Do AI Activity Indices Really Measure AI Activity?

Anonymous Author(s)
Affiliation
Address
email

Abstract

AI dashboards increasingly shape economic debate, yet their headline movements often conceal what changed. For a ratio, the same rise can reflect numerator growth, denominator decline, or both. This ambiguity creates structural non-identification: adding observations or richer fixed effects cannot recover the missing information. We propose three propositions showing that a relative-index regression blends the causal effect of interest with co-movement between focal and aggregate activity and shocks shared with the outcome. These components can give the coefficient either sign under a zero effect, or erase it under a nonzero effect. Publishing denominator levels changes what can be learned by removing the co-movement channel exactly, although shared-shock confounding remains. The problem extends to power: if most identifying variation comes from a handful of units, those units determine what the design can detect, however large the nominal sample looks. At a fixed standard error, the minimum detectable effect diverges as the effective cluster count approaches one, leaving a null result uninformative. We apply this framework to OECD.AI's hiring index and Eurostat employment across 24-27 European countries, where the detection floor exceeds the reported association. In a refreshed panel, that association disappears once the 2021 survey break is excluded. Ratio indices can still reveal concentration, timing, and co-movement, but establishing displacement requires more; our framework turns that boundary into a practical standard by making denominator levels a minimum disclosure requirement.

1 Introduction

Almost everything we know about how AI is spreading through the economy arrives as a ratio: a model family's standing is its share of arena battles; a provider's reach is its share of served requests; adoption is the share of surveyed firms reporting use. The most quoted labour-market indicator, the OECD.AI relative AI hiring index, is the growth of AI-skilled hiring relative to the growth of all hiring [29].

The ratio is published so that nuisances such as coverage or the survey frame cancel, and a raw numerator would be no better defined. The problem is not the ratio, but that the provider does not release the two counts. If we ask whether AI is replacing workers, or whether a model family is falling behind, we are given a number that answers neither question, and no subsequent estimator can correct it. This can be shown with the Chatbot Arena, which samples pairs of models for human comparison [54] and releases its battle logs, so both counts are observable. Across 106,134 battles among 55 models between June and August 2024 [35], the Claude family's published share of battles fell from 34.6% to 16.4%. Over the same window its win rate against the 21 opponents present in both periods moved by -0.01 points once the family's own model mix is held fixed, inside a bootstrap interval of ± 4 points. The primary change occurred in the sampling process: fourteen entrants accounted for 57% of late-window battles. Although daily volume remained stable, the family's own count declined, resulting in a decreased ratio (Figure 1).

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

These examples motivate a narrow question: what can a regression on a relative activity index establish when the counts inside it are withheld? We give a formal answer, and we measure how the resulting identification and power constraints bind in the regression.

Contributions and scope. We study regressions of outcomes on published relative indices when the component levels are withheld, and we characterise the additional disclosure required for identification. Instead of estimating the employment effect of AI, we focus on whether the published indices could identify it at any size. Our contributions are:

- We derive an exact decomposition of the coefficient from a regression of any outcome on a published relative index: the causal effect enters summed with a cyclicality term, which is the classical ratio confound, and a shock covariance term. The decomposition survives every fixed-effects design, because each within-transformation adds one equation and one unknown, so the effect is not identified from the index at any magnitude (Proposition 1).
- We prove that publishing the denominator in levels removes the cyclicality term exactly while the covariance term survives, and that where the two have opposite signs disclosure can leave the total bias larger than before. The textbook remedy, entering the ratio alongside its parts [61], is not available to the analyst here, because the parts are proprietary; only the provider can enable it (Proposition 2).
- We establish a lower bound on the smallest detectable effect that depends on the effective number of clusters rather than on the sample size (Proposition 3).

Empirically, we join the hiring index to Eurostat employment for 24 to 27 European countries from 2018 to 2025. A one-standard-deviation higher index is associated with about 0.5 percentage points lower employment growth, but inference varies across procedures and two or three countries carry most of the signal. In the panel reaching 2025, the association survives only through the 2021 survey break. Its minimum detectable effect is 1.8 to 4.5 times the estimate. The dashboards reveal concentration, timing and co-movement, but not displacement.

2 Related work

Pearson showed that indices sharing a denominator correlate even when the underlying quantities do not [60], and Kronmal restated the problem for regression: a ratio may enter only alongside its parts as main effects [61]. Kuh and Meyer, and then Belsley, developed the same point for deflated variables in econometrics [62, 63], and Aitchison's log-ratio analysis is its compositional-data form [64]. Our published index is an Aitchison log-ratio, and Proposition 2 is Kronmal's prescription written out for a setting where the parts are withheld.

