-technical categories. All 120 terms extracted and reported in frequency analyses. Shaded rows excluded from co-occurrence and MBA analyses due to high base rates (see Methodology).

Skill Extraction

Skill extraction used a four-stage hybrid keyword–regex workflow. Job descriptions were lowercased, then scanned for canonical skills and common abbreviations with multi-option patterns (e.g., (machine learning|ml), (google cloud|gcp)). As shown in Appendix Table A2, word boundaries (\b) were applied to single-word or short abbreviations to avoid false substring matches (e.g., blocking “html” within “xhtml”), while multi-word phrases did not require them. All variant expressions were normalized to a canonical analytical label before counting. The dual-frequency design retains both measures because they capture distinct employer signals: mention-level frequency counts every occurrence of a skill expression, measuring emphasis and salience; posting-level frequency deduplicates results so each skill is counted at most once per posting, measuring breadth of demand across employers. All 120 skills pass through extraction and frequency reporting; co-occurrence and MBA analyses are restricted to the 88-term technical subset because non-technical skills (Collaboration in 71.6% of postings; Communication in 52.0%) produce uniformly

dense co-occurrence structures driven by prevalence rather than genuine skill complementarity. Table 2 summarizes the two frequency measures.

Co-Occurrence and Market Basket Analysis

The posting-level dataset restricted to the 88-term technical subset was used for two downstream structural analyses. For co-occurrence analysis, each posting was converted to its set of technical skills and all pairwise combinations were enumerated; joint support was calculated as the proportion of postings containing both skills. Market basket analysis applied the Apriori algorithm (mlxtend; Raschka, 2018) with rules filtered to support $\geq 5\%$ and lift > 1.2 to maintain interpretability. All analyses were conducted in Python 3 (Van Rossum & Drake, 2009) using pandas (McKinney, 2010), NetworkX (Hagberg et al., 2008), mlxtend (Raschka, 2018), and matplotlib (Hunter, 2007).

Table 2. Dual Frequency Measures

Measure	Counting Rule	Interpretation	Corpus Result
Mention-level	Every occurrence counted	Employer emphasis within postings	11,250 total mentions
Posting-level	Each skill counted once per posting	Prevalence across employers; MBA input	4,106 posting-level occurrences

Results

Mention-Level Skill Frequency

The mention-level analysis identified 11,250 skill expressions across the 500-posting corpus (Table 3). Artificial Intelligence dominated with 4,966 mentions, an average of 9.9 per posting, confirming that AI is not merely a required skill but the pervasive contextual frame within which every other competency in these postings is situated. Machine Learning followed at 1,309 mentions (2.6 per posting). The 3.8:1 AI-to-ML mention ratio suggests AI is invoked repeatedly as narrative context while ML is named more specifically when technical requirements are articulated.

Analytics (360) and Large Language Models (350) ranked third and fourth, reflecting simultaneous demand for traditional analytical reasoning and generative AI capabilities. The three cloud platforms appeared at nearly identical mention rates, Azure (159), AWS (154), Google Cloud (154), signaling multi-cloud literacy as a baseline expectation. PyTorch (160) slightly led TensorFlow (155), but their proximity suggests employers do not differentiate meaningfully between the two frameworks.

Table 3. Top 15 Skills by Mention-Level Frequency (N = 500 postings; 11,250 total mentions)

Rank	Skill	Mentions
1	Artificial Intelligence	4,966
2	Machine Learning	1,309
3	Analytics	360
4	Large Language Models	350
5	Algorithms	285
6	Python	279
7	Generative AI	278
8	Data Science	276
9	Natural Language Processing	191

10	Deep Learning	182
11	PyTorch	160
12	Azure	159
13	TensorFlow	155
14	AWS	154
15	Google Cloud	154

Posting-Level Skill Frequency

Posting-level frequency reveals how broadly a skill is required across employers (Table 4). Artificial Intelligence appeared in 483 postings (96.6%), confirming it as the near-universal defining characteristic of the corpus. Machine Learning followed at 329 postings (65.8%), meaning it appeared in two-thirds of all postings. Python stood alone as the dominant programming language at 241 postings (48.2%), a substantial lead over the next language in the taxonomy; no other programming language exceeded 20% posting prevalence, suggests that Python is a foundational requirement in this corpus for AI work in 2025.

