

Technical data and construction note

Fatih Kansoy

9 September 2026

Abstract

This note documents the construction, keys, denominators, classifications, missingness rules and public files for AI Use and Occupations. The release combines pinned country-level platform records, a fixed *ONET task catalogue*, *alternative ONET–ESCO–ISCO* and UK SOC routes, and official employment context. It contains 121 countries and areas with direct occupation observations in May 2026 and 114 in April. Direct source shares remain distinct from semantic donor means and native employment joins. The note is a file-level reference for reproduction and reuse; substantive results and their relation to the literature are developed in the companion paper.

Keywords: data documentation; occupational classification; O*NET; ESCO; ISCO; SOC; reproducibility; missing data

1. Release identity and scope

The research release is `1.1.0`. Its primary platform source is the Anthropic Economic Index revision `2ea58ff75e4247d26810c37f10c179edc2466cac`, released on 26 June 2026. The source commit, release directories, retrieval times and hashes are recorded in the project manifest. The fixed occupation task catalogue is O*NET 30.2.

The public catalogue contains 27 curated evidence tables. Each entry records the source path, public download, format, row count, primary key, grain, units, schema, file size and SHA-256 hash. A catalogue preview is not a sample file: each linked download contains the complete frozen public table.

The observations concern classified consumer platform activity. The source assigns conversations to occupational activities; it does not observe the user's job. The first-party API source has no country breakdown and is not allocated to countries in this release.

2. Observation windows

The archive contains five country observation windows drawn from four selected releases.

Observation window	Source release	Published content used here
4–11 August 2025	15 September 2025	One-week task observations; 113 geographies
13–20 November 2025	15 January 2026	One-week task observations; 116 geographies
5–12 February 2026	24 March 2026	One-week task observations; 117 geographies

Observation window	Source release	Published content used here
1 April–1 May 2026	26 June 2026	Monthly country and direct occupation observations; 114 geographies
1 May–1 June 2026	26 June 2026	Monthly country and direct occupation observations; 121 geographies

The June source documents the monthly end dates as exclusive. Earlier files retain the publisher's displayed weekly date fields. Sampling, classifiers and taxonomy differ across releases, so the construction does not fill missing months or treat the five windows as a uniform panel.

The [Source country period inventory](#) table contains the source availability by geography and period. The [Country coverage](#) table records derived task support for the same 581 geography-period observations.

3. Geography registry and coverage universes

The registry has 250 entries: 248 countries and areas extracted from UNSD M49, plus Kosovo and Taiwan added as explicit source geographies. Geography keys use ISO alpha-3 style identifiers where available. Publisher labels that are not geographic, including `NONE` and `not_classified`, do not become registry entries.

Coverage counts refer to distinct universes:

Count	Definition
250	Registry entries
180	Recognised geographies with any country row in the four pinned releases
128	Geographies with a positive task cell in at least one archived window
121	Geographies with May direct occupation rows
114	Geographies with April direct occupation rows
182	Registry entries with a selected official employment series
119	Historical task geographies with an employment series
113 / 106	May / April occupation geographies with an employment series

The May roster includes 118 UN member states, the State of Palestine, Puerto Rico and Taiwan. The categories describe the statistical source and are not used to infer political status.

The [Country registry](#) table is the geography spine. The [Global availability](#) table summarizes source and derived coverage. The [Labour availability](#) table records labour-statistics availability separately.

4. Source selection and row keys

The primary country values select the consumer surface and retain the publisher's date, geography, facet, hierarchy, metric and category fields. Direct occupation evidence uses the `soc_occupation` facet at hierarchy level 0 with metric `pct`. Source labels, codes and displayed precision are retained before any mapping.

At source grain, a row is identified by the available combination of release, surface, observation dates, geography, facet, hierarchy, metric and category. Derived tables use narrower documented keys. Key checks reject duplicates rather than choosing an arbitrary row.

The core public tables have these grains:

Table	Grain	Principal identifiers
Country usage	Geography-month	geography, start date
Direct occupation source	Geography-month-source occupation	geography, start date, O*NET occupation
O*NET occupation panel	Geography-month-O*NET occupation	country, date_start, onet_code
ESCO occupation panel	Geography-month-route-ESCO occupation-ISCO4 link	country, date_start, mapping_route, esco_uri, isco4
ISCO occupation panel	Geography-month-route-ISCO group	country, date_start, mapping_route, occupation_level, occupation_code
US native soc panel	Observation-window-native SOC occupation	date_start, occ_code

The O*NET, ESCO and ISCO occupation panels above contain monthly observations. The native US panel also retains earlier weekly task descriptors; its `period` field is the BLS employment year, while `date_start` identifies the platform observation window. Measures are columns in these wide tables, not components of their row keys. The data catalogue is authoritative for the exact primary key of each released table.