The task framework of Autor, Levy and Murnane [13] and the robot-exposure design of Acemoglu and Restrepo [14] set the template that generative AI inherited. Experimental and firm-level work finds productivity gains on cognitive tasks [15, 16, 43]; task-based measures imply broad exposure [17, 52, 53, 24, 25, 48] against modest aggregate effects [18]; and direct employment evidence is mixed and mostly national [19, 20, 33, 42, 50, 51, 49, 44, 45], focusing on how large the effect is. We ask whether the published indicators can identify that employment effect at any magnitude. Proposition 1 shows they cannot: with only the ratio index and the outcome observed, the estimated coefficient is consistent with every value of the effect: the coefficient can take either sign when the true effect is zero, and be zero when it is not.

Our object is the measurement layer these dashboards constitute [56]: OECD.AI's LinkedIn-derived hiring, migration and skills series [29], venture-capital aggregates [30], enterprise adoption surveys [32] and vacancy rates [47]. The same form governs how AI systems themselves are watched: Chatbot Arena publishes shares of pairwise battles [54, 35], and Singh et al. [34] document unequal sampling, selective disclosure of private variants and silent deprecation behind its ratings. It establishes that one platform's numbers are distorted, whereas we show that a published ratio cannot carry a causal reading however well the platform behaves.

When many decision-makers run the same algorithm, its mistakes stop cancelling out and they share the same errors [57]. The same logic applies to measurement: one dashboard supplies most public claims about AI and jobs, so having many readers gives no independent check on it. Separately, a prediction that gets acted on changes what it predicted [58], and people who know they are being

scored move to improve the score [59]. A ratio is especially easy to influence, since it rises either when AI-skilled hiring goes up or when total hiring goes down.

Cluster-robust asymptotics fail when a handful of clusters carry the identifying variation [21, 22, 38], effective cluster counts diagnose it [55], and randomization inference trades the reference distribution for an exchangeability assumption [36, 37]. The shift-share literature supplies the analogous warning for exposure designs [41, 40], as does pre-testing for parallel trends [39].

3 Methods

3.1 Data

Three providers supply the labour-market analysis, OECD.AI, Eurostat and the ILO, and Table 1 gives the series and their coverage, including a policy score of our own construction. Section 3.2's arena exhibit uses a further source, the released Chatbot Arena battle logs [35], which the table omits. We exclude 2026 everywhere as a partial year.

Table 1: Primary sources and their coverage. The text below states what each is used for.

Source and variable	Coverage
OECD.AI relative AI hiring index [5]	2018-01 to 2026-01, monthly, averaged to years
OECD.AI AI skills migration [3]	Per 10,000 LinkedIn members; extracts spanning 2019-2024 (Panel A) and 2021-2025 (Panel B)
OECD.AI VC in AI start-ups [1, 2]	By country and by industry, 2012-2026; the industry table's country dimension is the investor's country
OECD.AI AI skills penetration [4]	By country and industry, 2016-2024
Eurostat employment by occupation [6, 65]	Ages 15-64; Panel A uses an extract of table lfsa_egai2d, the other samples lfsa_epgais over 2016-2025
ILO generative-AI exposure scores [48]	427 ISCO-08 unit groups, aggregated to major groups with Eurostat employment weights
Eurostat employment by NACE two-digit [7]	Ages 20-64, 2018-2025
Eurostat covariates [8, 9, 10, 11, 12] and own policy score	Country-year; policy score hand-coded 0-3 (one point each: national AI strategy in force, major update, EU AI Act in force)

No single join spans 2019 to 2025, so we build two samples. Panel A (24 countries, 2019 to 2024, 144 country-years) has the longer window and the full covariate set and is the primary test; Panel B (26 countries, 2021 to 2025, 130 country-years) is the only sample that observes 2025 and the only one industry investment joins to. They overlap in 88 of Panel B's 130 country-years but differ in window, covariate set and index extract, so Panel B tests Panel A's coefficient on a different extract rather than resampling it, and a pooled panel would have concealed the disagreement. The occupation designs use no migration series, which is why they reach 27 countries.

3.2 Identification

Dashboards report activity as a ratio. Let $A_{i,t}$ be a focal count and $B_{i,t}$ an aggregate that contains it, both observed on one platform, and let the published series be $R_{i,t} = \Delta \log A_{i,t} - \Delta \log B_{i,t}$. For the hiring index A is AI-skilled hires and B total hires; for a leaderboard A is one family's battles and B all battles. This object has two key properties.

- Invariant to whatever scales A and B together, which is exactly why providers publish it: platform coverage, the member base and any common reporting convention cancel.
- Level information is discarded by the invariance: the same value may reflect a collapsing numerator, a growing denominator, or both, and the series cannot distinguish among them.

Two departures from this idealised object matter.

- The published index is smoothed, so $R = \Delta \log A - \Delta \log B + \omega$, where the slack ω is one further free nuisance: it leaves Proposition 1 untouched but weakens Proposition 2, since disclosing B then recovers the AI-specific shock only up to ω.

