India Iceberg Index: Measuring AI Occupational Exposure Across 744 Districts

Somesh Mohapatra[1]

[1]Independent.

Contributing authors: someshm@alum.mit.edu

Abstract

Generative artificial intelligence represents a general-purpose technology targeting cognitive capabilities rather than routine manual labor, yet discourse on AI's workforce impact in India remains concentrated on the information technology sector. We introduce the India Iceberg Index, adapting the Project Iceberg methodology to quantify AI occupational exposure across 744 districts and 36 states and union territories using Periodic Labour Force Survey (PLFS) 2023–24 microdata covering 415,549 individuals. Our analysis reveals that approximately 41.5% of India's wage-weighted employment faces technical exposure to current-generation AI systems. The national mean Iceberg Index is 37.74 (SD = 9.38) at the district level, with values ranging from 13.41 to 70.56. Critically, visible technology-sector exposure (Surface Index) constitutes only 0.96% of total employment, while hidden cognitive exposure in administrative, financial, and clerical occupations—the "submerged mass"—extends to districts conventionally considered outside AI's reach. Districts exhibiting maximum "surprise" include The Dangs (Gujarat), Central Delhi, and Chitrakoot (Uttar Pradesh), where Iceberg Index values exceed 60 despite near-zero technology-sector presence. Rural areas show higher mean exposure (49.52) than urban areas (31.46), driven by formal administrative employment in government and banking. Industry concentration analysis using the Herfindahl-Hirschman Index reveals that 55.6% of districts exhibit distributed exposure patterns requiring multi-sector policy coordination. These findings challenge the prevailing "tech problem" framing and suggest that India's service-sector development model faces structural vulnerability as AI capabilities expand into entry-level cognitive work.

Keywords: Artificial Intelligence, Labor Markets, Occupational Exposure, India, Automation, Service Sector, Digital Economy

1 Introduction

Artificial intelligence is transforming global labor markets through mechanisms fundamentally different from prior waves of automation. Where industrial robots and computerization targeted routine manual and computational tasks, large language models (LLMs) demonstrate capability across cognitive functions: information synthesis, text generation, scheduling, coordination, and analysis [1]. This technological shift carries particular implications for economies that have built development strategies around service-sector expansion.

India represents a critical case for understanding these dynamics. The country's post-liberalization growth model has relied substantially on service-sector employment as the primary channel for upward mobility [3]. The information technology and business process outsourcing (IT-BPO) industry, employing approximately 5.4 million workers directly and generating $245 billion in revenue, has anchored this strategy [4]. Entry-level positions in call centers, data processing, and software development have served as "first rungs" on the formal employment ladder for graduates from India's expanding higher education system.

The prevailing discourse on AI's labor market impact in India frames it as a sectoral challenge confined to technology workers. Headlines focus on potential disruption to IT services, coding jobs, and data science roles. This framing treats AI as a vertical shock to specific industries rather than a horizontal transformation of cognitive work across the economy.

We argue this interpretation fundamentally mischaracterizes the nature of LLM-based AI systems. Unlike previous automation technologies, current AI capabilities target the core cognitive tasks that define white-collar employment: processing text, synthesizing information, generating reports, and coordinating workflows. These tasks are not exclusive to technology workers; they constitute the operational foundation of government administration, banking, education, and professional services throughout India.

To quantify this broader exposure, we adapt the Project Iceberg methodology [2], which measures the wage-weighted share of occupational tasks that AI systems can technically perform. The framework distinguishes between the "surface"—visible technology-sector disruption concentrated in software development and data science—and the "iceberg"—the larger submerged mass of cognitive exposure extending through administrative, financial, and clerical occupations nationwide.

Our analysis processes PLFS 2023–24 microdata covering 415,549 individuals across 744 districts, mapping 3,445 National Classification of Occupations (NCO-2015) codes to AI Occupational Exposure (AIOE) scores derived from the Felten et al. framework [1]. We calculate wage-weighted exposure indices at district, state, and sectoral levels, decomposing by urban-rural geography and industry concentration.

The results reveal that approximately 41.5% of India's wage-weighted employment faces technical exposure to current AI systems—a figure substantially higher than visible technology-sector adoption would suggest. The technology sector accounts for only 0.96% of total employment, yet cognitive exposure extends to districts with minimal

technology presence. This pattern indicates structural vulnerability in India's service-sector development model that transcends the boundaries of IT parks and startup ecosystems.

This paper makes three contributions. First, we provide the first comprehensive district-level mapping of AI occupational exposure for India, enabling geographically targeted policy analysis. Second, we identify substantial "hidden" exposure in administrative and clerical occupations that may not be recognized in technology-focused workforce planning. Third, we characterize the industry concentration structure of exposure, distinguishing districts requiring sector-specific intervention from those requiring broad-based coordination.

2 Results

2.1 National Exposure Profile

Analysis of the PLFS 2023–24 microdata yields AI occupational exposure estimates for 744 districts across 36 states and union territories. The wage-weighted national Iceberg Index stands at 41.5%, indicating that approximately two-fifths of India's formal labor market wage value involves tasks where current AI systems demonstrate technical capability.