5. Missingness and source precision

Three publication states remain distinguishable:

State	Source evidence	Stored meaning
Published positive	Row exists with a positive displayed share	Numeric source value
Published rounded zero	Row exists and displays zero	Numeric zero plus published status
Not published	No row exists for an otherwise supported geography-period-category	Missing raw share; actual use unknown

An absent category contributes zero to the sum of *published* mass but is not assigned zero actual use. The residual outside the published rows may combine unpublished activity, unclassified activity and rounding. The public source does not identify their separate contributions.

Source percentages are released to two decimal places. A reported value has a row-level rounding half-width of at most 0.005 percentage points. Summed rounding ranges accumulate these arithmetic bounds over numeric rows. They are not sampling intervals and do not account for classification or disclosure uncertainty.

6. Direct occupation measures

For geography c , month m and published occupation o , the direct share is

$$S_{cmo} = \frac{\text{pct}_{cmo}}{100}.$$

The denominator is total consumer usage for the geography in that observation window. Retained occupation mass is

$$M_{cm} = \sum_{o \in \mathcal{P}_{cm}} S_{cmo},$$

where $\mathcal{P}_{cm}$ is the set of published occupation rows. The source shares are summed before mapping and are never renormalized to the released categories.

The [Direct occupation source](#) download preserves only published source rows. The larger [O*NET occupation panel](#) expands to the fixed occupation universe and carries publication-state fields so that absent rows can be distinguished from rounded zeros.

7. Task catalogue measures and historical bridge

For fixed task set T_o , let s_{cmt} be the source task percentage divided by 100 and let $\mathcal{P}_{cm}^{\text{task}}$ be the set of published task rows. Task coverage is

$$C_{cmo} = \frac{\sum_{t \in T_o} \mathbf{1}\{t \text{ has a positive published row in } (c, m)\}}{|T_o|},$$

and task-share intensity is

$$I_{cmo} = \frac{\sum_{t \in T_o \cap \mathcal{P}_{cm}^{\text{task}}} s_{cmt}}{|T_o|}.$$

All catalogue tasks remain in the denominator. A task may be listed under several occupations, so these measures are not additive across occupations.

Historical task text first follows a legacy-ID bridge into the current catalogue. Only unmatched source cells may use exact normalized current-task text as a fallback. The procedure recovers 1,839 source cells, with zero observed convergent collisions in the frozen audit. Strictly matched rows are not expanded using a union of every text candidate. Changed text under a retained ID and identical text attached to multiple occupation-specific tasks remain flagged.

Source-share floors of 0.01, 0.05 and 0.10 percentage points are supplementary sensitivity checks. They remove small positive cells while holding the catalogue denominator and mapping route fixed. They do not reconstruct the publisher's count threshold.

8. European mapping construction

The typed O*NET–ESCO library contains 4,253 relations, including 498 exact links. Routes retain exact, exact-plus-narrow and all-typed relations separately; an 8,627-link untyped alternative supplies a sensitivity route.

For route a and ESCO occupation e , with available donor set $D_e^{(a)}$, the descriptor is

$$\bar{S}_{cme}^{(a)} = \frac{1}{|D_e^{(a)}|} \sum_{o \in D_e^{(a)}} S_{cmo}^{\text{published}}.$$

Here $S_{cmo}^{\text{published}}$ equals the source share when a row exists and zero published mass for an absent row within a supported geography-month. The donor set includes all linked O*NET catalogue occupations with a defined published-mass value, including donors with no published row. Donors outside the available catalogue are excluded. This convention measures published mass; actual use for an unpublished donor remains unknown.

The same donor-mean operation is applied to task descriptors. A missing donor set remains missing. ISCO4 values are equal means across represented ESCO occupations. ISCO1 and ISCO2 values are direct means over represented ISCO4 groups, avoiding a nested average that would assign implicit extra weight to small branches.

Mapped values are semantic donor means and are labelled as such. They do not conserve source mass and cannot be interpreted as allocated national shares. Alternative routes describe mapping sensitivity rather than sampling uncertainty. The [ESCO occupation panel](#) and [ISCO occupation panel](#) files retain route, donor count and support fields.

9. Employment joins and conditional summaries

For ISCO sources, the pipeline selects the finest eligible detail while retaining the actual year, sex and age scope, survey population, ICLS definition and publication flags. Finer observations can be older than an available coarse series. No source is silently relabelled to a common population or year.

