Supplementary Information for: Digitised workflows and a workforce segment overlooked by conventional AI exposure measures

Contents

- LLM scoring providers

Supplementary References

Appendix A. Scoring prompts

A.1 Original scoring prompt (GPT-4.1-mini, NB02) We used the following prompt to score all 18,796 O*NET tasks, with `{occupation}` and `{task}` populated at runtime, and then sent the prompt as a user message (temperature = 0.1, max_tokens = 100). We scored each task three times independently and took the majority vote as the final score.

> You are an expert in AI capabilities and occupational task analysis.
>
> Given the following occupational task, score it on 5 dimensions related to whether an autonomous AI agent with computer-use capabilities (able to see the screen, control mouse/keyboard, browse the web, manage files, and execute multi-step digital workflows) could perform this task end-to-end without human intervention.
>
> **Occupation:** {occupation} **Task:** {task}
>
> Score each dimension from 0 (lowest) to 5 (highest):
>
> D1. Digital Interface Dependency 0 = task is entirely physical/in-person with no digital component 5 = task occurs entirely within digital/screen-based environments
>
> D2. GUI Executability 0 = impossible for a computer-use agent to perform via GUI interaction 5 = fully executable end-to-end through screen interaction (clicking, typing, browsing, file management)
>
> D3. Judgment & Accountability Requirement (REVERSE SCORED) 0 = requires heavy subjective judgment, ethical decisions, or legally accountable sign-off 5 = minimal judgment needed; task has clear, verifiable success criteria
>
> D4. Workflow Composability 0 = task is holistic, unstructured, and cannot be broken into discrete steps 5 = task can be fully decomposed into a sequence of discrete, verifiable sub-steps
>
> D5. Physical-World Requirement (REVERSE SCORED) 0 = requires physical manipulation, in-person presence, or sensory input beyond a screen 5 = no physical-world component; entirely achievable through digital means
>
> Respond ONLY with a JSON object in this exact format (no explanation, no markdown): {"D1": X, "D2": X, "D3": X, "D4": X, "D5": X}
>
> Where X is an integer from 0 to 5.

A.2 Reworded scoring prompt (prompt sensitivity analysis, NB06) To assess sensitivity to prompt wording and message packaging, we applied a reworded prompt to a random sample of 100 tasks. We renamed dimension labels using synonymous phrasing, paraphrased anchor descriptions, and placed the reworded rubric in a system message with the task and occupation provided separately as the user message.

> You are an expert evaluating whether an AI agent operating a computer could autonomously perform a given work task.
>
> For the task described below, rate it on five aspects using a 0 to 5 integer scale.
>
> Aspect 1 — Digital System Presence: 0 means the task has zero connection to any computer or digital system. 5 means the task is conducted entirely within digital platforms or cloud services.
>
> Aspect 2 — Screen-Based Workflow Feasibility: 0 means there is no way to accomplish this through clicking, typing, or navigating a screen interface. 5 means every step can be completed through standard screen-based interaction.

Aspect 3 — Judgment Independence: 0 means the task demands extensive subjective reasoning, ethical weighing, or legally accountable decisions. 5 means the task follows clear-cut rules with objectively verifiable outcomes.

Aspect 4 — Step Decomposability: 0 means the task is a single indivisible action that cannot be broken into substeps. 5 means the task naturally splits into a clear sequence of discrete executable steps.

Aspect 5 — Freedom from Physical Presence: 0 means the task absolutely requires a human body to be physically present. 5 means the task can be done entirely remotely with no physical interaction.

Respond with ONLY a JSON object: {"D1": <int>, "D2": <int>, "D3": <int>, "D4": <int>, "D5": <int>}

Table S1. Exploratory factor analysis loadings (occupation level, N = 923)

Dimension	1-Factor loading	2-Factor: Factor 1	2-Factor: Factor 2
D1 Digital Interface	−0.975	0.992	−0.158
D2 GUI Executability	−0.980	0.976	−0.003
D3 Judgement (reverse-coded)	0.058	−0.086	0.782
D4 Composability	−0.817	0.813	0.415
D5 Physical (reverse-coded)	−0.953	0.953	−0.061

Note. Maximum likelihood estimation. 2-Factor model uses oblimin rotation. KMO = 0.759. Bartlett's chi² = 122,247.2 ($p < 0.001$). 1-Factor variance explained = 69.8%. 2-Factor cumulative variance explained = 86.5%. Factor correlation (2-Factor) = 0.087. D3 Judgement loads exclusively on Factor 2; this separates it from the executability cluster (D1, D2, D4, D5).

Table S2. Confirmatory factor analysis model comparison (occupation level, N = 923)

Model	CFI	RMSEA	Chi²	DoF
1-Factor (all 5 dimensions)	0.876	0.402	15,223.7	5
2-Factor (Tech Exec + Judgement)	0.876	0.450	15,223.7	4
2-Factor with cross-loading (D4 → both)	0.953	0.319	5,757.8	3

Note. CFA estimation used semopy on importance-weighted occupation-level scores. Tech Exec factor: D1, D2, D4, D5. Judgement factor: D3. The 2-Factor cross-loading model allows D4 to load on both factors. The 4-dimension model (D1, D2, D4, D5; reported in main text) yields CFI = 0.966 and RMSEA = 0.371 (DoF = 2). Very low degrees of freedom inflate RMSEA values; in near-saturated models, CFI provides a more reliable fit signal (Kenny et al 2015).

Table S3. D4 controlled incremental validity — full specification (M4: AIOE + D1 + D4)

Dependent variable	M3: R^2(AIOE + D1)	M4: (D4)	SE	p	ΔR^2(D4)	VIF(D4)
Job Zone	0.510	−1.240	0.083	< 0.001	0.135	31.7
Cognitive Demand	0.361	−0.541	0.048	< 0.001	0.114	31.6
Physical Demand	0.770	0.132	0.052	0.011	0.002	31.6
Routine Cognitive	0.328	−0.056	0.063	0.370	0.001	31.6
Computer Use	0.665	0.166	0.066	0.012	0.003	31.6

Note. M3: OLS regression DV ~ AIOE + D1. M4: OLS regression DV ~ AIOE + D1 + D4. All models use cluster-robust standard errors at the 6-digit SOC level (683 clusters). Two-tailed tests, = 0.05. D4 is significant for 4 of 5 DVs in M4 (Job Zone, Cognitive Demand, Physical Demand, Computer Use), with Routine Cognitive non-significant ($p = 0.370$). D4 sign-consistent on all 5 DVs (5/5). Three correlated regressors inflate VIF to 31.6–31.7; M4 read off across-DV significance pattern, single coefficients non-diagnostic at this collinearity.

Table S4. Heterotrait-monotrait (HTMT) discriminant validity matrix (4-dimension set)

	D1 Digital	D2 GUI	D4 Composability	D5 Physical (reverse-coded)
D1 Digital	—	1.047	0.907	1.033
D2 GUI	1.047	—	0.967	1.039
D4 Composability	0.907	0.967	—	0.922
D5 Physical (reverse-coded)	1.033	1.039	0.922	—

Note. HTMT ratios on importance-weighted occupation-level scores (N = 923). Threshold: > 0.85 = failed discriminant validity. All pairwise ratios > 0.85 here. Shared variance at this magnitude reads as formative; the D1 headline + D4 modifier architecture follows, with a flat composite ruled out.

Table S5. Variance inflation factor diagnostics (4-dimension formative model)

Dimension	VIF	Tolerance	Status
D1 Digital Interface	53.4	0.019	Problematic
D2 GUI Executability	46.0	0.022	Problematic
D4 Composability	4.3	0.231	Acceptable
D5 Physical (reverse-coded)	23.9	0.042	Problematic

Note. VIF from OLS regression of each dimension on the remaining three. Cut-off: > 10 = problematic multicollinearity. D1, D2, D5 cluster on a shared "digital presence" factor; D4 alone carries non-redundant information (VIF = 4.3). Architecture roles: D1 absorbs the shared variance as headline, D4 is the companion modifier, D2/D5 stay in the appendix as diagnostic subcomponents.