Data Science (33.6%), Analytics (32.4%), and Algorithms (29.6%) formed a secondary analytical tier. Large Language Models (29.2%), PyTorch (29.0%), and TensorFlow (27.2%) each appeared in roughly one in four postings. Deep Learning and Generative AI were co-equal at 24.4%. The three cloud platforms clustered tightly: AWS (23.6%), Azure (21.4%), Google Cloud (20.4%), supporting the MBA finding that multi-cloud literacy appears to be a recurring structural expectation in this corpus.

Table 4. Top 15 Skills by Posting-Level Frequency — Technical Subset (N = 500 postings)

Rank	Skill	Postings (n)	% Postings
1	Artificial Intelligence	483	96.6%
2	Machine Learning	329	65.8%
3	Python	241	48.2%
4	Data Science	168	33.6%
5	Analytics	162	32.4%
6	Algorithms	148	29.6%
7	Large Language Models	146	29.2%
8	PyTorch	145	29.0%
9	TensorFlow	136	27.2%
10	Deep Learning	122	24.4%
11	Generative AI	122	24.4%
12	AWS	118	23.6%
13	Azure	107	21.4%
14	Google Cloud	102	20.4%
15	Natural Language Processing	100	20.0%

Category-Level Summary

Table 5a aggregates counts by technical taxonomy category. AI/Machine Learning Core dominated both measures (7,417 mentions; 1,529 summed posting-level occurrences), reflecting the centrality and intensity of AI/ML demand. NLP / LLM / GenAI ranks second with 1,058 mentions, Cloud / MLOps / DevOps ranks third (776 mentions) and Data/Analytics/Statistics fourth (720 mentions). Data

Engineering/Databases ranked last in mentions (184) but maintained a meaningful posting footprint (157), consistent with its role as a specialized submarket rather than a universal requirement.

Non-technical categories, excluded from structural analysis, recorded substantial frequency counts: Communication and Collaboration totaled 2,550 mentions and 1,607 summed posting-level occurrences; Leadership and Management totaled 2,600 mentions and 1,461 occurrences. At the individual skill level (Table 5b), Collaboration appeared in 358 postings (71.6%), Innovation in 283 (56.6%), and Communication in 260 (52.0%), confirming that these skills pervade job descriptions regardless of technical role and precisely, this high-base-rate pattern justifies their exclusion from co-occurrence and MBA analyses.

Table 5a. Technical Category-Level Frequency Totals (N = 500 postings; 88-skill technical subset)

Category	Mention Count	Summed Posting-Level Occurrences
AI / Machine Learning Core	7,417	1,529
NLP / LLM / GenAI	1,058	492
Programming Languages	651	544
Cloud / MLOps / DevOps	776	573
ML / Data Tools & Libraries	444	389
Data Engineering / Databases	184	157
Data / Analytics / Statistics	720	422

Note. Summed posting-level occurrences total across all skills in each category; a single posting may contribute to multiple skills within a category.

Table 5b. Non-Technical Skill Frequency (reported for completeness; excluded from structural analysis)

Skill	Postings (n)	% Postings	Category
Collaboration	358	71.6%	Communication & Collaboration
Innovation	283	56.6%	Leadership & Management
Communication	260	52.0%	Communication & Collaboration
Strategic Thinking	196	39.2%	Leadership & Management
Stakeholder Management	187	37.4%	Communication & Collaboration

Note. Near-universal posting-level presence confirms these skills' role as baseline organizational expectations. Their high base rates justify exclusion from co-occurrence and MBA analyses.

Skill Co-Occurrence Analysis

Pairwise co-occurrence analysis of the 88-term technical subset identified the combinations most commonly appearing together (Table 6). The Artificial Intelligence–Machine Learning pair led with 64.2% joint support, which is nearly two in three postings required both simultaneously, confirming that employers treat AI and ML as a single productive bundle. Artificial Intelligence–Python (46.6%) and Machine Learning–Python (40.2%) established Python as the implementation language for this core triadic bundle.