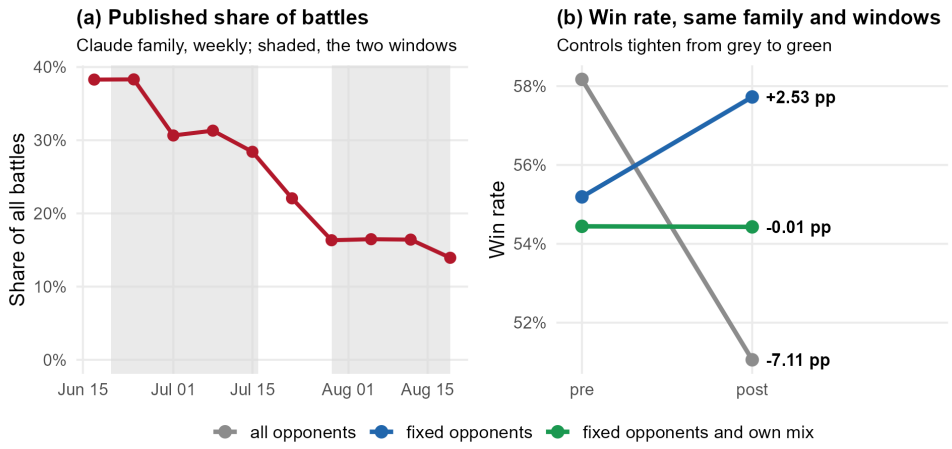

Figure 1: A published index and the signal it is read as. The Claude family's share of Chatbot Arena battles halves as entrants take 57% of a flat daily volume, while its win rate is flat once the opponent set and the family's own model mix are both held fixed [35, 27].

- This instrument is published as a growth differential, whilst most others of this kind are published as level shares $S = A/B$. The two are one object, since $\Delta \log S = \Delta \log A - \Delta \log B$, so a regression on $\Delta \log S$ inherits (2) term for term, and one on ΔS differs by a unit- and period-specific rescaling that changes no term's status.

Consequently, the results apply whether the provider publishes a level share or a growth differential.

For the hiring index we can argue that the two counts move for unrelated reasons but cannot show it, because neither is published. By contrast, the same analysis is possible for Chatbot Arena because it releases its battle logs [54]. The figures reported in the introduction locate the change in the numerator. Daily battle volume remained flat, and the roster fell from 41 models to 38, ruling out an expansion of B. Yet observing B would establish only that the movement occurred through A, not why A moved, as Proposition 2 shows. And holding opponents fixed, but not the family's own model mix, yields $+2.5$ points (95% CI -0.4 to $+5.5$) rather than -0.01, showing that the performance measure is sensitive to changes in within-family composition (Figure 1).

The leaderboard's headline is a Bradley-Terry rating, and it does not escape the problem, since it conditions on a pairing policy and prompt distribution the platform does not publish [34]. The required disclosure is the analogous one: per-family battle counts, the eligible model set per period, and the sampling policy, in levels.

We develop the algebra for the hiring case because the employment data needed to evaluate displacement are available and the index's denominator enters directly into employment dynamics. The identification problem therefore arises from the index itself rather than from a lack of outcome data. The identity $E_{i,t} = E_{i,t-1} + H_{i,t} - S_{i,t}$ makes aggregate hiring a common component of both the index and employment growth $g_{i,t}$, while the numerator measures the AI-skilled hiring margin whose effect on employment is the parameter of interest. Assume finite second moments and a common relative-cyclicality coefficient across the units entering the pooled moments below. Write the within-country projection of AI-skilled on total hiring as $\Delta \log h = \gamma \, \Delta \log H + u$, where γ is a relative cyclicality parameter and u an AI-specific shock orthogonal to $\Delta \log H$ by construction, and let

$$g_{i,t} \;=\; \delta \, u_{i,t} \;+\; \eta_{i,t}, \tag{1}$$

where δ is the displacement parameter of interest and η collects everything else, including the cycle. We define δ outside (1), as the causal effect of intervening on the AI-specific hiring shock while holding η, and with it aggregate hiring, fixed; (1) is a linearity restriction on that structure, not a definition of η, which would make it an accounting identity holding for any δ. We place no restriction on $\mathrm{Cov}(u, \eta)$: a common shock can move AI-specific hiring and employment together without either causing the other. Substituting the projection into the index gives $R = (\gamma - 1)\Delta \log H + u$, and

hence

$$\underbrace{\frac{\mathrm{Cov}(R,g)}{\mathrm{Var}(R)}}_{\beta,\ \text{estimated}} = \underbrace{\frac{(\gamma-1)\,\mathrm{Cov}(\Delta\log H,g)}{\mathrm{Var}(R)}}_{\text{relative cyclicality}} + \underbrace{\frac{\delta\,\mathrm{Var}(u)}{\mathrm{Var}(R)}}_{\text{displacement}} + \underbrace{\frac{\mathrm{Cov}(u,\eta)}{\mathrm{Var}(R)}}_{\text{common shock}}. \tag{2}$$

Our only object of interest is the displacement term.