Table S6. Prompt reword sensitivity analysis (N = 100 tasks)

Dimension	Pearson r	ICC	MAD
D1 Digital Interface	0.877	0.848	0.61
D2 GUI Executability	0.895	0.893	0.51
D3 Judgement (reverse-coded)	0.576	0.546	0.70
D4 Composability	0.782	0.740	0.50
D5 Physical (reverse-coded)	0.910	0.904	0.43

Note. Per-dimension ICCs are ICC(2,1) — single-rater, two-way random effects, absolute agreement (McGraw–Wong notation; equivalent to the `ICC(A,1)` label in the `pingouin` package used for computation). MAD = mean absolute deviation between original and reworded scores. *Overall composite ICC = 0.918 (95 % BCa CI [0.887, 0.943], N = 100 tasks × 2 raters; see Table S17)*, computed as the single-rater ICC(A,1) on the *per-task 5-dimension mean composite*. For transparency, alternative formulations on the same data are also reported: the stacked 500-observation pooling (N = 100 tasks × 5 dimensions treated as pseudo-replicates) yields ICC(A,1) = 0.844; the k = 2 average-rater Spearman–Brown formulation on the mean composite yields ICC(A,k) = 0.957. The headline 0.918 corresponds to the per-task mean-composite single-rater formulation. 4-dimension ICC (excluding D3) = 0.927. D3 shows the lowest prompt stability. Judgement assessment is ambiguous, which supports treating D3 as a separate axis.

Table S7. Aggregation method sensitivity (AIOE correlation)

Method	Pearson r(AIOE)	Spearman (AIOE)
Equal-weight	0.831	0.832
Importance-weighted	0.829	0.830
PCA	0.898	0.892
Two-layer	0.750	0.759

Note. Pearson and Spearman correlations between each aggregation method's occupation-level D1 score and AIOE (N = 683 matched at 6-digit SOC level). Importance-weighted and equal-weight produce nearly identical results. PCA maximises AIOE correlation by construction. This comes at the cost of interpretability.

Cross-method correlation: Pearson r between importance-weighted and equal-weight = 0.979; between importance-weighted and high-importance-only = 0.937 (N = 923 occupations).

Table S8. Quadrant robustness: newly exposed count across methods and thresholds

Method	P25	Median	P75
D1 headline (importance-wt)	**43**	**37**	**40**
Equal-weight	58	45	52
Importance-weighted	59	45	53
PCA	43	36	43
Two-layer	69	57	61
4-dim equal-weight	48	37	45
4-dim importance-weighted	47	36	45

Note. Number of 'newly exposed' occupations (above-median D1, the headline digital-interface exposure proxy, and at-or-below-median AIOE) at each threshold level (P25, Median, P75). The first row (bold) is the main-text specification: occupation-level D1 headline proxy aggregated from task-level scores using O*NET importance weights. Rows 2–7 use composite scores across multiple dimensions. At the median

threshold, the count ranges from 36 to 57 across methods. A core of 33 occupations appears under every composite method (Jaccard stability = 0.614). N = 683 occupations matched at 6-digit SOC level.

Table S9. Top 20 and bottom 20 occupations by ATEI 4-dim composite

Panel A. Top 20 (highest four-dimension composite)

Rank	O*NET code	Occupation title	ATEI 4-dim composite	D3 (ATEI-J)	Tasks
1	41-3041.00	Travel Agents	4.68	3.74	8
2	29-2072.00	Medical Records Specialists	4.60	3.41	17
3	43-9041.00	Insurance Claims and Policy Processing Clerks	4.57	3.48	25
4	43-4051.00	Customer Service Representatives	4.54	3.25	13
5	15-1243.01	Data Warehousing Specialists	4.52	3.11	18
6	15-1253.00	Software Quality Assurance Analysts and Testers	4.51	3.19	30
7	15-2051.01	Business Intelligence Analysts	4.51	2.95	17
8	13-1081.02	Logistics Analysts	4.49	3.29	31
9	15-1299.01	Web Administrators	4.48	3.07	35
10	43-3051.00	Payroll and Timekeeping Clerks	4.47	3.67	21
11	43-4021.00	Correspondence Clerks	4.47	3.64	17
12	15-1254.00	Web Developers	4.47	2.95	29
13	13-2053.00	Insurance Underwriters	4.46	2.42	7
14	43-3031.00	Bookkeeping, Accounting, and Auditing Clerks	4.46	3.93	28
15	15-2051.00	Data Scientists	4.45	2.81	16
16	43-4151.00	Order Clerks	4.43	3.84	19
17	15-1242.00	Database Administrators	4.39	2.67	18
18	43-9022.00	Word Processors and Typists	4.39	4.62	19
19	15-2099.01	Bioinformatics Technicians	4.37	3.15	19
20	43-4131.00	Loan Interviewers and Clerks	4.37	3.23	18

Panel B. Bottom 20 (lowest four-dimension composite)

Rank	O*NET code	Occupation title	ATEI 4-dim composite	D3 (ATEI-J)	Tasks
904	47-2072.00	Pile Driver Operators	0.60	3.59	5
905	47-2082.00	Tapers	0.59	3.92	16
906	49-9063.00	Musical Instrument Repairers and Tuners	0.59	3.10	24
907	47-2042.00	Floor Layers, Except Carpet, Wood, and Hard Tiles	0.57	4.21	14
908	47-3012.00	Helpers—Carpenters	0.56	4.06	18
909	51-4071.00	Foundry Mold and Coremakers	0.56	4.46	13
910	47-5043.00	Roof Bolters, Mining	0.55	3.55	14
911	47-2181.00	Roofers	0.55	3.49	27

Rank	O*NET code	Occupation title	ATEI 4-dim composite	D3 (ATEI-J)	Tasks
912	47-2071.00	Paving, Surfacing, and Tamping Equipment Operators	0.54	3.73	20
913	47-5071.00	Roustabouts, Oil and Gas	0.52	3.96	13
914	47-2051.00	Cement Masons and Concrete Finishers	0.52	3.48	26
915	47-2022.00	Stonemasons	0.51	3.29	16
916	47-3014.00	Helpers—Painters, Paperhangers, Plasterers, and Stucco Masons	0.51	5.00	11
917	51-6051.00	Sewers, Hand	0.51	4.43	11
918	51-3022.00	Meat, Poultry, and Fish Cutters and Trimmers	0.50	3.92	12
919	47-2053.00	Terrazzo Workers and Finishers	0.50	4.19	26
920	49-9045.00	Refractory Materials Repairers, Except Brickmasons	0.47	4.17	10
921	47-2043.00	Floor Sanders and Finishers	0.46	3.39	7
922	47-3016.00	Helpers—Roofers	0.45	4.31	18
923	51-3023.00	Slaughterers and Meat Packers	0.45	2.63	14

Note. ATEI 4-dim composite = importance-weighted occupation-level mean of D1, D2, D4, D5 (0–5 scale). D3 (ATEI-J) = judgement constraint axis. Tasks = number of O*NET task statements. Top-ranked occupations are concentrated in Computer and Mathematical (SOC 15) and Office and Administrative Support (SOC 43). Bottom-ranked occupations are concentrated in Construction and Extraction (SOC 47) and Production (SOC 51).