For an available mapped score x_g and employment E_g , a conditional weighted mean is

$$\bar{x}^E = \frac{\sum_{g \in \mathcal{G}_{\text{matched}}} E_g x_g}{\sum_{g \in \mathcal{G}_{\text{matched}}} E_g}.$$

The matched-employment fraction is

$$F_E = \frac{\sum_{g \in \mathcal{G}_{\text{matched}}} E_g}{E_{\text{official total}}}.$$

Different routes may have different matched denominators. Weighting a donor mean does not convert it into a national conversation-share allocation.

The United Kingdom uses APS employment and an open ONS SOC 2020 coding-index route. Counts of coding-index job titles supply lexical weights; they are not observed transition probabilities between SOC and ISCO. Earnings sensitivities use Eurostat SES 2022 and UK ASHE only where compatible values are available. Combining earnings and employment from different source populations does not produce a measured wage bill.

10. Native US construction

O*NET children are aggregated to their native six-digit SOC code. Direct occupation shares are summed at source integer precision, while task coverage and intensity are averaged within SOC code. The BLS National Employment Matrix then joins on exact detailed codes.

The Matrix contains 831 line items and an official 2025 total of 170,280,800 jobs. Exact matching supports 772 detailed groups and 92.53% of the total. Aggregate and unmatched BLS codes are not split through a synthetic concordance. The Matrix population counts jobs, including unincorporated self-employment.

For matching occupation o , relative concentration is

$$R_{cmo} = \frac{S_{cmo}}{E_o / E_{\text{national}}}.$$

The national denominator retains unmatched employment. The ratio is a descriptive comparison of shares, not a probability of worker use. Construction inputs and outputs are in [US bls employment](#), [US native soc panel](#) and [US native weighted summary](#).

11. Exposure proxy and acceptance gate

For task t , let C_t^{work} be work-related consumer counts, A_t global API counts, q_t consumer automation share and β_t capability. The published construction begins with

$$W_t = C_t^{\text{work}} + A_t$$

and, when $W_t > 0$,

$$\alpha_t = 0.5 + 0.5 \left(\frac{C_t^{\text{work}} q_t + A_t}{W_t} \right).$$

The gated task value is

$$r_t = \mathbf{1}\{W_t \geq 100\} \mathbf{1}\{\beta_t \geq 0.5\} \alpha_t,$$

with $r_t = 0$ when $W_t = 0$, without evaluating the zero denominator. For occupation-specific task-time weights w_{ot} , exposure is

$$X_o = \frac{\sum_{t \in T_o} w_{ot} r_t}{\sum_{t \in T_o} w_{ot}}.$$

Both sums range over the occupation's full task set. Tasks that fail either gate contribute zero to the numerator but remain in the time-weight denominator; weights are not renormalized over eligible tasks.

The frozen independent proxy reaches Spearman 0.89465 on 756 occupations with 6 common top-ten occupations. The prespecified numerical gate requires correlation at least 0.95 and at least 8 common top-ten occupations, so the proxy fails. Equal aggregation of already-published task scores reaches 0.87154 but is a downstream diagnostic, since those inputs embed publisher outcomes.

Exact reconstruction is also blocked by unavailable original time fractions, work-share imputation, similar-task grouping, employment allocation and uncensored intermediates. The failed local-model time weights are excluded from substantive results. Their preservation in the audit archive permits deterministic checking of why they failed; it does not turn them into valid estimates.

12. Validation and comparison tables

Comparison outputs retain pairwise common support and the direction of each source measure. They do not flip signs or select coefficients according to favourability.

Purpose	Summary table	Underlying observations
Publication breadth and usage volume	Volume regression	Country usage, Country coverage
Task versus occupation retained mass	Facet comparison	Direct and task source panels
Native US exposure comparison	US measure validation	US validation pairs
European mapping routes	Mapping correlation distribution	Occupation mapping correlations
Theoretical exposure measures	Comparator distribution	Occupation comparator correlations
PIAAC work activities	PIAAC across occupation summary	PIAAC detailed country results

PIAAC public activity cells require at least 30 valid respondents. Survey weights enter activity shares, while the released rank correlations use occupations as their comparison units. Restricted respondent records are not included.

13. Reporting grades

Grades combine observable publication and employment components: positive task-cell breadth, retained occupation mass, two monthly occupation observations, employment year, classification detail and matched-employment fraction. The rules yield 29 grade A, 11 grade B, 81 grade C and 129 grade D registry entries.