At the district level, the mean Iceberg Index is 37.74 (SD = 9.38), ranging from a minimum of 13.41 to a maximum of 70.56 (Table 1). The distribution exhibits moderate positive skew, with the median (37.91) falling slightly above the mean. The interquartile range spans 31.33 to 44.16, indicating substantial variation in exposure across districts.

Table 1 Summary Statistics: District-Level Iceberg Index (N = 744)

Statistic	Mean	SD	Min	Median	Max
Iceberg Index	37.74	9.38	13.41	37.91	70.56
Surface Index	22.63	28.53	0.00	0.00	68.05
Surprise Index	15.10	28.74	−36.52	32.44	70.56
Industry HHI	1645	956	372	1380	9573

The Surface Index, measuring exposure within explicitly technology-sector occupations (software developers, database professionals, ICT technicians), averages 22.63 across districts. However, this figure is misleading: only 267 of 744 districts (35.9%) show any technology-sector employment at all. Among districts with technology-sector presence, the Surface Index averages 63.2%, reflecting the high automatability of software development and data processing tasks. Nationally, technology-sector employment constitutes merely 0.96% of total weighted employment.

The Surprise Index—the difference between Iceberg and Surface indices—quantifies hidden cognitive exposure beyond visible technology adoption. The national mean Surprise Index of 15.10 indicates that the typical district's cognitive automation potential exceeds its technology-sector exposure by approximately 15 percentage points.

2.2 Geographic Distribution

State and union territory-level aggregation reveals substantial regional variation in AI occupational exposure (Table 2). The highest Iceberg Index values appear in Chandigarh (56.43), Puducherry (52.42), Sikkim (50.44), Goa (49.97), and Kerala (49.01). These states and union territories share characteristics of high urbanization, service-sector concentration, and formal employment prevalence.

Table 2 State and Union Territory-Level Iceberg Index: Top and Bottom Five

State	Iceberg	Surface	Surprise	Districts
Highest Exposure				
Chandigarh	56.43	63.47	−7.04	1
Puducherry	52.42	64.05	−11.63	4
Sikkim	50.44	66.19	−15.75	2
Goa	49.97	68.05	−18.08	2
Kerala	49.01	58.42	−9.41	14
Lowest Exposure				
Bihar	29.91	14.57	15.34	38
Madhya Pradesh	33.83	17.09	16.74	50
Uttar Pradesh	34.35	16.31	18.04	71
Jharkhand	36.75	17.62	19.13	24
Assam	37.35	9.37	27.98	33

The lowest state-level exposure appears in Bihar (29.91), Madhya Pradesh (33.83), Uttar Pradesh (34.35), Jharkhand (36.75), and Assam (37.35). These states exhibit lower formal-sector employment rates, larger agricultural workforce shares, and less penetration of administrative and financial services. Notably, all five states show positive Surprise Index values, indicating that their cognitive exposure substantially exceeds visible technology adoption.

Regional aggregation weighted by employment reveals the following pattern: South India (42.06) exhibits the highest exposure, followed by West India (39.29), North India (37.28), and East India (36.66). The South-East differential of 5.40 percentage points reflects the concentration of IT services, financial services, and formal administrative employment in southern states.