Deep-learning framework pairings: AI–PyTorch (28.2%), PyTorch–Python (27.6%), Python–TensorFlow (26.6%), AI–TensorFlow (26.2%), ML–PyTorch (25.6%), and ML–TensorFlow (24.6%), clustered in a

narrow 24–29% range. This tight clustering indicates that PyTorch and TensorFlow participate nearly interchangeably in the core technical stack, and posting language does not appear to differentiate strongly between the two frameworks in their hiring. Large Language Models–Machine Learning (23.8%) confirmed that LLM skills are not a standalone specialty but an extension of the broader ML infrastructure. AWS–Artificial Intelligence (23.6%) tied cloud deployment capability directly to the AI/ML core.

Table 6. Strongest Skill Co-Occurrences in the Posting-Level Matrix (N = 500; technical subset)

Skill A	Skill B	Joint Support
Artificial Intelligence	Machine Learning	64.2%
Artificial Intelligence	Python	46.6%
Machine Learning	Python	40.2%
Artificial Intelligence	Large Language Models	28.6%
Artificial Intelligence	PyTorch	28.2%
PyTorch	Python	27.6%
Python	TensorFlow	26.6%
Artificial Intelligence	TensorFlow	26.2%
Machine Learning	PyTorch	25.6%
Machine Learning	TensorFlow	24.6%
Large Language Models	Machine Learning	23.8%
AWS	Artificial Intelligence	23.6%

Market Basket Analysis

Market basket analysis of the 88-term technical subset produced 12 high-lift association rules (Table 7b; metrics defined in Table 7a). The strongest group links cloud platforms: Azure → Google Cloud (lift = 3.62), AWS → Azure (lift = 3.48), and AWS → Google Cloud (lift = 3.32) show that requiring any one cloud platform is a near-directional predictor of requiring the others, with confidence values of 67–82%.

Table 7a. Association-Rule Metrics

Measure	Meaning	Interpretive Use
Support	Proportion of postings containing both skills	Tables emphasize $\geq 5\%$
Confidence	$P(\text{consequent} \mid \text{antecedent})$	Stable directional rules
Lift	Association strength relative to chance	Tables emphasize > 1.2

Table 7b. Selected High-Lift Association Rules (support $\geq 5\%$; lift > 1.2 ; N = 500; technical subset)

Rule	Support	Confidence	Lift
Azure → Google Cloud	15.8%	73.8%	3.62
AWS → Azure	17.6%	74.6%	3.48
AWS → Google Cloud	16.0%	67.8%	3.32
PyTorch → TensorFlow	23.6%	81.4%	2.99
Deep Learning → PyTorch	17.8%	73.0%	2.52
Deep Learning → TensorFlow	15.8%	64.8%	2.38
Google Cloud → TensorFlow	12.4%	60.8%	2.23

AWS → SQL	7.4%	31.4%	2.15
Deep Learning → Natural Language Processing	10.4%	42.6%	2.13
Google Cloud → SQL	6.2%	30.4%	2.08
AWS → TensorFlow	13.2%	55.9%	2.06
Azure → SQL	6.4%	29.9%	2.05

PyTorch → TensorFlow (lift = 2.99, confidence = 81.4%) shows four in five PyTorch-requiring postings also require TensorFlow. Deep Learning → PyTorch (lift = 2.52) and Deep Learning → TensorFlow (lift = 2.38) confirm deep learning as a strong antecedent for both frameworks. Google Cloud → TensorFlow (lift = 2.23) links GCP to TensorFlow, reflecting Google’s role in developing the ecosystem. AWS → SQL (lift = 2.15), Google Cloud → SQL (lift = 2.08), and Azure → SQL (lift = 2.05) establish relational querying as a consistent co-requirement across all three cloud platforms. Deep Learning → NLP (lift = 2.13) reflects field convergence since the transformer architecture (Vaswani et al., 2017).