Proposition 1 (Non-identification)**.** Let $\beta = \mathrm{Cov}(R,g)/\mathrm{Var}(R)$, and let the data identify only β and $\mathrm{Var}(R)$. Then δ is not identified from (R,g):

$$\forall\,\delta\in\mathbb{R},\ \forall\,b\in\mathbb{R},\quad \exists\,\gamma\in\mathbb{R},\ \exists\,c=\mathrm{Cov}(u,\eta)\in\mathbb{R}:\ (2)\text{ holds with }\beta=b.$$

In particular β takes either sign with $\delta = 0$, and $\beta = 0$ is attainable with $\delta \neq 0$. The statement is unchanged under any linear within-transformation M, including country and year fixed effects and country-specific trends.

Proof. Write $C_{Hg} = \mathrm{Cov}(\Delta\log H, g)$ and $V = \mathrm{Var}(R)$, so (2) reads $\beta V = (\gamma-1)C_{Hg} + \delta\,\mathrm{Var}(u) + c$: one equation, three unknowns (γ,δ,c). Given any target (δ,b), set $\gamma = 1$ and $c = bV - \delta\,\mathrm{Var}(u)$; then $\beta = b$. Taking $\delta = 0$, $c = bV$ gives $\beta = b$ of either sign; taking $b = 0$, $c = -\delta\,\mathrm{Var}(u) \neq 0$ gives $\beta = 0$ with $\delta \neq 0$. Equivalently, any orthogonal split of R is admissible, and δ rescales freely as $\mathrm{Var}(u) \to 0$. For the fixed-effects claim let M be symmetric and idempotent, removing country and year means or any further within-country regressors. Then $MR = (\gamma-1)M\Delta\log H + Mu$ and $Mg = \delta Mu + M\eta$, so

$$\beta_M \;=\; \frac{(\gamma-1)\,\mathrm{Cov}(M\Delta\log H,\ Mg)\ +\ \delta\,\mathrm{Var}(Mu)\ +\ \mathrm{Cov}(Mu,\ M\eta)}{\mathrm{Var}(MR)},$$

which is (2) with transformed moments and a *new* free nuisance $\mathrm{Cov}(Mu, M\eta)$. Each M buys one equation and one unknown, and adds no restriction unless $\mathrm{Cov}(u,\eta)$ is itself restricted. □

Three restrictions bear on the proposition; the first two jointly overturn it, the third alone suffices:

- $\gamma = 1$, so AI-skilled and total hiring share a cycle exactly, eliminates the first term, but it is an assumption about the very quantity the index conceals and no published series tests it.
- $\mathrm{Cov}(u,\eta) = 0$, asserting that nothing moving employment also moves AI-specific hiring; however, European hiring cooled and AI-specific hiring fell together over 2022 and 2023.
- Using an instrument that shifts u while moving neither η nor the withheld denominator $\Delta\log H$. We find none in the published data; the likeliest candidates, venture capital and skills migration, may affect employment through channels other than the AI-specific hiring shock, and both are procyclical, so they cannot be assumed to shift u alone.

Section 2 places this decomposition in the ratio literature, leaving the sign of the cyclicality term open. The sign of γ is not determined a priori: adjustment costs make specialist hiring lumpy, arguing for $\gamma < 1$ and a negative β with no displacement at all, while specialist hiring fell further than aggregate hiring in 2022 and 2023, arguing for $\gamma > 1$. Published data cannot separate the two. Figure 2 shows how, with δ held at zero, the coefficient changes sign once $\mathrm{Cov}(u,\eta)$ crosses $-(\gamma-1)\mathrm{Cov}(\Delta\log H,\eta)$. Panel a is not calibrated to the data, so it establishes only that the bias can run in either direction.

Proposition 2 (What disclosure buys, and what it does not)**.** Suppose the provider also publishes $\Delta\log B$, here total platform hiring. Then γ and u are identified, and the coefficient on u in the projection of g on $(\Delta\log H, u)$ equals $\delta + \mathrm{Cov}(u,\eta)/\mathrm{Var}(u)$. The relative-cyclicality channel is removed exactly and γ becomes measurable; δ remains confounded by $\mathrm{Cov}(u,\eta)$ and is still not identified.