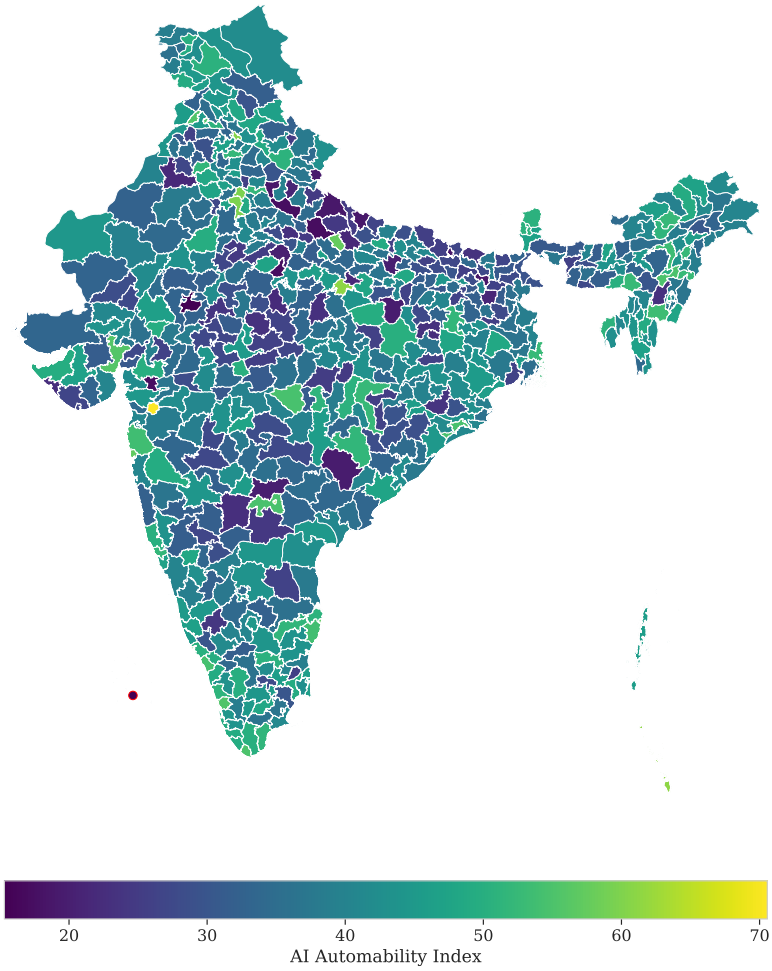

Fig. 1 Geographic distribution of AI Automability Index across India's 744 districts. The choropleth map displays district-level Iceberg Index values ranging from approximately 20 (dark purple, lowest exposure) to 70 (yellow, highest exposure). Districts with high exposure cluster in administrative centers and service-sector regions, while lower exposure appears in agriculturally dominated areas. The color scale represents the wage-weighted proportion of occupational tasks where current AI systems demonstrate technical capability. The visualization reveals that AI exposure extends nationwide rather than concentrating solely in traditional technology hubs.

2.3 Districts with Maximum Hidden Exposure

The Surprise Index identifies districts where AI occupational exposure extends far beyond visible technology-sector presence. Table 3 presents the ten districts with highest Surprise Index values—all exceeding 52 percentage points and all showing zero or negligible technology-sector employment.

Table 3 Districts with Highest Surprise Index (Hidden Exposure)

State	District	Iceberg	Surface	Surprise
Gujarat	The Dangs	70.56	0.00	70.56
Delhi	Central	67.72	0.00	67.72
Uttar Pradesh	Chitrakoot	61.19	0.00	61.19
Delhi	New Delhi	59.75	0.00	59.75
Nagaland	Kohima	54.82	0.00	54.82
Jammu & Kashmir	Ramban	54.43	0.00	54.43
Manipur	Churachandpur	53.45	0.00	53.45
Assam	Golaghat	53.14	0.00	53.14
Assam	Hojai	52.62	0.00	52.62
Delhi	South West	52.54	0.00	52.54

The Dangs district in Gujarat registers the highest Iceberg Index nationally (70.56) despite having no recorded technology-sector employment. This tribal-majority district, conventionally associated with agriculture and forestry, contains substantial formal administrative employment through government offices, banking correspondents, and educational institutions. The cognitive tasks performed by clerks, accountants, and administrative assistants in these establishments exhibit high AI exposure.

Chitrakoot in Uttar Pradesh, a district primarily known for pilgrimage tourism and rural heritage, registers an Iceberg Index of 61.19. The formal employment in this district concentrates in government administration and public banking, where clerical and data-processing tasks dominate the occupational structure.

Central Delhi and New Delhi districts show Iceberg values of 67.72 and 59.75 respectively, with zero Surface Index. These administrative districts house government ministries and public sector headquarters where employment concentrates in clerical, administrative, and coordination roles rather than software development.

The northeastern states contribute multiple high-Surprise districts (Kohima, Churachandpur, Golaghat, Hojai), reflecting workforce structures dominated by government employment and formal administrative services with limited private technology-sector presence.

2.4 Urban-Rural Decomposition

Contrary to the assumption that AI exposure concentrates in urban technology hubs, our analysis reveals higher mean exposure in rural areas. Among 633 districts with both urban and rural employment data, the mean rural Iceberg Index (49.52) substantially exceeds the urban mean (31.46)—a difference of 18.06 percentage points.

This counterintuitive finding reflects the composition of formal employment in rural India. While urban areas contain both high-exposure (financial services, administration) and low-exposure (retail, hospitality, informal services) occupations, rural formal employment concentrates in government offices, public banking (including banking correspondents under financial inclusion programs), primary education, and agricultural extension services. These occupations involve substantial text processing, record-keeping, and coordination tasks with high AI exposure scores.

The urban-rural pattern varies by state. In Uttar Pradesh, rural Iceberg Index (42.38) exceeds urban (31.97) in most districts, driven by formal government employment in district headquarters and tehsil offices. In contrast, Maharashtra shows more balanced urban-rural exposure, reflecting the broader occupational diversity in both sectors.

2.5 Industry Concentration Analysis

The Herfindahl-Hirschman Index (HHI) of exposure shares across NIC industry codes reveals how AI occupational exposure distributes within districts (Table 4). National mean HHI is 1645 (SD = 956), with values ranging from 372 (highly distributed) to 9573 (highly concentrated).

Table 4 Distribution of Industry Concentration Tiers

Concentration Tier	Districts	Percentage	HHI Range
Distributed	414	55.6%	< 1500
Moderate	234	31.5%	1500–2000
Concentrated	96	12.9%	> 2000

The majority of districts (55.6%) exhibit distributed exposure patterns, where AI-exposed employment spreads across multiple industries including education, public administration, financial services, and wholesale trade. These districts require multi-sector policy coordination rather than industry-specific intervention.

Concentrated districts (12.9%) show exposure dominated by one or two industries—typically government administration or financial services. Districts like Doda (HHI = 3539) and Faridkot (HHI = 3451) in this category can target reskilling and adaptation programs at specific sectors.