Discussion

Association rules are interpreted as co-occurrence patterns, not causal relationships. They indicate that skills are advertised together more often than expected by chance within this corpus; they do not establish why employers combine them or whether one skill requirement causes another.

The AI–ML–Python triadic core

The co-occurrence results establish a clear hierarchical structure in AI skill demand. The AI–ML–Python triad, present in 40–65% of postings depending on the pair, appears to function as a central technical foundation in the corpus of modern AI roles. Python’s 48.2% posting prevalence, combined with its position as the implementation language for both ML modeling (ML–Python co-occurrence: 40.2%) and deep-learning frameworks (PyTorch–Python: 27.6%), confirms that Python literacy is not merely desirable but foundational. IS programs that allow students to develop AI competencies in other languages without Python proficiency may leave graduates less aligned with observed employer requirements in this corpus.

The near-parity of PyTorch and TensorFlow across both posting-level frequency (29.0% vs 27.2%) and MBA confidence values reflects an industry-wide expectation of framework agnosticism. Employers are not hiring for a single framework but for deep-learning capability that spans both. This pattern is consistent with organizational technical debt and mixed production/development environments. Enterprises maintain TensorFlow-based production pipelines built during TensorFlow’s earlier dominance while adopting PyTorch for new model development (Paszke et al., 2019). Graduates who know only one framework may be unable to service existing systems while building new ones.

Multi-cloud literacy as structural expectation

The three cloud platform MBA rules: AWS → Azure (lift = 3.48), Azure → Google Cloud (lift = 3.62), AWS → Google Cloud (lift = 3.32), are the highest-lift rules in the dataset and together constitute the most actionable curriculum finding. Confidence values of 67–82% mean that requiring one major cloud platform is a near-directional predictor of requiring the others. This pattern is consistent with common enterprise architecture practices. Enterprise AI workloads are distributed across providers by design for regulatory compliance, training/serving splits, and disaster-recovery redundancy. IS programs that teach a single cloud platform as a specialization are underestimating the breadth of multi-cloud literacy that employers already treat as a baseline expectation.

SQL as the AI pipeline boundary skill

The SQL rules: AWS → SQL (lift = 2.15), Google Cloud → SQL (lift = 2.08), Azure → SQL (lift = 2.05), establish relational querying as a consistent co-requirement across all three cloud platforms. This challenges the assumption that SQL is a legacy skill displaced by ML pipelines. Production AI systems interface with organizational data infrastructure through SQL: feature stores, data warehouses, and model evaluation pipelines all depend on relational queries. The uniform lift values across platforms indicate a stable structural pattern, not a cloud-specific quirk. AI practitioners without SQL fluency cannot fully participate in the data infrastructure layer that production systems require.

Field convergence: Deep Learning and NLP

Deep Learning → NLP (lift = 2.13) captures the structural shift since the transformer architecture (Vaswani et al., 2017). Pre-transformer NLP was largely separable from deep learning. Post-transformer NLP is definitionally deep learning: BERT, GPT, and their successors are deep neural networks requiring the same PyTorch or TensorFlow foundations as other DL applications. In this corpus, NLP appears less as a standalone specialization and more as a deep-learning-linked application area. Curricula that teach NLP as a separate track without DL foundations are organized around a pre-2017 division of the field.

LLM and Generative AI skills as an application-layer extension

The results suggest that LLM and Generative AI skills are emerging as a distinct but dependent layer within AI hiring, rather than replacing the traditional AI/ML foundation. Large Language Models appear in 29.2% of postings and Generative AI in 24.4%, placing them near PyTorch, TensorFlow, and Deep Learning in posting-level prevalence. However, the co-occurrence results show that LLMs remain connected to the broader machine-learning infrastructure: the Large Language Models–Machine Learning pair appears in 23.8% of postings, and Artificial Intelligence–Large Language Models appears in 28.6%. This pattern suggests that employers often treat LLM/GenAI capability as an extension of the AI/ML stack, requiring practitioners to combine prompt-oriented, RAG, or transformer-related capabilities with conventional modeling, programming, cloud, and data skills. Curriculum design should therefore avoid treating Generative AI as a standalone module detached from ML, Python, data engineering, and governance foundations.