Table S10. Newly exposed occupations (N = 37)

SOC (6-digit)	Occupation title	SOC major group	D1 score	AIOE	ATEI-J
29-2051	Dietetic Technicians	Healthcare Practitioners and Technical	3.44	−0.13	2.36
33-1012	First-Line Supervisors of Police and Detectives	Protective Service	3.29	−0.19	1.57
43-5071	Shipping, Receiving, and Inventory Clerks	Office and Administrative Support	3.25	−0.75	3.74
47-1011	First-Line Supervisors of Construction Trades and Extraction Workers	Construction and Extraction	3.09	−0.31	2.19
11-9051	Food Service Managers	Management	3.03	−0.12	2.25
35-1012	First-Line Supervisors of Food Preparation and Serving Workers	Food Preparation and Serving Related	2.95	−0.26	2.45
37-1011	First-Line Supervisors of Housekeeping and Janitorial Workers	Building and Grounds Cleaning and Maintenance	2.81	−0.59	2.74
43-5051	Postal Service Clerks	Office and Administrative Support	2.72	−0.52	3.74
29-2032	Diagnostic Medical Sonographers	Healthcare Practitioners and Technical	2.72	−0.26	2.05
43-5041	Meter Readers, Utilities	Office and Administrative Support	2.72	−0.97	4.01
29-1123	Physical Therapists	Healthcare Practitioners and Technical	2.69	−0.36	1.49
27-1013	Fine Artists, Including Painters, Sculptors, and Illustrators	Arts, Design, Entertainment, Sports, and Media	2.68	−0.57	2.15
49-2094	Electrical and Electronics Repairers, Commercial and Industrial Equipment	Installation, Maintenance, and Repair	2.67	−0.46	2.91
29-2099	Neurodiagnostic Technologists	Healthcare Practitioners and Technical	2.67	−0.28	2.28
17-3024	Electro-Mechanical and Mechatronics Technologists and Technicians	Architecture and Engineering	2.63	−0.30	3.12
41-2011	Cashiers	Sales and Related	2.63	−0.25	3.97
37-1012	First-Line Supervisors of Landscaping, Lawn Service, and Groundskeeping Workers	Building and Grounds Cleaning and Maintenance	2.62	−0.55	2.28
53-2011	Airline Pilots, Copilots, and Flight Engineers	Transportation and Material Moving	2.61	−0.21	1.74
27-1012	Craft Artists	Arts, Design, Entertainment, Sports, and Media	2.60	−0.93	2.56
53-2012	Commercial Pilots	Transportation and Material Moving	2.59	−0.18	1.77

SOC (6-digit)	Occupation title	SOC major group	D1 score	AIOE	ATEI-J
19-1032	Foresters	Life, Physical, and Social Science	2.59	−0.18	1.75
25-2059	Adapted Physical Education Specialists	Educational Instruction and Library	2.55	−0.66	1.81
31-9095	Pharmacy Aides	Healthcare Support	2.55	−0.16	3.97
49-2095	Electrical and Electronics Repairers, Powerhouse, Substation, and Relay	Installation, Maintenance, and Repair	2.47	−0.55	3.06
29-1011	Chiropractors	Healthcare Practitioners and Technical	2.46	−0.19	1.47
11-9013	Farmers, Ranchers, and Other Agricultural Managers	Management	2.44	−0.14	2.52
51-9151	Photographic Process Workers and Processing Machine Operators	Production	2.42	−0.36	4.06
27-4031	Camera Operators, Television, Video, and Film	Arts, Design, Entertainment, Sports, and Media	2.42	−0.45	2.48
45-1011	First-Line Supervisors of Farming, Fishing, and Forestry Workers	Farming, Fishing, and Forestry	2.40	−0.48	2.48
35-1011	Chefs and Head Cooks	Food Preparation and Serving Related	2.40	−0.27	1.93
53-4031	Railroad Conductors and Yardmasters	Transportation and Material Moving	2.39	−0.47	2.70
53-6051	Transportation Inspectors	Transportation and Material Moving	2.38	−0.18	2.37
41-9012	Models	Sales and Related	2.37	−1.12	2.61
29-2034	Radiologic Technologists and Technicians	Healthcare Practitioners and Technical	2.36	−0.56	2.36
29-1124	Radiation Therapists	Healthcare Practitioners and Technical	2.36	−0.22	2.07
27-2032	Choreographers	Arts, Design, Entertainment, Sports, and Media	2.35	−1.58	1.59
27-4021	Photographers	Arts, Design, Entertainment, Sports, and Media	2.35	−0.17	3.11

Note. 'Newly exposed' = above-median D1, the headline digital-interface exposure proxy, and at-or-below-median AIOE at the 6-digit SOC level. D1 = importance-weighted headline proxy. AIOE = AI Occupational Exposure index (Felten et al 2021). ATEI-J = D3 Judgement constraint axis. Occupations are sorted by D1 score (descending). These 37 occupations account for approximately 8.3 million workers (6.8% of the matched workforce; BLS OES May 2024). Occupational mix: first-line supervisors, healthcare technicians, transportation roles, administrative-support clerks.

Table S11. Dimension-level Pearson correlations with occupation characteristics

Variable	D1 Digital	D2 GUI	D3 Judgement (reverse-coded)	D4 Composability	D5 Physical (reverse-coded)
Job Zone	0.637***	0.575***	−0.718***	0.278***	0.561***
Computer Use	0.815***	0.789***	−0.467***	0.681***	0.761***
Physical Demand	−0.815***	−0.784***	0.441***	−0.593***	−0.826***
Cognitive Demand	0.591***	0.522***	−0.602***	0.275***	0.527***
Routine Cognitive	0.557***	0.527***	−0.521***	0.399***	0.482***
AIOE	0.908***	0.879***	−0.596***	0.676***	0.881***

Note. Pearson r at occupation level. N = 782 (AIOE) or 894 (O*NET characteristics). *** $p < 0.001$ (two-tailed). D3 reverse-coded: higher = less judgement, hence negative signs throughout; the negative D3-Job Zone and D3-Cognitive Demand pairings flag judgement requirements rising with training and cognition. D4 weakest on Job Zone (0.278) and Cognitive Demand (0.275), consistent with its unique contribution past D1.

Table S12. Inter-dimension Pearson correlation matrix (occupation level, N = 923)

	D1 Digital	D2 GUI	D3 Judgement (reverse-coded)	D4 Composability	D5 Physical (reverse-coded)
D1 Digital	1.000	0.985	−0.499	0.819	0.978
D2 GUI	0.985	1.000	−0.452	0.861	0.969
D3 Judgement (reverse-coded)	−0.499	−0.452	1.000	−0.034	−0.401
D4 Composability	0.819	0.861	−0.034	1.000	0.826
D5 Physical (reverse-coded)	0.978	0.969	−0.401	0.826	1.000

Note. Importance-weighted occupation-level scores. D1–D2 ($r = 0.985$), D1–D5 ($r = 0.978$), and D2–D5 ($r = 0.969$) form a collinear cluster. D3 carries negative signs against every other dimension and sits near-orthogonal to D4 ($r = -0.034$), consistent with D3 as a separate constraint axis. D4 correlates moderately with D1 ($r = 0.819$) and retains partial independence.

Table S13. Imputation sensitivity (missing importance weights)

Comparison	Pearson r	p	MAD
Importance-weighted vs equal-weight	0.979	< 0.001	0.292
Importance-weighted vs high-importance-only	0.937	< 0.001	0.372
Equal-weight vs high-importance-only	0.941	< 0.001	0.249

Note. N = 923 occupations. MAD = mean absolute deviation on the 0–5 scale. 845 tasks (4.5%) have missing importance weights and are imputed with equal weights within the affected occupations. Importance-weighted and equal-weight aggregations correlate at $r = 0.979$; the imputation choice has minimal effect on occupation rankings.