Moderate concentration (31.5%) characterizes districts where exposure spans several industries but with uneven distribution. Policy approaches for these districts may combine sector-specific and broad-based strategies.

2.6 Correlation with Socioeconomic Indicators

District-level Iceberg Index values correlate significantly with Census 2011 socioeconomic indicators (Table 5). Literacy rate shows the strongest positive correlation ($r = 0.50$), followed by internet access ($r = 0.49$) and urbanization rate ($r = 0.47$). Youth population share (ages 0–29) correlates negatively ($r = -0.35$).

These correlations align with the mechanism underlying AI exposure: districts with higher literacy and connectivity have larger formal service sectors where cognitive tasks concentrate. The negative youth correlation reflects the inverse relationship between demographic structure and economic development level in India's districts.

Importantly, these correlations describe cross-sectional associations rather than causal relationships. High-literacy districts face greater AI exposure because their

Table 5 Correlation Between Iceberg Index and Census Indicators

Indicator	r	Interpretation
Literacy Rate	+0.50	High literacy districts have more automatable knowledge work
Internet Access	+0.49	Digital connectivity correlates with formal service employment
Urbanization	+0.47	Urban service economies more exposed than rural agricultural
Youth % (0–29)	−0.35	Younger demographic districts tend to be less developed

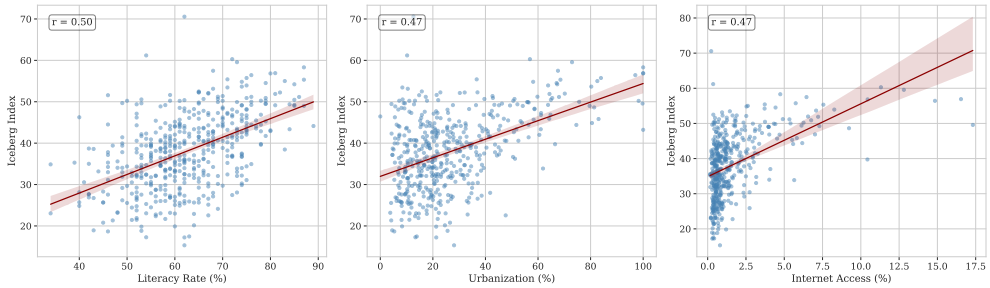

Fig. 2 Correlation between Iceberg Index and Census 2011 socioeconomic indicators across Indian districts. Each panel shows a scatter plot with linear regression trend line (dark red) and 95% confidence interval (shaded region). Left: Literacy rate shows moderate positive correlation ($r = 0.50$), indicating that districts with higher educational attainment have greater concentrations of automatable knowledge work. Center: Urbanization rate exhibits positive correlation ($r = 0.47$), reflecting the concentration of formal service-sector employment in urban areas. Right: Internet access demonstrates the strongest positive relationship ($r = 0.49$), as digital connectivity enables both the cognitive work that AI can automate and the infrastructure for AI deployment. $N = 463$ matched districts.

economies have developed toward knowledge work—the same pattern that previously drove economic advancement now creates automation vulnerability.

2.7 Validation with Real-World Patterns

Our methodology produces exposure estimates consistent with observed AI adoption patterns. Districts containing major IT hubs (Hyderabad, Bangalore Urban, Pune, Gurgaon) rank among the highest in Surface Index, confirming alignment between our framework and actual technology-sector presence.

The NCO-to-O*NET crosswalk achieves 100% coverage of 3,445 occupation codes through a tiered matching strategy: semantic keyword matching (29.5%), NCO hierarchical prefix matching (70.2%), and division-level fallback (0.3%). AIOE score assignment achieves 87.6% direct match with the Felten et al. database, with version reconciliation and prefix averaging covering the remainder.

The wage estimation procedure successfully imputes values for 94.7% of employed individuals in the PLFS sample, drawing on Current Weekly Status earnings where

available and household consumption expenditure otherwise. This approach addresses the substantial informal employment share while weighting formal employment appropriately in exposure calculations.

3 Methods

This study adapts the Iceberg Index methodology [2] to measure district-level AI workforce exposure in India. The analysis integrates four data sources through a multi-stage pipeline: (1) cross-national occupation mapping from India's NCO to the U.S. O*NET system, (2) AI automatability scoring using the Language Modeling AIOE framework [1], (3) workforce exposure calculation from PLFS microdata, and (4) validation against Census socioeconomic indicators. Figure 3 illustrates the complete data flow.