Non-technical skills as organizational baselines

Collaboration (71.6%), Innovation (56.6%), and Communication (52.0%) pervade AI job descriptions regardless of the specific technical role. Their near-universal presence is why they were excluded from structural analysis, with base rates above 50%, any co-occurrence would be driven by prevalence rather than complementarity. Their ubiquity is nonetheless a substantive finding: IS programs that treat collaboration, communication, and stakeholder management as peripheral rather than core requirements are misreading the employer signal. These are threshold expectations, not differentiating signals.

Implications for IS curriculum

Three curriculum recommendations follow. First, the AI–ML–Python–frameworks bundle should be taught as an integrated co-requisite stack. Second, multi-cloud literacy should be a program requirement given that 67–82% of cloud-requiring postings require competency on at least two platforms. Third, SQL should be positioned as a production-critical bridge skill, not legacy content. These recommendations extend IS2020 guidelines (Topi et al., 2010) with empirical specificity about which skill combinations matter most in the 2025 AI labor market.

Limitations

Four limitations apply. First, the corpus reflects 500 postings from PostJobFree during a single Q1 2025 collection window. This creates two external-validity limits. Platform composition may differ from LinkedIn, Indeed, Dice, company career sites, and sector-specific boards in employer mix, geography,

seniority distribution, and posting detail. Temporal scope is also limited because AI skill demand changes rapidly as generative AI tools, model platforms, and enterprise adoption patterns evolve. The findings should therefore be interpreted as a cross-sectional snapshot of skill-bundle structure in this corpus rather than as a definitive representation of the entire AI job market. Future research should replicate the same taxonomy and MBA framework across multiple platforms and repeated collection windows. Second, the rule-based pipeline may miss contextual synonyms and emerging terminology outside the defined taxonomy; transformer-based NER (Devlin et al., 2019) could complement this approach. Third, structural analyses are restricted to the 88-term technical subset; non-technical skill relationships remain a complementary research opportunity. Fourth, the taxonomy is anchored to ESCO v1.2.1 and future updates may introduce new concepts not captured here. Association rules throughout describe co-occurrence structure, not causation, and cannot distinguish equally required from aspirationally preferred skills.

Future work

A related extension would analyze non-technical skills in a separate matrix rather than mixing them with technical skills. Because communication, collaboration, innovation, and stakeholder-management terms have very high base rates, including them in the same structural analysis would make technical association rules less interpretable. A non-technical-only co-occurrence analysis could nevertheless reveal whether leadership, communication, stakeholder, and organizational skills form distinct soft-skill bundles.

Conclusion

This study provides an integrated empirical analysis of AI job skill bundles using 500 job postings, a 120-skill ESCO/O*NET-validated taxonomy, and three complementary analytical approaches. The core findings are consistent across all methods: AI hiring is organized around stable bundles of complementary technical skills, the AI–ML–Python triadic core is effectively universal, multi-cloud literacy is a structural employer expectation rather than a specialization, SQL is a production-critical boundary skill between AI pipelines and organizational data infrastructure, and non-technical competencies are threshold organizational expectations present in every role type.

The dual-frequency design, distinguishing mention intensity from posting prevalence, and the mechanistic interpretation of each MBA rule group are methodological contributions that extend beyond the specific findings. The most valuable extension of this work would apply the same framework longitudinally and across multiple platforms to track how AI skill bundles shift as generative AI matures, providing a richer evidence base for curriculum evolution decisions.

Acknowledgements

The authors acknowledge the use of Claude (Anthropic) as a generative AI tool to assist with document preparation during the production of this manuscript. All analytical decisions, interpretations, and written content are the authors' own.