Table S14. D1 incremental validity over O*NET Computer Use and Physical Demand

DV	N	R^2 baseline	R^2 augmented	ΔR^2	D1	D1 t	D1 p
AIOE	781	0.850	0.887	0.036	0.307	15.773	< 0.001***
Job Zone	894	0.460	0.466	0.006	0.146	3.256	0.001**
Cognitive Demand	894	0.452	0.460	0.008	0.077	3.611	< 0.001***
Routine Cognitive	894	0.572	0.573	0.000	0.014	0.646	0.518

Note. Baseline model: DV ~ Computer Use + Physical Demand. Augmented model: DV ~ Computer Use + Physical Demand + D1. ΔR^2 is the incremental variance explained by D1 after controlling for both O*NET descriptors. D1 is significant for three of four dependent variables. The D1 headline digital-interface exposure proxy captures occupational characteristics beyond basic computerisation and physical-activity levels. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S15. D1 incremental validity over O*NET Computer Use alone

DV	N	R^2 (CU only)	R^2 (CU + D1)	ΔR^2	D1 p
AIOE	781	0.618	0.832	0.215	< 0.001***
Job Zone	894	0.369	0.436	0.067	< 0.001***
Cognitive Demand	894	0.451	0.457	0.006	0.002**
Routine Cognitive	894	0.552	0.559	0.007	< 0.001***
Physical Demand	894	0.413	0.666	0.254	< 0.001***

Note. Baseline model: DV ~ Computer Use. Augmented model: DV ~ Computer Use + D1. D1 adds significant incremental variance for all five dependent variables, with the largest increments for AIOE ($\Delta R^2 = 0.215$) and Physical Demand ($\Delta R^2 = 0.254$). Across all five DVs, D1 carries variance that the Computer Use descriptor alone does not. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S16. D1 headline proxy and the Anthropic Economic Index

Benchmark	Pearson r	Spearman	p	N
Observed exposure	0.571	0.650	< 0.001	474
Automation share	0.169	0.189	< 0.001	495
Augmentation share	−0.169	−0.188	< 0.001	495

Note. Anthropic Economic Index metrics (Anthropic 2025) are derived from observed AI tool usage in real-world interaction logs. Observed exposure measures the share of tasks within an occupation that show evidence of AI tool engagement. The D1 side of each correlation uses the headline D1 aggregation (task-level D1 scores averaged to occupation level at 6-digit SOC), matched to the Anthropic benchmark by SOC code. The moderate-to-strong correlation with D1 (= 0.650) is criterion-adjacent evidence that the D1 headline proxy relates to observed AI tool usage; observed usage reflects deployment rather than verified agent task-completion outcomes. Automation share and augmentation share are complementary (sum to 1.0). High-D1 occupations correlate positively with automation share and negatively with augmentation share. The bootstrap 95 % CI for the observed-exposure Pearson correlation is reported in Table S17.

Table S17. Bootstrap 95 % confidence intervals for key validity and reliability estimates

#	Estimate	Point	95 % BCa CI	SE	N
1	D1 × AIOE Pearson r	0.908	[0.895, 0.919]	0.006	782
2	D4 ΔR^2 over D1, Job Zone (M2)	0.181	[0.149, 0.221]	0.018	923
3	D4 ΔR^2 over D1, Cognitive Demand (M2)	0.131	[0.097, 0.170]	0.019	894
4	Three-model stratified ICC(A,1)	0.913	[0.892, 0.930]	0.011	220
5	Prompt sensitivity composite ICC(A,1)	0.918	[0.887, 0.943]	0.015	100
6	D1 (headline) × Anthropic observed exposure Pearson r	0.571	[0.526, 0.607]	0.021	474

Note. Bias-corrected and accelerated (BCa) bootstrap confidence intervals with 1,000 resamples (random seed = 42), computed via `scipy.stats.bootstrap` with `paired = True` for statistics over paired arrays (Pearson, ICC, and the regression ΔR^2 estimates). For Estimate 4, the three-model stratified ICC is computed on $N = 220$ stratified tasks (22 SOC strata × 10 tasks per stratum), each independently scored by GPT-4.1-mini, GPT-4o and Claude Sonnet 4. For Estimate 5, the prompt sensitivity ICC is the single-rater ICC(A,1) on the per-task 5-dimension mean composite (see Table S6). Rows 2–3 use the full sample (N = 923 / 894); the corresponding AIOE-matched subsample estimates in Table 4 (N = 782) are 0.183 and 0.134—the minor discrepancy reflects sample restriction, not a specification difference. All six point estimates are bit-exact reproducible from the scoring CSVs.

Table S18. Temporal stability of the D1 headline proxy across O*NET versions

Panel A. Task-level Pearson correlation on D1 (strict task-ID overlap)

Version pair		Pearson r	95 % BCa CI	N tasks
O*NET 28.2	29.1	0.992	[0.991, 0.993]	2,002
O*NET 29.1	30.2	0.955	[0.950, 0.958]	2,002
O*NET 28.2	30.2	0.955	[0.951, 0.959]	2,002

Panel B. Occupation-level Spearman rank correlation on full ATEI ranking

Version pair		Spearman	95 % Fisher-z CI	N occupations
O*NET 28.2	29.1	0.993	[0.991, 0.994]	804
O*NET 29.1	30.2	0.962	[0.957, 0.967]	804
O*NET 28.2	30.2	0.963	[0.957, 0.968]	804

Note. A stratified sample of $N = 2{,}002$ tasks was rescored under each O*NET version using the identical GPT-4.1-mini scoring prompt (see Appendix A.1). The sample is proportional to SOC major group (minimum 20 tasks per group). Panel A reports Pearson correlations on task-level D1 scores, restricted to tasks whose task identifier appears in all three versions (strict overlap). Panel B reports Spearman rank correlations on the full occupation-level ATEI ranking, after aggregating to the occupation level within each O*NET version. The two panels use different denominators (strict task-ID overlap for Panel A, full occupation set for Panel B), so the sample sizes differ. The ATEI ranking is stable across O*NET versions: adjacent-version correlations are 0.955 and the two-step (28.2 30.2) correlation is 0.955. The scoring prompt and model are held fixed across versions; any across-version variation comes from the task inventory.

Table S19. Leave-one-out SOC major-group sensitivity for the D1 headline proxy × AIOE correlation

SOC code	SOC major group (held out)	r (D1, AIOE)	N remaining
11	Management	0.9043	736
13	Business and Financial Operations	0.9031	738
15	Computer and Mathematical	0.9065	777
17	Architecture and Engineering	0.9075	727
19	Life, Physical and Social Science	0.9082	733
21	Community and Social Service	0.9099	768
23	Legal	0.9076	775
25	Education, Training and Library	0.9063	730
27	Arts, Design, Entertainment, Sports and Media	0.9132	746
29	Healthcare Practitioners and Technical	0.9152	726
31	Healthcare Support	0.9081	768
33	Protective Service	0.9100	758
35	Food Preparation and Serving Related	0.9069	768
37	Building and Grounds Cleaning and Maintenance	0.9074	774
39	Personal Care and Service	0.9122	755
41	Sales and Related	0.9089	761
43	Office and Administrative Support	0.9096	731
45	Farming, Fishing and Forestry	0.9069	771
47	Construction and Extraction	0.8948	726
49	Installation, Maintenance and Repair	0.9060	732
51	Production	0.9028	680
53	Transportation and Material Moving	0.9058	742

Summary statistics across the 22 iterations: minimum 0.8948 (SOC 47 Construction and Extraction held out), median 0.9074, maximum 0.9152 (SOC 29 Healthcare Practitioners held out), standard deviation 0.0041. Full-sample baseline (no group held out): $r = 0.9076$ (Table S17).

Note. Each iteration drops one 2-digit SOC major group and recomputes D1 × AIOE Pearson on the remaining 6-digit occupations. N remaining varies across iterations because the 22 major groups cover different counts of matched 6-digit rows. Full 22-iteration range: [0.8948, 0.9152], SD = 0.0041; no single major group anchors the headline 0.9076. Minimum lands on Construction and Extraction (SOC 47), a low-D1, low-AIOE cluster at the scatter's lower-left whose removal compresses the score range. To three decimals, this minimum (0.8948) coincides with the BCa 95% lower bound (0.895) reported in Table S17.