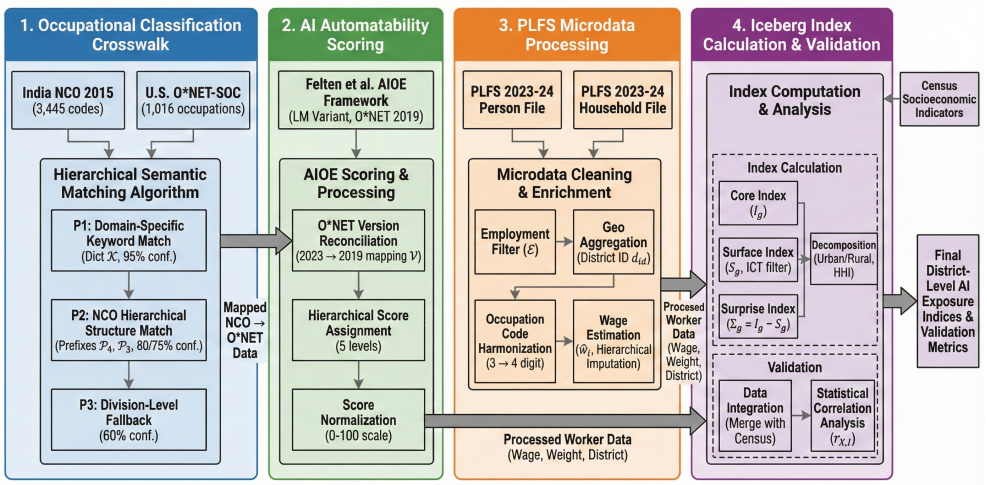

Fig. 3 Methodological pipeline for computing the India Iceberg Index. The framework integrates four components: (1) Occupational Classification Crosswalk mapping 3,445 India NCO-2015 codes to 1,016 U.S. O*NET-SOC occupations through hierarchical semantic matching; (2) AI Automatability Scoring using the Felten et al. AIOE framework with O*NET version reconciliation and hierarchical score assignment; (3) PLFS Microdata Processing including employment filtering, geographic aggregation, occupation code harmonization, and wage estimation; and (4) Iceberg Index Calculation & Validation computing core, surface, and surprise indices with decomposition by urban-rural sector and industry concentration (HHI), validated against Census socioeconomic indicators.

3.1 Occupational Classification Crosswalk

3.1.1 Data Sources

The primary classification systems are India's National Classification of Occupations (NCO) 2015, which provides 3,445 occupation codes structured hierarchically following

ISCO-08, and the U.S. Bureau of Labor Statistics O*NET-SOC system comprising 1,016 occupations with detailed skill and ability taxonomies.

3.1.2 Semantic Mapping Approach

Traditional fuzzy text matching produces semantically incorrect mappings (e.g., matching "mycologist" to "dentist" based on suffix similarity rather than to "biologist" based on occupational domain). We therefore implement a hierarchical semantic matching algorithm with three priority levels:

Priority 1: Domain-Specific Keyword Matching

We construct a dictionary $\mathcal{K}$ of approximately 200 domain-specific keywords mapped to O*NET codes. Keywords span legal professions, biological sciences, engineering specializations, healthcare, and India-specific occupations (e.g., Ayurveda practitioners, tabla makers). For a given NCO title t, keywords are matched in descending order of string length to ensure specific terms take precedence:

$$\text{match}_1(t) = \arg\max_{k \in \mathcal{K}} \{|k| : k \subseteq t_{\text{lower}}\} \tag{1}$$

where t_{lower} denotes the lowercase transformation of t and $|k|$ is keyword length. Matches at this level receive confidence score $s = 95$.

Priority 2: NCO Hierarchical Code Structure

When no keyword match exists, the algorithm exploits NCO's hierarchical structure. The first four digits of an NCO code identify the unit group (occupational family), while the first three digits identify the minor group. We maintain lookup tables $\mathcal{P}_4$ and $\mathcal{P}_3$ mapping these prefixes to appropriate O*NET codes:

$$\text{match}_2(c) = \begin{cases} \mathcal{P}_4[c_{1:4}] & \text{if } c_{1:4} \in \mathcal{P}_4, \quad s = 80 \\ \mathcal{P}_3[c_{1:3}] & \text{if } c_{1:3} \in \mathcal{P}_3, \quad s = 75 \end{cases} \tag{2}$$

where $c_{1:n}$ denotes the first n digits of NCO code c.

Priority 3: Division-Level Fallback

For unmatched occupations, the algorithm assigns default O*NET codes based on the NCO major division (first digit), ensuring complete coverage. These mappings receive confidence score $s = 60$.

3.1.3 Crosswalk Validation

The final crosswalk achieves 100% coverage across 3,445 NCO occupations with the following distribution: semantic keyword matches 29.5%, NCO prefix matches 70.2%, and division-level fallbacks 0.3%. The NCO 2015 to NCO 2004 concordance preserves 77.6% backward compatibility.

3.2 AI Automatability Scoring

3.2.1 Felten et al. Language Modeling AIOE Framework

We adopt the Artificial Intelligence Occupational Exposure (AIOE) scores from Felten et al. [**?**], specifically the Language Modeling variant calibrated to GPT-3/ChatGPT-era capabilities. This framework maps 10 AI application areas to 52 human abilities from O*NET via a crowd-sourced relatedness matrix $R \in \mathbb{R}^{10\times 52}$. For each O*NET occupation o with ability importance vector $\mathbf{a}_o \in \mathbb{R}^{52}$ and prevalence vector $\mathbf{p}_o \in \mathbb{R}^{52}$, the Language Modeling AIOE score is:

$$\text{AIOE}_{\text{LM}}(o) = \sum_{j=1}^{52} a_{o,j} \cdot p_{o,j} \cdot \left(\sum_{i=1}^{10} w_i \cdot R_{i,j} \right) \tag{3}$$

where w_i represents the weight assigned to AI application i for language modeling capabilities. The original scores range from -1.854 (least exposed) to $+1.926$ (most exposed) across 774 O*NET occupations.