References

- ACM/IEEE-CS Joint Task Force on Computing Curricula. (2020). Computing curricula 2020: Paradigms for global computing education. Association for Computing Machinery.
<https://doi.org/10.1145/3467967>

- Alekseeva, L., Azar, J., Gimé, M., Samila, S., & Taska, B. (2021). The demand for AI skills in the labor market. *Labour Economics*, 71, 102002. <https://doi.org/10.1016/j.labeco.2021.102002>
- Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4), 1279–1333. <https://doi.org/10.1162/003355303322552801>
- Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40(S1), S293–S340. <https://doi.org/10.1086/718327>
- Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago Press.
- Bone, M., González Ehlinger, A., & Stephany, F. (2025). Skills or degree? The rise of skill-based hiring for AI and green jobs. *Technological Forecasting and Social Change*, 214, 124042. <https://doi.org/10.1016/j.techfore.2025.124042>
- Burning Glass Technologies. (2021). The new foundational skills of the digital economy. Burning Glass Technologies. https://www.bhef.com/sites/default/files/BHEF_2018_New_Foundational_Skills.pdf
- Samek, L., Squicciarini, M., & Cammeraat, E. (2021). The human capital behind AI: Jobs and skills demand from online job postings. *OECD Science, Technology and Industry Policy Papers*, No. 120, OECD Publishing. <https://doi.org/10.1787/2e278150-en>
- Chiarello, F., Spada, I., Marraccini, E., & Fantoni, G. (2021). Towards ESCO 4.0: Is the European classification of skills in line with Industry 4.0? *Technological Forecasting and Social Change*, 173, 121177. <https://doi.org/10.1016/j.techfore.2021.121177>
- Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, 96(1), 108–116.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *Science*, 384(6702), 1306–1308. <https://doi.org/10.1126/science.adj0998>
- European Commission. (2022). Digital skills and knowledge concepts labelling initiative. *ESCO Data Science Blog*. <https://esco.ec.europa.eu/en/about-esco/data-science-and-esco/digital-skills-and-knowledge-concepts>
- Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Wang, D., Youn, H., & Rahwan, I. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116(14), 6531–6539. <https://doi.org/10.1073/pnas.1900949116>

- Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference*, 11–15.
- 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>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Lassébie, J., Morikawa, L., Sakha, S., Squicciarini, M., Samek, L., & Venturini, C. (2021). A machine learning approach to classify skills in online job postings. *OECD Social, Employment and Migration Working Papers*, No. 255. [https://one.oecd.org/document/DELSA/ELSA/WD/SEM\(2021\)10/en/pdf](https://one.oecd.org/document/DELSA/ELSA/WD/SEM(2021)10/en/pdf)
- Litecky, C., Aken, A., Ahmad, A., & Nelson, H. J. (2010). Mining for computing jobs. *IEEE Software*, 27(1), 78–85. <https://doi.org/10.1109/MS.2009.150>
- Squicciarini, M., & Nachtigall, H. (2021). Demand for AI skills in jobs: Evidence from online job postings. *OECD Science, Technology and Industry Working Papers*, No. 2021/03, OECD Publishing. <https://doi.org/10.1787/3ed32d94-en>
- Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R., & Sanghvi, S. (2017). *Jobs lost, jobs gained: Workforce transitions in a time of automation*. McKinsey Global Institute.
- McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 56–61. <https://doi.org/10.25080/Majora-92bfl922-00a>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32. <https://doi.org/10.48550/arXiv.1912.01703>
- Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software*, 3(24), 638. <https://doi.org/10.21105/joss.00638>
- Siau, K., & Yang, Y. (2017). Impact of artificial intelligence, robotics, and automation on higher education. *Proceedings of the Twenty-Third Americas Conference on Information Systems (AMCIS)*. <https://www.proceedings.com/content/038/038301webtoc.pdf>
- Topi, H., Valacich, J. S., Wright, R. T., Kaiser, K., Nunamaker, J. F., Sipior, J. C., & de Vreede, G. J. (2010). IS 2010 curriculum guidelines for undergraduate degree programs in information systems. *Communications of the Association for Information Systems*, 26(1), 359–428. <https://doi.org/10.17705/1CAIS.02618>
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. <https://doi.org/10.48550/arXiv.1706.03762>

Hui, X., Reshef, O., & Zhou, L. (2024). The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market (July 31, 2023).
<http://dx.doi.org/10.2139/ssrn.4527336>

Appendix

AI Skill Taxonomy, ESCO/O*NET Crosswalk, and Extraction Rules

This appendix documents the complete 120-skill taxonomy. Appendix Table A1 presents representative ESCO/O*NET crosswalk entries for the 88 technical skills. Appendix Table A2 provides the full technical taxonomy with regex extraction patterns and frequency counts. Appendix Table A3 documents the 32 non-technical skill terms. All 120 terms were extracted and reported in frequency analyses; only the 88 technical terms in Table A2 were used in co-occurrence and association-rule analyses.