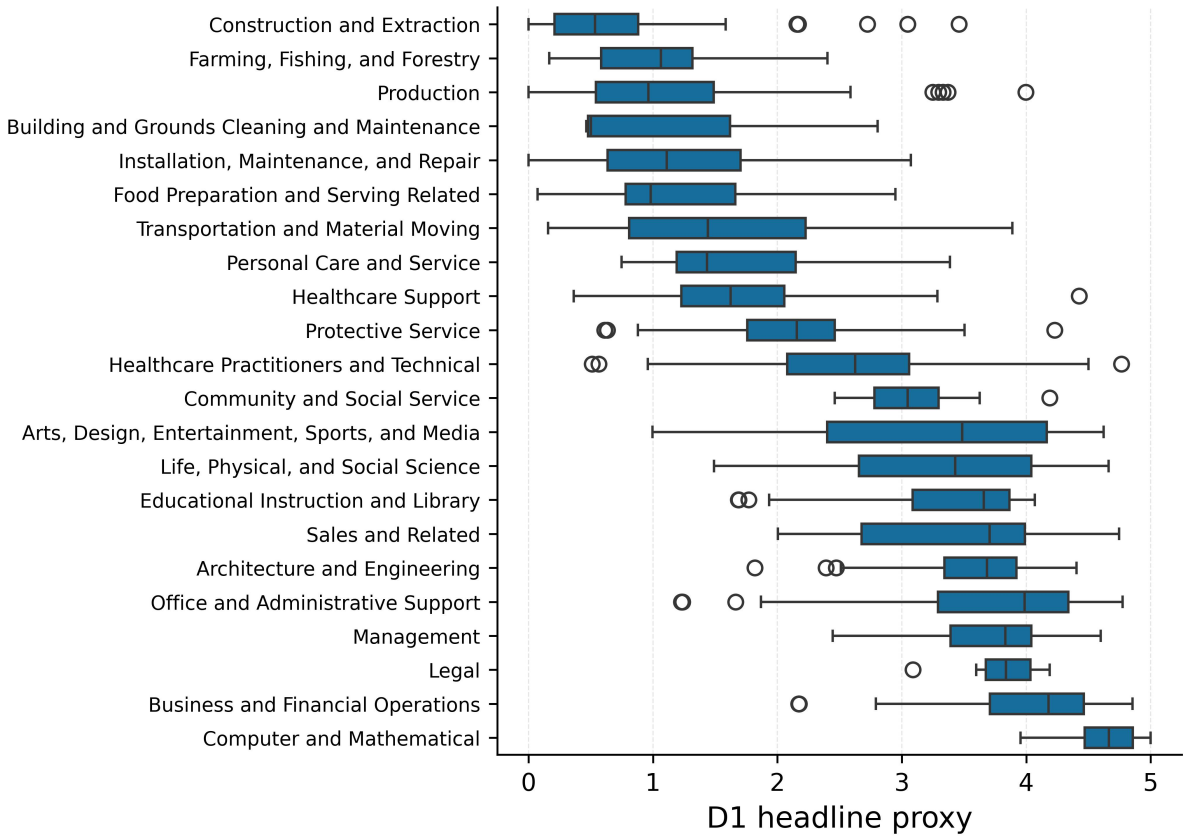

Figure S1. D1 headline proxy by SOC major occupation group. Box plots display D1 headline proxy distributions across the 22 SOC major occupation groups (boxes: IQR with median line; whiskers: 1.5× IQR). Computer and Mathematical lead the median ordering; Construction and Extraction trail.

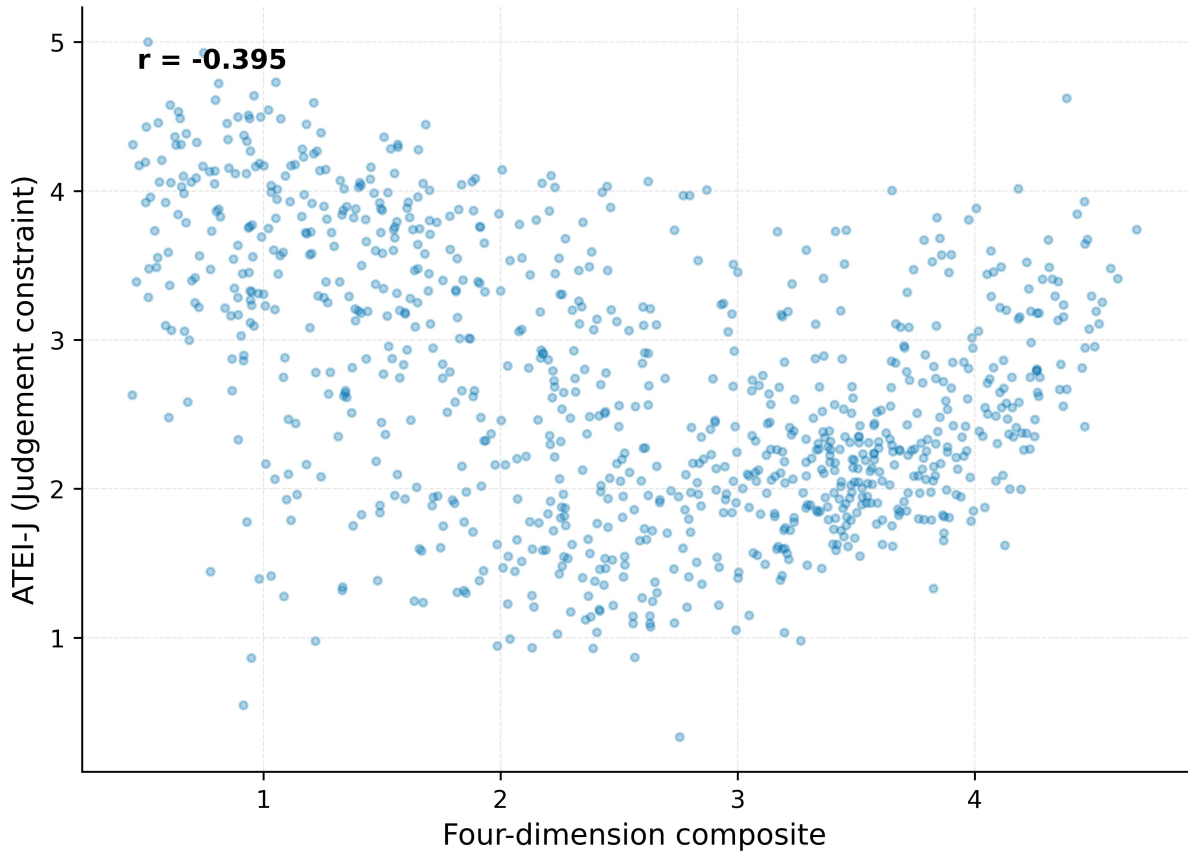

Figure S2. Four-dimension composite vs. ATEI-J (Judgement constraint). Scatter plot of occupation-level four-dimension composite (D1, D2, D4, D5) against ATEI-J (judgement constraint) for 923 occupations ($r = -0.395$). The two axes are near-orthogonal; ATEI-J therefore stays a separate constraint axis, kept outside the D1 headline proxy.