3.2.2 O*NET Version Reconciliation

The Felten scores use O*NET 2019 codes, while our NCO crosswalk targets O*NET 2023 codes. We implement a version mapping table $\mathcal{V}$ for 97 restructured codes:

$$\text{code}_{2019} = \mathcal{V}[\text{code}_{2023}] \tag{4}$$

Key restructurings include computer systems analysts (15-1211.00 $\rightarrow$ 15-1121.00), software developers (15-1252.00 $\rightarrow$ 15-1132.00), and medical specializations (29-1216.00 $\rightarrow$ 29-1063.00 for internal medicine physicians).

3.2.3 Hierarchical Score Assignment

For each NCO occupation mapped to O*NET code c, automatability scores are assigned through a five-level hierarchy:

1. **Direct match**: Full O*NET code in Felten database (87.6% of records)
2. **Version mapping**: O*NET 2023 $\rightarrow$ 2019 code conversion (5.0%)
3. **Base SOC**: Match on 6-digit SOC without decimal suffix (2.6%)
4. **Prefix-5**: Mean score of 5-digit SOC prefix group (4.8%)
5. **Major group**: Mean score of 2-digit occupational category (0%)

3.2.4 Score Normalization

Raw AIOE scores are normalized to a 0–100 scale for interpretability:

$$\text{AI}_{\text{auto}}(o) = \frac{\text{AIOE}_{\text{LM}}(o) - \text{AIOE}_{\min}}{\text{AIOE}_{\max} - \text{AIOE}_{\min}} \times 100 \tag{5}$$

where $\text{AIOE}_{\min} = -1.854$ and $\text{AIOE}_{\max} = 1.926$. Exposure categories are defined as: Low (< 25), Medium-Low (25–50), Medium-High (50–75), and High (> 75).

The normalized distribution across 3,445 NCO occupations shows mean score 40.6 ($\sigma = 25.0$), with 39.1% Low, 28.5% Medium-Low, 19.3% Medium-High, and 13.0% High exposure categories.

3.3 PLFS Microdata Processing

3.3.1 Data Sources

The Periodic Labour Force Survey (PLFS) 2023–24 provides individual-level employment data through two files:

- **Person-level file** (cperv1.csv): 415,549 records with occupation codes, employment status, earnings, demographics, and geographic identifiers
- **Household-level file** (chhv1.csv): 101,957 records with monthly consumer expenditure and household attributes

3.3.2 Employment Filter

Workers are classified as employed based on Principal Activity Status codes:

$$\mathcal{E} = \{11, 12, 21, 31, 41, 51\} \tag{6}$$

representing self-employed (own account and employer), unpaid family helpers, regular wage/salary workers, and casual laborers.

3.3.3 Geographic Aggregation

Unique district identifiers combine state and district codes:

$$d_{\text{id}} = \text{zfill}_2(\text{State\_UT\_Code}) \parallel \text{zfill}_2(\text{District\_Code}) \tag{7}$$

where $\parallel$ denotes string concatenation and zfill_2 pads to 2 digits with leading zeros.

3.3.4 Occupation Code Harmonization

PLFS records 3-digit NCO division codes while our automatability dataset uses 4-digit codes. Aggregation proceeds as:

$$\bar{A}_{\text{div}} = \frac{1}{|\mathcal{O}_{\text{div}}|} \sum_{o \in \mathcal{O}_{\text{div}}} \text{AI}_{\text{auto}}(o) \tag{8}$$

where $\mathcal{O}_{\text{div}}$ is the set of 4-digit occupations within 3-digit division div.

3.3.5 Wage Estimation

India's large informal sector requires hierarchical wage imputation. For worker i, estimated monthly wage $\hat{w}_i$ follows priority order:

$$\hat{w}_i = \begin{cases} \text{CWS}_{\text{salaried},i} \times 4.33 & \text{if available} \\ \text{CWS}_{\text{self-emp},i} \times 4.33 & \text{elif available} \\ \sum_{d=1}^{7} \sum_{a=1}^{2} w_{i,d,a} \times 4.33 & \text{elif daily wages available} \\ \text{MCE}_h / \text{HH\_Size}_h & \text{otherwise} \end{cases} \tag{9}$$

where CWS denotes Current Weekly Status earnings, the factor 4.33 converts weekly to monthly values, $w_{i,d,a}$ represents daily wages for day d and activity a, and MCE_h is monthly consumer expenditure for household h.

3.4 Iceberg Index Calculation

3.4.1 Core Index Definition

Following the Project Iceberg methodology [2], the Iceberg Index for geographic unit g measures the wage-weighted proportion of occupational value where AI systems demonstrate technical capability:

$$I_g = \frac{\sum_{i \in g} \hat{w}_i \cdot m_i \cdot A_i}{\sum_{i \in g} \hat{w}_i \cdot m_i} \times 100 \tag{10}$$

where $\hat{w}_i$ is estimated monthly wage, m_i is the survey multiplier (sampling weight), and $A_i \in [0, 1]$ is the normalized automatability score for worker i's occupation.