Appendix Table A1. Representative ESCO/O*NET Crosswalk for Technical Skills

Category	Employer Term	ESCO Source Concept	ESCO Type	O*NET Reference	Mentions	Postings
AI / ML Core	Artificial Intelligence (AI)	Artificial intelligence	Knowledge	Artificial Intelligence	4,966	483
AI / ML Core	Machine Learning	Machine learning	Skill	Machine Learning	1,309	329
AI / ML Core	Deep Learning	Deep learning	Knowledge	—	182	122
AI / ML Core	Computer Vision	Computer vision	Knowledge	—	124	76
NLP / LLM / GenAI	Natural Language Processing	Natural language processing	Skill	—	191	100
NLP / LLM / GenAI	Large Language Models	Language model development	Skill	—	350	146
NLP / LLM / GenAI	Generative AI	Generative artificial intelligence	Skill	—	248	114
Programming	Python	Python programming	Skill	Programming Languages	279	241
Programming	SQL	Structured Query Language	Skill	Databases	102	73
ML Tools	PyTorch	Use deep learning frameworks	Skill	—	160	145
ML Tools	TensorFlow	Use deep learning frameworks	Skill	—	155	136
Analytics	Analytics	Data analytics	Skill	Data Analysis	382	170
Data Engineering	Data Engineering	Data engineering	Skill	Data Engineering	51	40
Cloud / MLOps	AWS	Use cloud platforms	Skill	Cloud Platforms	154	118
Cloud / MLOps	Docker	Containerisation	Skill	DevOps	31	29
Cloud / MLOps	Kubernetes	Container orchestration	Skill	DevOps	45	34

Note. ESCO concept type: Skill = operational capability; Knowledge = domain knowledge concept. O*NET reference: — = no clean one-to-one mapping. Full 88-term crosswalk available from authors upon request.

Appendix Table A2. Full Technical Skill Taxonomy with Regex Extraction Patterns and Frequency Counts (88 terms)

Canonical Skill	Regex Pattern	Mentions	Postings
AI / ML Core			
Artificial Intelligence	(artificial intelligence ai)	4,966	483
Machine Learning	(machine learning ml)	1,309	329
Data Science	(data science)	269	166
Algorithms	(algorithms? algorithm design)	285	148
Deep Learning	(deep learning)	182	122
Computer Vision	(computer vision)	124	76
Model Training	(model training train models)	58	50
Reinforcement Learning	(reinforcement learning)	36	32
Neural Networks	(neural networks?)	34	27
Feature Engineering	(feature engineering)	32	25
Model Deployment	(model deployment deploy models)	30	28
Model Evaluation	(model evaluation)	22	17
NLP / LLM / GenAI			
Large Language Models	(large language models? llms?)	350	146
Generative AI	(generative ai genai)	248	114
Natural Language Processing	(natural language processing nlp)	191	100
RAG	(rag retrieval augmented generation)	69	34
Transformers	(transformers?)	44	32
Prompt Engineering	(prompt engineering)	20	20
Chatbots	(chatbots? conversational agents)	83	19
Embeddings	(embeddings?)	15	12
Named Entity Recognition	(named entity recognition ner)	2	2
Sentiment Analysis	(sentiment analysis)	2	2
Text Mining	(text mining)	2	2
Speech Processing	(speech recognition speech processing)	2	2
Information Extraction	(information extraction)	1	1
Programming Languages			
Python	\bpython\b	279	241
R	R programming	89	76
SQL	(sql structured query language)	102	73
Java	\bjava\b	74	62
Go	(golang go programming)	49	43
JavaScript	\bjavascript\b	30	25
TypeScript	\btypescript\b	7	7
Scala	\bscala\b	1	1
C++	c\+\+	1	1
C#	c#	0	0