Proof. Given R and $\Delta\log H$, the series $\Delta\log h = R + \Delta\log H$ is observed, so γ is the projection coefficient of one observed series on another and u is its residual. Regressing g on $(\Delta\log H, u)$ with $u \perp \Delta\log H$ gives a coefficient on u of $\mathrm{Cov}(u,g)/\mathrm{Var}(u)$, which by (1) equals $\delta + \mathrm{Cov}(u,\eta)/\mathrm{Var}(u)$. □

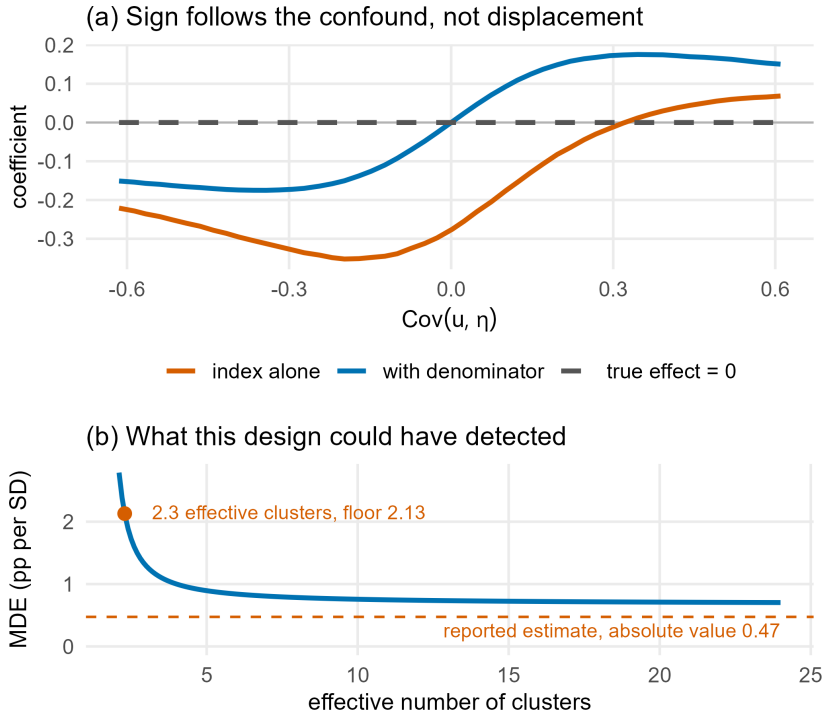

Figure 2: (a) With $\delta = 0$ everywhere, the published-index coefficient changes sign as $\mathrm{Cov}(u, \eta)$ varies; disclosure removes the cyclicality channel only, and over part of the range sits further from the truth (simulation details in Appendix B). (b) The Proposition 3 detection floor at $\sigma = 0.2405$: the minimum detectable effect at 80% power against the effective cluster count, with this design marked at $G^* = 2.3$ (floor 2.13) and the reported estimate 0.47 dashed.

The removal is exact because u becomes the residual of an observed projection. Disclosure does not however shrink the bias: where the channels have opposite signs, removing one enlarges the net error, as Figure 2 shows, so a disclosed index can sit further from the truth than the ratio it replaced while being far more informative about why. One caveat is substantive: the provider publishes rates, not counts, so recovering $\Delta \log H$ also needs the member-base growth that cancels out in the ratio. The required disclosure is therefore the underlying levels. Even with those levels, however, identification is not the only constraint.

Proposition 3 (Detection floor)**.** For a clustered design with standard error σ and G^* effective clusters, the smallest effect detectable at the 5% level with power $1 - \kappa$ is approximately $\mathrm{MDE} = (t_{0.975,\, G^*-1} + t_{1-\kappa,\, G^*-1})\, \sigma$, which diverges as $G^* \to 1$ at fixed σ, irrespective of the number of observations.

G^* is the effective cluster count implied by the concentration of the regression's own influence: it falls below the nominal count whenever a few clusters carry most of the squared score mass, so it is a property of the design: adding observations to the same few clusters does not raise it; only variation spread across more clusters would, and the panel already spans the continent.

4 Results

4.1 Dashboards

Financing is deal-concentrated, though less than the aggregate share implies. IT infrastructure takes 65.21% of European AI venture capital in 2025, but one investor-country cell carries most of that, and the share falls to 27.8% on EU27 member rows, where there is no rebound above the 2021 level. One

robust finding remains: mega-deals were about 73% of global AI investment value [30]. Appendix A gives the series and the EU27 aggregation gap.

The published index increases in 2024, eighteen months after the ChatGPT release [28]. We hold two extracts of the index: the medians here use the refreshed one and Panel A's regressions the other, so agreement between them is corroboration only within one data ecosystem. The median across 27 countries ran 3.30, 1.48, 3.43 over 2018 to 2020, turned negative at -1.31, -4.93, -4.32 over 2021 to 2023, then rose to 6.80 and 7.96. The upturn is common to both extracts and survives changes of country set, means for medians, and no year-averaging. The pre-2021 level does not, so the rise partly recovers the earlier level rather than establishing a new one.