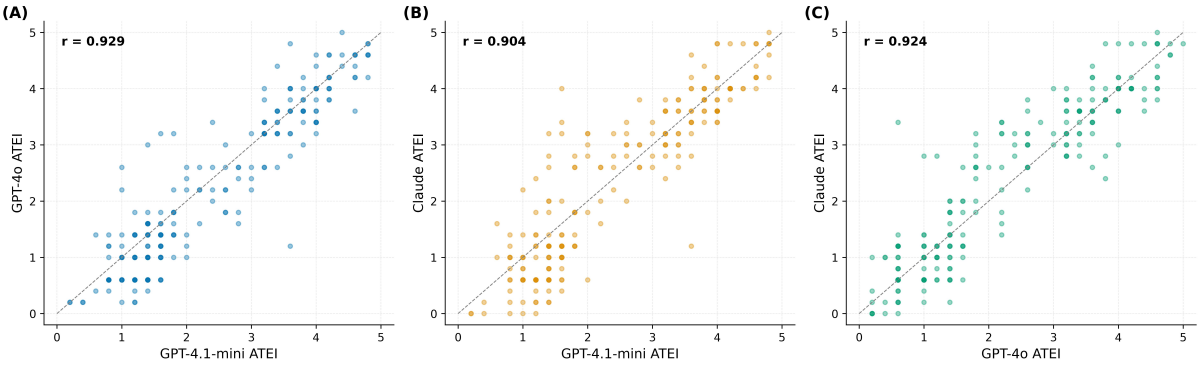

Figure S3. Cross-model agreement scatter plots. Pairwise scatter plots of occupation-level ATEI scores across the three scoring models: (A) GPT-4.1-mini vs. GPT-4o, $r = 0.929$; (B) GPT-4.1-mini vs. Claude, $r = 0.904$; (C) GPT-4o vs. Claude, $r = 0.924$. Dashed line in each panel marks perfect agreement. Pairwise correlations track the three-model ICC of 0.913 (main text).

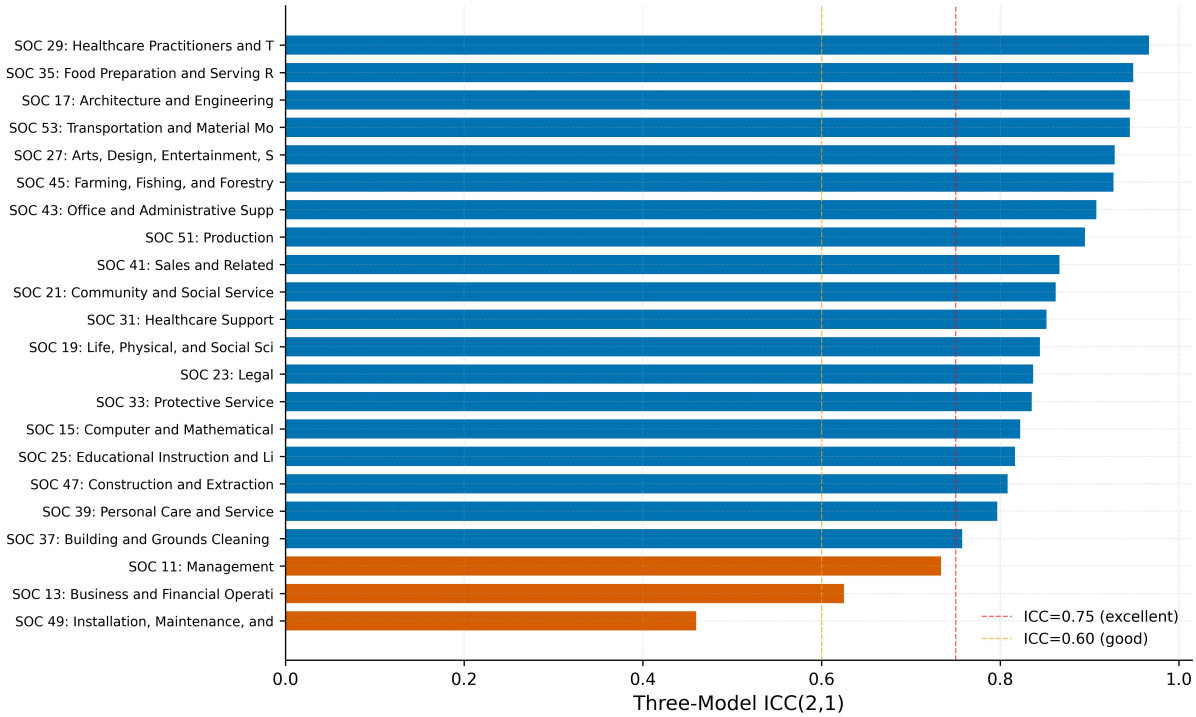

Figure S4. Three-model ICC by SOC major occupation group. Three-model ICC(2,1) values across the 22 SOC major occupation groups. Reference lines: ICC = 0.50 (moderate cut-off), 0.75 (good); excellent 0.90. Most groups land at good or excellent. Three exceptions drop to acceptable: Installation, Maintenance and Repair; Business and Financial Operations; Management.

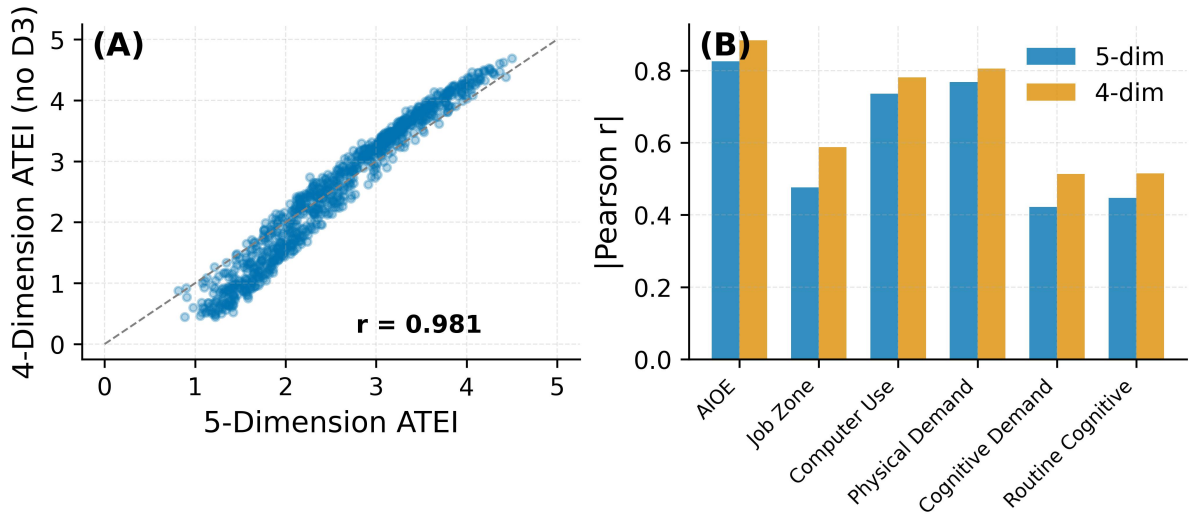

Figure S5. Four-dimension vs. five-dimension ATEI comparison. Panel A: scatter of four-dimension ATEI (excluding D3) against five-dimension ATEI for 923 occupations ($r = 0.981$). Panel B: paired bars of the five- and four-dimension composites against six external benchmarks (AIOE, Job Zone, Computer Use, Physical Demand, Cognitive Demand, Routine Cognitive); bar height = $|r|$. The four-dimension composite matches or exceeds the five-dimension composite on all six. D3 reports separately as ATEI-J; the four-dimension composite stays in the appendix as a diagnostic/sensitivity comparator, outside the headline.