Rust	\brust\b	0	0
Bash	(bash shell scripting)	1	1
ML / Data Tools			
PyTorch	\bpytorch\b	160	145
TensorFlow	\btensorflow\b	155	136
scikit-learn	(scikit-learn sklearn)	23	18
Hugging Face	hugging face	23	17
LangChain	langchain	19	16
Keras	\bkeras\b	15	13
Pandas	\bpandas\b	12	12
NumPy	\bnumpy\b	8	8
Matplotlib	\bmatplotlib\b	8	8
XGBoost	\bxgboost\b	6	3
spaCy	\bspacy\b	2	2
Jupyter	(jupyter notebooks? jupyterlab)	1	1
OpenCV	\bopencv\b	1	1
Data / Analytics			
Analytics	\banalytics?\b	382	170
Predictive Modeling	predictive modeling	26	26
Data Visualization	data visualization	26	24
Statistical Analysis	statistical analysis	20	20
Anomaly Detection	anomaly detection	15	14
Business Analytics	business analytics	12	7
Experimental Design	experimental design	7	7
A/B Testing	a/b testing	4	4
Data Mining	data mining	4	3
Clustering	clustering	2	2
Hypothesis Testing	hypothesis testing	2	2
Time Series Analysis	time series analysis	2	2
Data Engineering			
Data Pipelines	data pipelines	61	51
Data Engineering	data engineering data engineer	51	40
Data Warehousing	data warehousing	13	12
Apache Spark	spark	12	12
PostgreSQL	postgresql	12	12
NoSQL	nosql	12	10
ETL	etl	7	7
Relational Databases	relational databases	6	6
Data Lakes	data lakes	8	5
Database Design	database design	5	5
Hadoop	hadoop	2	2

Kafka	kafka	2	2
MongoDB	mongodb	1	1
Cloud / MLOps / DevOps			
AWS	aws amazon web services	154	118
Azure	azure	159	107
Google Cloud	google cloud gcp	154	102
CI/CD	ci/cd	61	41
MLOps	mlops	56	39
Kubernetes	kubernetes k8s	45	34
Docker	docker	31	29
Microservices	microservices	13	12
Infrastructure as Code	infrastructure as code iac	19	11
Terraform	terraform	11	10
Airflow	airflow	2	2
Deployment Pipelines	deployment pipelines? deployment automation	6	6
Model Monitoring	model monitoring model observability	0	0

Note 1. Regex patterns use Python re syntax with re.IGNORECASE flag. Mention count = all occurrences in corpus. Posting count = deduplicated occurrences across 500 postings. Terms with zero posting counts are retained because they represent widely recognized AI-ecosystem technologies. All 88 terms used in co-occurrence and association-rule analyses.

Note 2. To prevent false positives from common English words, single-letter or common-word languages (R, Go) were extracted using contextual modifiers (e.g., 'R programming').

**Appendix Table A3. Non-Technical Skill Taxonomy with Extraction Patterns and Frequency Counts
(representative sample of 32 terms)**

Canonical Skill	Regex Pattern	Mentions	Postings
Communication & Collaboration			
Communication	\bcommunication\b	342	260
Collaboration	(collaboration collaborat(e ing) teamwork cross-functional)	817	358
Stakeholder Management	(stakeholders? management stakeholders? engagement stakeholders?)	332	187
Mentoring	(mentor(ing ship)? coach(ing)?)	250	152
Problem Solving	(problem.solving problem solving)	138	123
Leadership & Management			
Innovation	(innovation innovative)	592	283
Strategic Thinking	(strategically? strategy)	472	196
Leadership	(leadership lead(ing)? team)	338	177
Organizational Skills	(organizational skills? organiz(e ing ation))	311	166
Roadmap	(roadmaps? roadmapping)	118	93

Note. Representative sample of the 32 non-technical skills. These skills were retained in frequency reporting but excluded from co-occurrence and MBA analyses due to high base rates. Full 32-term crosswalk available from authors upon request.