Capital and talent co-move, and employment reallocates. Lagged venture capital goes with a higher hiring index the following year under country and year effects (3.60, SE 1.73, $p = 0.048$), but not without them and not on the industry panel, and the outcome in that model is the index itself, so Proposition 1 bounds it too (Figure 4). Employment shares shift toward IT infrastructure and healthcare and away from the residual. These are shares, so they describe reallocation, and the programming trend predates generative AI [31]. Appendix A gives the coefficients and the shares.

Treating November 2022 as the event date, three of four annual designs reject their pre-trend test; the fourth, whose outcome is employment growth rather than the index, does not ($p = 0.100$) and shows nothing after the date. A shift-share sensitivity agrees, since its bivariate first stage flips sign once controls enter [41, 40, 39]. All four designs reported in Appendix B.

4.2 AI hiring and employment growth

Proposition 1 says a coefficient of either sign is compatible with no displacement. We nonetheless estimate the regression and assess what it can support. The outcome $Y_{i,t}$ is total employment growth in percent year on year, and the baseline is

$$Y_{i,t} = \alpha_i + \lambda_t + \beta_1 \text{Hiring}_{i,t} + \beta_2 \text{Migration}_{i,t} + \beta_3 \text{Tertiary}_{i,t} + \beta_4 \text{Policy}_{i,t} + \varepsilon_{i,t},$$

with errors clustered by country. Twenty-four clusters is too few for reliable asymptotics, so we fix four procedures in advance: CRV1 intervals ($t(23)$)[1]; a wild cluster bootstrap-t with Rademacher weights, null imposed, $B = 99{,}999$, and seed 42 [21, 22]; the same t referred to the effective cluster count; and studentized randomization inference [36, 37]. A 200-variant specification curve leaves 198 estimates negative, but 192 of the 200 perturb one 144-observation regression (Appendix C).

Table 2: Robustness of the relative-AI-hiring coefficient (per 1 SD), Panel A. Intervals are CRV1 $t(23)$; p columns are analytic clustered and wild cluster bootstrap-t (99,999 draws for the baseline, 4,999 elsewhere, one shared draw sequence). Remaining preprocessing rows and full influence and inference diagnostics are in Appendix B. The text reports two further reference distributions [27].

Check	Coefficient	95% interval	p_{analytic}	$p_{\text{bootstrap}}$
Two-way FE, baseline	-0.47	$[-0.97, +0.02]$	0.061	0.037
+ GDP-growth and unemployment controls	-0.46	$[-0.83, -0.09]$	0.017	0.014
Without COVID years ($n = 96$)	-0.45	$[-0.99, +0.09]$	0.097	0.003
Without 2021, the LFS break year ($n = 120$)	-0.45	$[-1.07, +0.18]$	0.157	0.030

Relative AI hiring is the only baseline regressor with a recurring partial association: a 1 SD higher index goes with roughly 0.3 to 0.6 percentage points lower employment growth within a country, while migration, tertiary share and policy score show none. Table 2 reports four specifications; the coefficient stays negative throughout, but the 95% intervals reach or cross zero in most rows. Employment weighting strengthens it to -0.89 per SD ($p = 0.004$) and country trends shrink it to -0.34 ($p = 0.17$). Adding GDP growth and unemployment leaves it unchanged: the channel Proposition 1 formalises runs through total hiring, which neither control measures.

The identifying variation is concentrated in very few clusters based on the decomposition: Türkiye carries 47% of the squared mass on the restricted scores the bootstrap resamples, giving 3.2 effective

[1]The package supports two conventions. CRV1 charges absorbed fixed effects to the degrees of freedom, giving $\text{SE} = 0.241$ and $p = 0.061$, whereas the package default gives $\text{SE} = 0.219$ and $p = 0.041$. We use CRV1 for every reported result.

clusters, while on the unrestricted scores Estonia carries 61% and $G^* = 2.3$. Estonia and Latvia hold the panel's two largest regressor values, $+4.6$ and $+4.4$ standard deviations in 2019, so small-country volatility affects hiring as well as migration. Leave-one-cluster-out agrees: dropping Türkiye moves the baseline from -0.47 ($p = 0.061$) to -0.35, Latvia to -0.50 and Estonia to -0.73 ($p = 0.005$), so single-cluster deletions span -0.35 to -0.73.