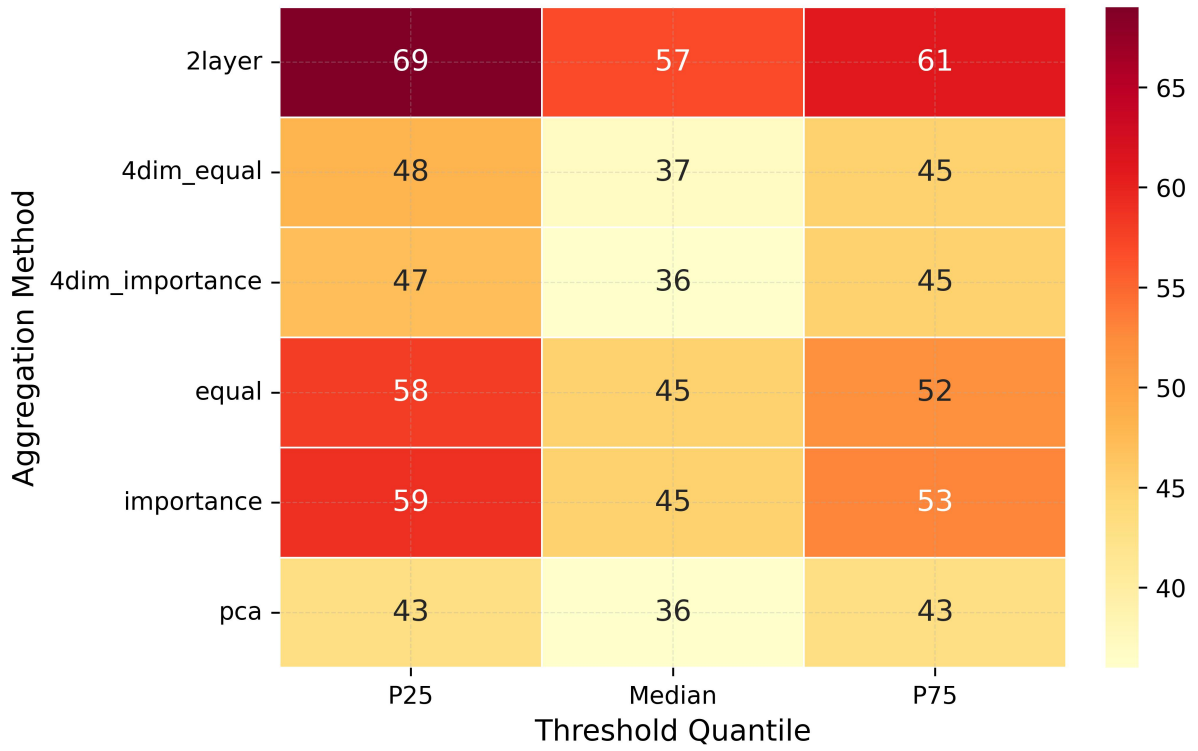

Figure S6. Newly exposed occupation count by aggregation method and threshold. Heatmap cells: newly exposed occupation counts under six aggregation methods (importance-weighted, equal-weight, PCA, four-dimension importance, four-dimension equal, two-layer) × three threshold quantiles (P25, Median, P75). Cell values span 36 to 69. At the median threshold, a 33-occupation core recurs under every composite method. The primary D1-based specification is shown separately in Table S8.

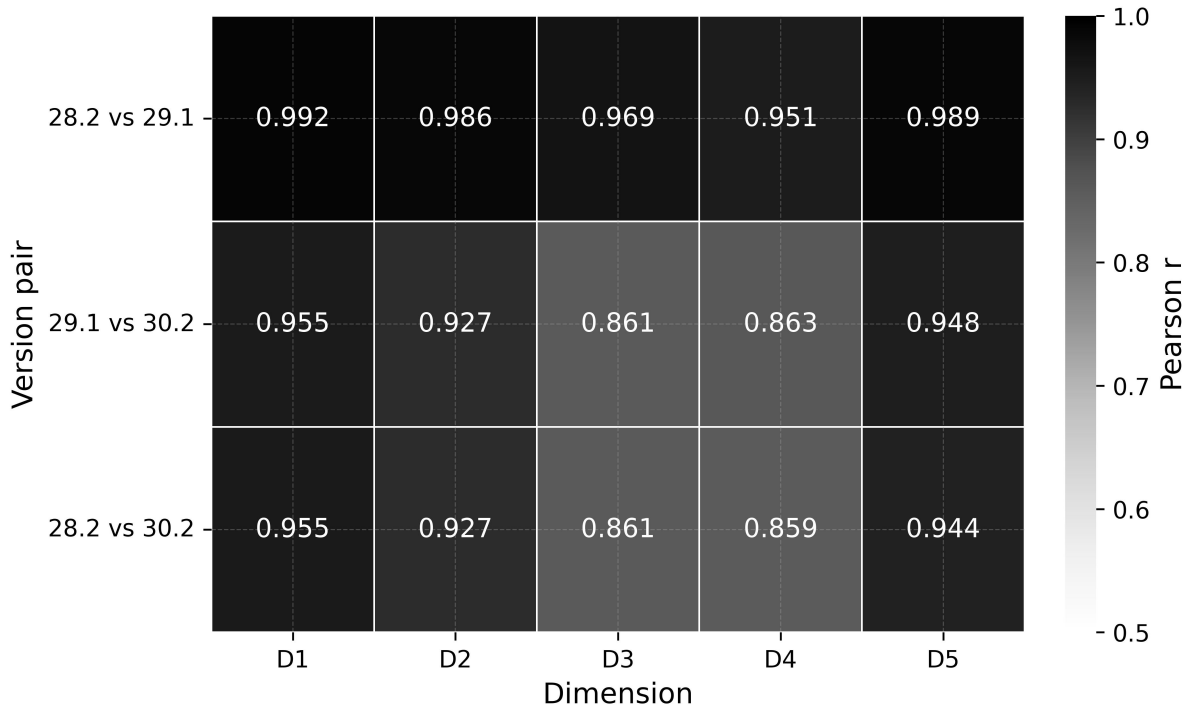

Figure S7. Temporal stability heatmap of the five ATEI dimensions across O*NET releases. Heatmap of pairwise Pearson correlations between task-level ATEI dimension scores across O*NET releases 28.2, 29.1, and 30.2. Correlations are computed on the strict Task ID subset present in all three releases ($N =$ 2,002). D1 r exceeds 0.95 across all three release pairs.

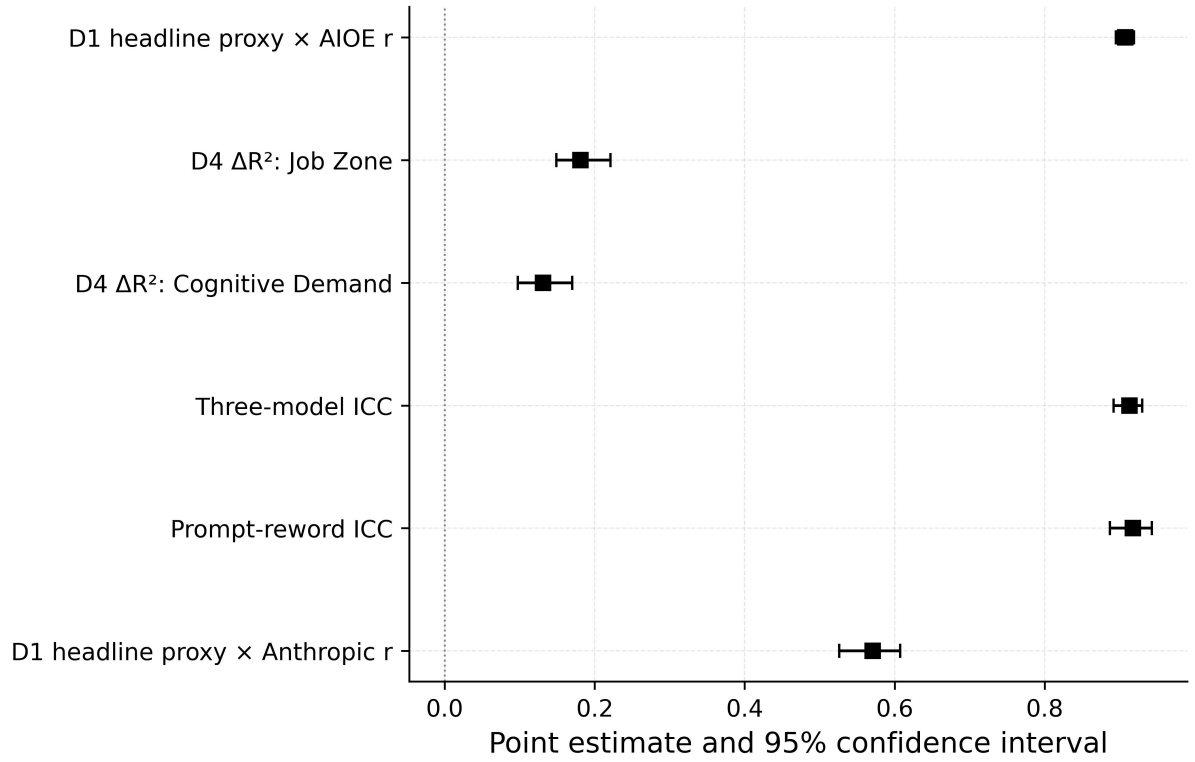

Figure S8. Bootstrap 95% confidence intervals for six key ATEI estimates. Forest plot of point estimates with bias-corrected and accelerated bootstrap 95% confidence intervals (1,000 resamples). Metrics include D1 × AIOE r, D4 ΔR^2 for Job Zone and Cognitive Demand (M2), three-model stratified ICC, prompt-sensitivity ICC, and D1 × Anthropic observed exposure r. All six intervals exclude zero.