3.4.2 Surface Index

The Surface Index captures visible technology-sector exposure, restricted to ICT occupation divisions:

$$S_g = \frac{\sum_{i \in g: o_i \in \mathcal{T}} \hat{w}_i \cdot m_i \cdot A_i}{\sum_{i \in g} \hat{w}_i \cdot m_i} \times 100 \tag{11}$$

where $\mathcal{T} = \{251, 252, 351, 352\}$ represents software developers, database/network professionals, ICT operations technicians, and telecommunications technicians.

3.4.3 Surprise Index

The Surprise Index quantifies hidden white-collar exposure beyond visible technology adoption:

$$\Sigma_g = I_g - S_g \tag{12}$$

High Σ_g values indicate districts with substantial cognitive automation potential in administrative, financial, and professional services that may not be recognized in technology-focused workforce planning.

3.4.4 Urban-Rural Decomposition

District-level indices are decomposed by sector using the PLFS sector indicator:

$$I_g^{\text{urban}} = \frac{\sum_{i \in g: \text{sector}_i = 1} \hat{w}_i \cdot m_i \cdot A_i}{\sum_{i \in g: \text{sector}_i = 1} \hat{w}_i \cdot m_i} \tag{13}$$

with analogous calculation for rural sectors ($\text{sector}_i = 2$).

3.4.5 Industry Concentration

To assess whether district exposure concentrates in few industries or distributes broadly, we compute the Herfindahl-Hirschman Index (HHI) of exposure shares across NIC industry codes j:

$$\text{HHI}_g = \sum_j \left(\frac{E_{g,j}}{\sum_j E_{g,j}} \right)^2 \times 10000 \tag{14}$$

where $E_{g,j} = \sum_{i \in g: \text{NIC}_i = j} \hat{w}_i \cdot m_i \cdot A_i$ is the total exposure value for industry j in district g.

3.5 Census Validation

3.5.1 Data Integration

District-level Iceberg Index values are merged with Census 2011 indicators (with 2024 projections where available) using normalized district names. The merge achieves 463 matched districts (67% of 694 total), providing robust sample size for correlation analysis.

3.5.2 Indicator Selection

We examine four socioeconomic indicators hypothesized to relate to AI exposure:

- **Literacy Rate**: Percentage of population age 7+ who are literate
- **Internet Access**: Percentage of households with internet connectivity
- **Urbanization Rate**: Percentage of population in urban areas
- **Youth Population**: Percentage of population age 0–29

3.5.3 Statistical Analysis

Pearson correlation coefficients quantify relationships between Iceberg Index and Census indicators:

$$r_{X,I} = \frac{\sum_g (X_g - \bar{X})(I_g - \bar{I})}{\sqrt{\sum_g (X_g - \bar{X})^2 \sum_g (I_g - \bar{I})^2}} \tag{15}$$

where X_g represents the Census indicator value and I_g the Iceberg Index for district g. Linear regression trend lines provide visual confirmation of relationships.

3.6 Implementation

All analyses are implemented in Python 3.10 using pandas for data manipulation, numpy for numerical computation, and matplotlib/seaborn for visualization. The complete pipeline executes in approximately 15 minutes on a standard workstation. Code and intermediate datasets are available at [repository URL].

3.6.1 Computational Pipeline

The analysis proceeds through five stages:

1. **Data Loading**: Parse PLFS microdata, automatability scores, and district reference files

2. **Data Preparation**: Filter employed workers, create identifiers, aggregate automatability to 3-digit NCO divisions
3. **Wage Estimation**: Apply hierarchical imputation, merge household consumption data
4. **Index Calculation**: Compute Iceberg, Surface, Surprise indices with urban-rural decomposition and industry concentration
5. **Aggregation & Output**: Generate district and state-level summaries, merge Census indicators

Table 6 NCO to O*NET Crosswalk Match Quality

Match Type	Count	Percentage	Confidence Score
Semantic keyword	1,016	29.5%	95
NCO prefix (4-digit)	2,417	70.2%	80
Division fallback	12	0.3%	60
Total	3,445	100.0%	—

Table 7 AIOE Score to O*NET Code Match Types

Strategy	Count	Percentage
Direct match	3,019	87.6%
Version mapping (2023→2019)	172	5.0%
Base SOC match	90	2.6%
Prefix-5 average	164	4.8%
Total	3,445	100.0%

Table 8 AI Automatability Score Distribution

Category	Score Range	Count	Percentage
Low	0–25	1,349	39.1%
Medium-Low	25–50	982	28.5%
Medium-High	50–75	666	19.3%
High	75–100	448	13.0%
Mean (SD)			40.6 (25.0)

4 Discussion

Our analysis reveals that AI occupational exposure in India extends substantially beyond the technology sector that dominates public discourse. The 41.5% wage-weighted national exposure, combined with technology-sector employment of less than 1%, indicates that the "submerged mass" of the iceberg—cognitive exposure in administrative, financial, and clerical occupations—constitutes the dominant share of automation potential.