As the estimate depends on the cluster, the p-value depends on the procedure. Each $p_{\text{bootstrap}}$ value in Table 2 undercuts its analytic counterpart p_{analytic}, in one row by 0.127, while referring the same t to $t(G^* - 1)$ gives $p = 0.25$ unrestricted and $p = 0.17$ restricted. It might seem that the bootstrap over-rejects and $t(G^* - 1)$ corrects it; however, simulated under the null on this design's own matrix, both the bootstrap and CRV1 are correctly sized, so their gap in Table 2 is unexplained, while $t(G^* - 1)$ rejects in none of the 353 replications as concentrated as the data, so $p = 0.25$ is uninformative, which restates Proposition 3. Appendix B gives the calibration in full, with two cautions about G^*: it is estimator-dependent (5.5 under an alternative weighting, $p = 0.11$) and noisy (median 3.4, 5 to 95% range 1.8 to 7.2). Studentized randomization inference [36, 37] gives $p = 0.045$ over 20,000 draws and is the least sensitive of the four to the error process, though the exchangeability it rests on strains this panel, since the heterogeneity driving the leave-one-out range is what a permutation null denies. Across procedures, the baseline runs from 0.037 to 0.25, but the 0.25 end is uninformative: it comes from a reference distribution on which this design could never have rejected. We do not need the estimate to be insignificant: Proposition 1 bars the displacement reading at every value.

At the reported $\sigma = 0.2405$ and G^* between 2.29 and 5.5, the Proposition 3 floor at 80% power runs from 2.13 down to 0.86, and to 0.70 with all 24 clusters effective (Figure 2b); every value exceeds the 0.47 the primary specification reports. Inverting the noncentral t gives exactly 2.38 and 0.87, so the closed form understates the constraint at the concentrated end. The comparison is not fixed in advance, since σ and G^* are both estimated from the regression being evaluated, so a floor above the estimate is implied by $p > 0.05$. That ratio, 1.8 to 4.5 times the estimate, therefore rescales the null result rather than testing it. Against a benchmark fixed in advance, even the floor's best case, 0.70 points of annual employment growth per standard deviation, exceeds any employment effect documented for earlier automation waves [14],so effects of every previously observed size were undetectable by design.

The natural probe of Proposition 1 is a measure whose denominator is not total hiring [32, 46, 47]. AI talent concentration, whose denominator is platform membership, is published in three encodings on the same 221 observations. Only one has power against the baseline effect (90%), and it is null: $+0.02$ per SD, $p = 0.91$. The other two are underpowered at 70% and 13%, so their negative point estimates are not evidence (Appendix B), and two further probes cannot discriminate at this sample size either (Appendix B). The second check is Panel B: with predictors in their own units and venture capital as $\log(1 + \text{deals})$, the hiring association holds in sign but weakens to -0.027 under two-way FE ($p = 0.059$, bootstrap 0.056), about -0.24 per SD in Panel A. Every 2021 observation carries Eurostat's break flag from the Labour Force Survey redesign, and 2021 growth is measured against the 2020 depression; excluding 2021 moves the coefficient to $+0.002$ ($t = 0.07$) and excluding all break-flags to $+0.003$, while the same exclusion leaves Panel A at -0.45 (Figure 6, Appendix D).

The two extracts of the index are not the same series. On the 208 country-years where they overlap they correlate 0.447; one is positive in 200 of those cells and the other in 119, so they disagree in sign on at least 81. This undocumented divergence is itself evidence: under the two-way demeaning the baseline uses, the extracts correlate 0.528, so the identifying variation behind the -0.47 is only half-reproducible.

Local projections [23] give cumulative associations between -0.36 and -0.62 over horizons 0 to 3, with only $h = 2$ excluding zero under CRV1, and the lead estimate is imprecise but larger than the contemporaneous one, so reverse causality is not excluded (Appendix B).

4.3 Occupations

A displacement hypothesis predicts losses concentrated in occupations more exposed to AI, which the country regressions cannot see. We test it on a country × occupation × year panel of nine ISCO-08 major groups, regressing occupation employment growth on the hiring index interacted with exposure, under country × year, country × occupation and occupation × year fixed effects, clustered by country. Exposure is the index of the published ILO Working Paper [48], 427 unit groups aggregated to nine

with Eurostat employment weights; an unweighted crosswalk and three author-coded rank proxies [24, 25] are robustness encodings.

The country $\times$ year effects absorb every country-level shock in levels, so identification is cross-occupation within country-year, but they do *not* absorb the channel of Section 3.2: substituting the decomposition gives $R_{i,t}X_o = (\gamma - 1)\Delta \log H_{i,t}X_o + u_{i,t}X_o$, and a country-year term times an occupation characteristic survives those fixed effects exactly as the displacement term does. Exposure indices load on cognitive and clerical content, where cyclical sensitivity also differs.

The interaction is negative across five exposure encodings and an employment-weighted variant, ranging from -0.34 to -1.69. Two pass the 5% threshold under CRV1 and one under the wild bootstrap. Excluding the COVID years, the largest economy, or 2021 preserves the sign. With $G^* = 4.7$, referring the statistic to $t(G^* - 1)$ gives $p = 0.16$. Appendix A reports all estimates and p-values.