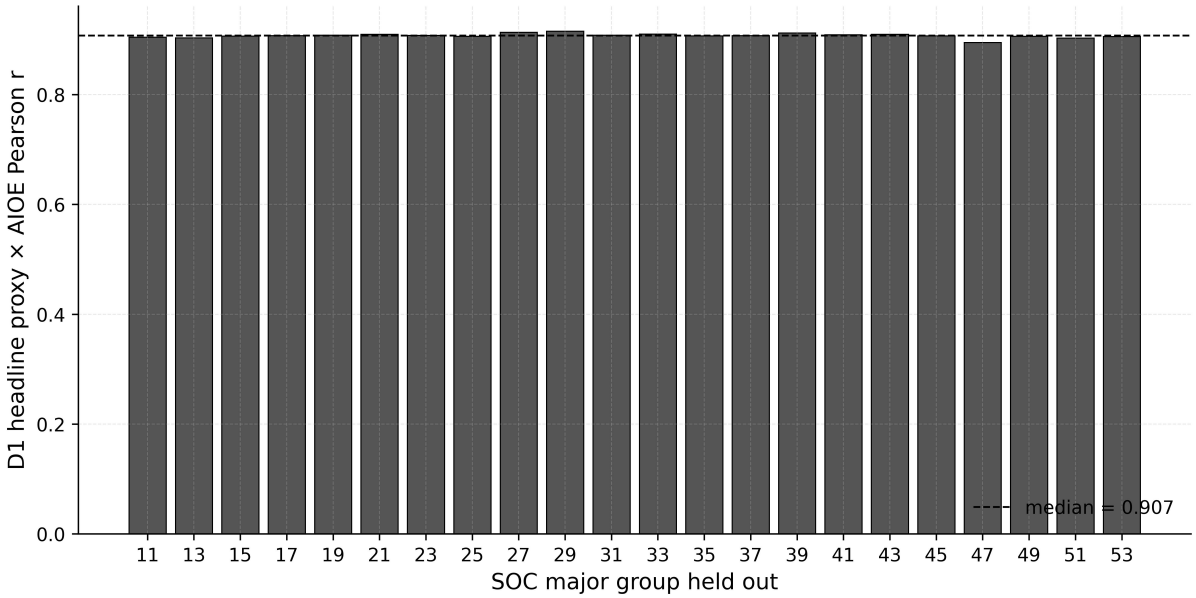

Figure S9. Leave-one-out SOC major-group sensitivity for the D1 headline proxy × AIOE correlation. Bar chart of D1 × AIOE Pearson correlations recomputed after dropping each of the 22 SOC major groups in turn. Values range from 0.895 to 0.915 with standard deviation 0.0041 across all 22 iterations. The headline correlation is robust to any single SOC major group exclusion.

Reproducibility, Code and Data Availability

Software environment All analyses run on Google Colab Pro+ (Linux, Python 3.12) with `pandas 2.2.2`, `numpy 2.0.2`, `scipy 1.16.3`, `scikit-learn 1.6.1`, `statsmodels 0.14.6`, and `pingouin 0.6.1`. The notebooks set `SEED = 42` for all random number generators; bootstrap analyses use 1,000 BCa replicates.

Code availability Nine Jupyter notebooks (`NB01`–`NB08`, `NB06s`) reproduce all primary analyses, validation, and supplementary robustness in this study. `notebooks/README.md` lists execution order, dependencies, and runtime requirements. Each notebook writes CSV tables, quad-format figures (JPEG + TIFF + PDF + EPS at 600 DPI), and per-notebook `nbXX_run_config.json` + `nbXX_environment.txt` + `table_manifest.txt` under `Results/NB0X_results/{tables,figures,intermediate,reproducibility}/`.

Notebook	Reproduces
NB01	Data preparation (O*NET 30.2 → 18,796 tasks across 923 occupations)
NB02	Task-level rubric scoring (five dimensions, three-call majority vote, GPT-4.1-mini)
NB03	Three-model cross-vendor reliability (reported in Methods, Table 2; ICC = 0.913)
NB04	Five aggregation methods; primary importance-weighted composite
NB05	External and construct validity (Tables S1, S2, S7, S11, S12)
NB06	Robustness diagnostics: VIF, HTMT, structure competition, D4 controlled validity, prompt sensitivity, imputation, quadrant robustness (Tables S3–S6, S8, S13)
NB06s	Anthropic convergent validity (Table S16); D1 incremental validity over O*NET descriptors (Tables S14–S15)
NB07	Descriptive statistics, quadrant analysis, BLS employment translation (Tables S9, S10); main Figures 1–5
NB08	Temporal stability across O*NET 28.2/29.1/30.2 (Table S18), bootstrap 95% confidence intervals (Table S17), leave-one-out SOC major-group sensitivity (Table S19)

Data availability All datasets are publicly available: O*NET 30.2 (also 28.2 and 29.1 for temporal stability; U.S. Department of Labor; CC-BY-4.0); the Anthropic Economic Index (Anthropic 2025); OpenAI Signals (OpenAI public release); the AIOE Data Appendix (Felten et al 2021); and the BLS OES May 2024 National Employment Estimates. The reproducibility package documents the expected `dataset/` folder layout; data files should be downloaded from the original public sources cited above.

LLM scoring providers Task-level scoring uses three independent commercial LLM APIs accessed during April–May 2026 (NB02 primary scoring and NB03 cross-vendor reliability used April 2026 calls; NB06 prompt-sensitivity reword and NB08 historical-version rescoring used May 2026 calls):

- OpenAI `gpt-4.1-mini` (primary scorer in NB02, prompt-sensitivity reword in NB06, and cross-version rescoring in NB08)
- OpenAI `gpt-4o` (cross-vendor reliability check in NB03)
- Anthropic `claude-sonnet-4-20250514` (cross-vendor reliability check in NB03)

All calls used a low temperature with three-call majority voting per dimension: `temperature = 0.1` for primary scoring (NB02), prompt-sensitivity reword (NB06), and historical-version rescoring (NB08); `temperature = 0.0` for the cross-vendor reliability check in NB03 to maximise determinism. Commercial LLM APIs have non-zero output randomness even at low temperature, and provider model weights may update between releases. A re-run of the same notebooks may produce ICC, correlation, or ΔR^2 values differing from those reported here in the third decimal place; such drift is within the precision reported in the manuscript.

Supplementary References

Anthropic (2025) The Anthropic Economic Index. Anthropic Research. https://www.anthropic.com/research/the-anthropic-economic-index. Accessed 7 May 2026

Felten E, Raj M, Seamans R (2021) Occupational, industry, and geographic exposure to artificial intelligence: a novel dataset and its potential uses. Strateg Manag J 42(12):2195–2217. https://doi.org/10.1002/smj.3286

Kenny DA, Kaniskan B, McCoach DB (2015) The performance of RMSEA in models with small degrees of freedom. Sociol Methods Res 44(3):486–507. https://doi.org/10.1177/0049124114543236