The geographic distribution of exposure challenges conventional assumptions about AI's reach. Districts like The Dangs, Chitrakoot, and Ramban—conventionally associated with agriculture, pilgrimage, and rural heritage—register among the highest Iceberg Index values nationally. Their exposure derives not from technology-sector employment but from formal administrative work in government offices, banking institutions, and educational establishments. The cognitive tasks performed by clerks, data entry operators, and administrative assistants in these settings exhibit high alignment with current AI capabilities.

The counterintuitive urban-rural pattern merits particular attention. Rural formal employment in India concentrates in occupations with high AI exposure: government clerks, banking correspondents, primary school teachers, and agricultural extension workers. Urban areas, by contrast, contain both high-exposure (financial services) and low-exposure (retail, hospitality) occupations. This pattern suggests that AI's impact on rural India may arrive through formal administrative channels rather than agricultural automation.

The positive Surprise Index in lower-income states (Bihar, Uttar Pradesh, Madhya Pradesh, Jharkhand, Assam) indicates a structural mismatch between preparation and exposure. These states show minimal technology-sector presence yet substantial cognitive exposure in government and public sector employment. Workforce planning focused on visible technology disruption may substantially underestimate transformation potential in these regions.

Our findings carry implications for India's service-sector development model. The entry-level cognitive jobs that have served as mobility pathways—data entry clerks, junior accountants, administrative assistants—represent precisely the tasks where AI demonstrates strongest capability. If these positions contract without corresponding growth in complementary roles, the "first rung" of the formal employment ladder may weaken.

Several limitations bound our analysis. The PLFS occupation codes aggregate to 3-digit NCO divisions, limiting within-division variation capture. The wage imputation procedure, while necessary given India's informal employment structure, introduces measurement error. The AIOE scores derive from US occupational characteristics and may not fully reflect Indian workplace contexts. Finally, the index measures technical capability overlap rather than adoption likelihood or displacement outcomes.

5 Conclusion

The India Iceberg Index provides the first comprehensive district-level quantification of AI occupational exposure for India's labor market. Our analysis of 744 districts reveals

that approximately 41.5% of wage-weighted employment faces technical exposure to current AI systems, with the "hidden" cognitive exposure in administrative and clerical occupations substantially exceeding visible technology-sector disruption.

The findings suggest that framing AI as a "tech problem" fundamentally mischaracterizes its scope. Districts without technology-sector presence register among the highest exposure values, driven by formal administrative employment in government, banking, and education. Rural areas show higher mean exposure than urban areas, reflecting the concentration of formal cognitive work in non-metropolitan settings.

Policy implications flow from the industry concentration analysis. The majority of districts (55.6%) exhibit distributed exposure patterns requiring coordinated multi-sector intervention rather than targeted industry programs. States with positive Surprise Index values face particular urgency in recognizing cognitive exposure that current workforce planning may overlook.

Future work should extend this analysis temporally to track exposure evolution, incorporate firm-level adoption data to connect capability to implementation, and examine occupation-specific reskilling pathways. The India Iceberg Index provides a baseline for such investigations and a framework for evidence-based workforce policy in the AI era.

Data Availability. The complete dataset, analysis code, and interactive visualization are publicly available at https://github.com/pikulsomesh/india-iceberg-index. The Streamlit application provides district-level exploration at https://india-iceberg-index.streamlit.app/.

Acknowledgements. The author thanks the Project Iceberg team at MIT for developing the methodological framework.

Declarations

- **Funding**: Not applicable.
- **Conflict of interest**: The author declares no competing interests.
- **Ethics approval**: Not applicable (secondary analysis of publicly available survey data).
- **Consent for publication**: Not applicable.
- **Code availability**: Available at https://github.com/pikulsomesh/india-iceberg-index.
- **Author contribution**: S.M. conceived the study, developed the methodology, conducted the analysis, and wrote the manuscript.

References

[1] Felten, E.W., Raj, M., Seamans, R.: How will language modelers like ChatGPT affect occupations and industries? arXiv preprint arXiv:2303.01157 (2023)

[2] Chopra, A., Bhattacharya, S., Salvador, D., et al.: The Iceberg Index: Measuring skills-centered exposure in the AI economy. arXiv preprint arXiv:2510.25137v2 (2025)

[3] Rodrik, D.: Premature deindustrialization. Journal of Economic Growth **21**(1), 1–33 (2016)

[4] NASSCOM: Technology Sector in India 2024: Strategic Review. National Association of Software and Service Companies, New Delhi (2024)

[5] National Sample Survey Office: Periodic Labour Force Survey (PLFS) Annual Report (July 2023–June 2024). Ministry of Statistics and Programme Implementation, Government of India (2024)

[6] Ministry of Labour and Employment: National Classification of Occupations (NCO-2015). Government of India (2015)

[7] National Center for O*NET Development: O*NET OnLine. U.S. Department of Labor, Employment and Training Administration (2024). https://www.onetonline.org/

[8] Office of the Registrar General & Census Commissioner: Census of India 2011. Ministry of Home Affairs, Government of India (2011)

[9] Azar, J.A., Marinescu, I., Steinbaum, M.I., Taska, B.: Concentration in US labor markets: Evidence from online vacancy data. Labour Economics **66**, 101886 (2020)