The grades were adopted after inspecting the public data. Their cutoffs are practical reporting thresholds rather than preregistered sample rules. A grade does not estimate representativeness, classifier accuracy or latent data quality. The exact components are released in [Geography support](#) and should be used instead of the letter alone when selecting a sample.

14. Integrity checks and provenance

The build checks unique keys, finite values, permissible ranges, publication states, aggregation denominators and source-to-public row counts. Numerical table generation, source licensing, PDF layout and scientific validation are separate checks with separate receipts.

File hashes identify the precise public bytes. Source manifests identify upstream URLs, commits and retrieval times. A hash establishes identity, not scientific validity. When source-derived files are updated, the catalogue, data dictionary and validation receipts must be regenerated together.

Restricted mapping vectors, comparator vectors and respondent-level PIAAC records remain outside the public package. Public aggregate files carry component-specific attribution and redistribution terms. The catalogue licence field links those terms rather than asserting a single licence for the combined archive.

15. Recommended entry points

Readers reproducing headline results should begin with [Facet comparison](#), [Volume regression](#) and [Geography support](#). Classification analysis begins with the O*NET, ESCO and ISCO panels. Native US work begins with the US SOC panel and BLS employment file. Validation work should use both the summary file and the underlying pair-level or geography-level observations.

The [data catalogue](#) supplies schemas and downloads for all 27 tables. The companion [paper](#) presents substantive results and scholarly interpretation. Shared terminology is defined in the [glossary](#), and all cited sources are collected in the [bibliography](#).

References

This bibliography is shared by the paper, technical data note and methodological pages.

- Anthropic. (2026). *Anthropic Economic Index report: Cadences*. 26 June.
- Anthropic. (2026). *EconomicIndex dataset*, revision `2ea58ff75e4247d26810c37f10c179edc2466cac`.
- Anthropic. (2026). *Labor market impacts of AI: A new measure and early evidence*.
- Appel, R., McCrory, P., Tamkin, A., McCain, M., Neylon, T., & Stern, M. (2025). *Anthropic Economic Index report: Uneven geographic and enterprise AI adoption*. arXiv:2511.15080.
- Bank Underground. (2026). *Canaries in the column: AI exposure and the UK's hiring slowdown*. 6 August.
- Bick, A., Blandin, A., & Deming, D. J. (2024; revised 2025). *The Rapid Adoption of Generative AI*. NBER Working Paper 32966.
- Brynjolfsson, E., Li, D., & Raymond, L. (2025). *Generative AI at Work*. *Quarterly Journal of Economics*, 140(2), 889–942.
- Centre for British Progress. (2026). *AI and the UK labour market: the evidence so far*.
- Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). *How People Use ChatGPT*. NBER Working Paper 34255.
- 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*. arXiv:2303.10130.
- European Commission. *Crosswalk between ESCO and ONET**.
- Eurostat. *EU Labour Force Survey employment by occupation (lfsa_ega2d)* and *Structure of Earnings Survey 2022*.
- Fan, R. Y., & Nguyen, H. M. (2026). *Aggregate Gains from AI and Their Distribution: Global Evidence from Usage Data*. IMF Working Paper 2026/147.

- Felten, E., Raj, M., & Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*.
- Felten, E., Raj, M., & Seamans, R. (2023). *How will Language Modelers like ChatGPT Affect Occupations and Industries?*. arXiv:2303.01157.
- Handa, K., et al. (2025). *Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations*. arXiv:2503.04761.
- Humlum, A., & Vestergaard, E. (2025). *Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI*. NBER Working Paper 33777.
- International Labour Organization. (2025). *Generative AI and Jobs: A Refined Global Index of Occupational Exposure*. ILO Working Paper 140.
- ILOSTAT. *Annual employment by sex and occupation, bulk dissemination facility*.
- OECD. *Survey of Adult Skills (PIAAC), Cycle 2 database*.
- OECD. *The OECD AI exposure measure*.
- Office for National Statistics. *SOC2020 Volume 2: coding rules and conventions*.
- O*NET Resource Center. *ONET Database 30.2 and archived releases**.
- Steele, J. L., & Cruz, I. (2026). *Helping People Choose Careers in the Age of AI*. arXiv:2607.15506.
- Tomlinson, K., Jaffe, S., Wang, W., Counts, S., & Suri, S. (2025). *Working with AI: Measuring the Applicability of Generative AI to Occupations*. arXiv:2507.07935.
- US Bureau of Labor Statistics. *Employment Projections Data Definitions*.
- US Bureau of Labor Statistics. (2026). *Table 1.2: Occupational projections, 2025–35, and worker characteristics, 2025*.