Fourteen of the 1,458 observations show annual growth above 20 log points, mostly 2021 survey-redesign values; trimming them halves the published-index estimate. The occupational levels, however, run in the opposite direction: growth rises monotonically with exposure in both eras, with the era-over-era gain largest for the *least* exposed tercile (Figure 5), so the negative interaction is a within-country-year co-movement with the index.

5 Limitations

Relative AI hiring measures how fast AI-skilled hiring changes relative to overall hiring, not the number of AI jobs. Skills migration is per 10,000 LinkedIn members, so small-country rates are volatile. The venture-capital data measure financing activity, so large deals can dominate.

Proposition 1 formalises one mechanism under a linear projection. It does not model platform selection or measurement error, and the published index is a smoothed percentage change, so Proposition 2 describes an idealised series. Our δ is a direct effect holding η, and so aggregate hiring, fixed: displacement operating through total hiring loads onto γ, so Proposition 2 is not a route to a total effect, which is not identified either.

Net employment cannot reveal gross displacement [26]: separations, hours, wages and task reallocation may move while headcount does not. Every AI-activity measure here derives from LinkedIn's member base, and the arena exhibit from a second platform with its own sampler, so platform composition can move the indicators independently of the quantity of interest. Our nulls are failures to reject.

6 Discussion

This work asked what a published activity ratio can identify. A regression of any outcome on the index returns the sum of three terms, only one of which is the causal effect, and no fixed-effects design separates them (Proposition 1). Publishing the denominator in levels removes the cyclicality term exactly and leaves the shared shock in place (Proposition 2), and the effective number of clusters bounds what any such design could detect (Proposition 3). Both exhibits show these limits binding: the Arena share halved with the win rate flat because the sampling changed, and the employment association rests on two or three countries, sits below its own detection floor, and reproduces on the 2025 panel only through the 2021 survey break. The dashboards do support concentration, timing and co-movement, as Section 4.1 reports.

More broadly, the decomposition applies wherever an ecosystem is watched through ratios, and each result names the disclosure that would answer it. Proposition 1 sets the disclosure minimum: both counts in levels, at the frequency of the ratio, including the base whose growth cancels in the published rate. The Arena exhibit adds the sampling layer: which units were eligible in each period and how they were sampled, without which a share confounds performance with exposure. Section 4.2 adds versioning: the divergence we document arrived with no published notice, so extracts should be citable by version. Proposition 3 then obliges the analyst rather than the provider: report the effective number of clusters and the smallest detectable effect alongside any estimate built on such an index.

References

[1] OECD.AI Policy Observatory. Flow of VC investments in AI from country of investor to industry and country of start-up. [Online]. Available: https://oecd.ai/en/data?selectedArea=investments-in-ai-and-data&selectedVisualization=flow-of-vc-investments-in-ai-from-country-of-investor-to-industry-and-country-of-start-up

[2] OECD.AI Policy Observatory. VC investments in AI by country. [Online]. Available: https://oecd.ai/en/data?selectedArea=investments-in-ai-and-data&selectedVisualization=vc-investments-in-ai-by-country

[3] OECD.AI Policy Observatory. Between-country AI skills migration. [Online]. Available: https://oecd.ai/en/data?selectedArea=ai-jobs-and-skills&selectedVisualization=between-country-ai-skills-migration

[4] OECD.AI Policy Observatory. Cross-country AI skills penetration by industry. [Online]. Available: https://oecd.ai/en/data?selectedArea=ai-jobs-and-skills&selectedVisualization=cross-country-ai-skills-penetration-by-industry-2

[5] OECD.AI Policy Observatory. AI hiring over time. [Online]. Available: https://oecd.ai/en/data?selectedArea=ai-jobs-and-skills&selectedVisualization=ai-hiring-over-time

[6] *Eurostat*. Persons in full-time and part-time employment by sex, age and occupation, table lfsa_epgais. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/lfsa_epgais/default/table?lang=en&category=labour.employ.lfsa.lfsa_empftpt

[7] *Eurostat*. Employed persons by detailed economic activity (NACE Rev. 2 two-digit level), table lfsa_egan22d. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/lfsa_egan22d/default/table?lang=en

[8] *Eurostat*. Population in private households by educational attainment level, table edat_lfse_03. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/edat_lfse_03/default/table?lang=en

[9] *Eurostat*. Real GDP growth rate, volume, table tec00115. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/tec00115/default/table?lang=en

[10] *Eurostat*. Unemployment by sex and age, annual data, table une_rt_a. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/une_rt_a/default/table?lang=en

[11] *Eurostat*. Employed information and communications technology (ICT) specialists, table isoc_sks_itspt. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/isoc_sks_itspt/default/table?lang=en

[12] *Eurostat*. Graduates in tertiary education in science, mathematics, computing, engineering, manufacturing and construction, table educ_uoe_grad04. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/educ_uoe_grad04/default/table?lang=en

[13] D. H. Autor, F. Levy, and R. J. Murnane. "The skill content of recent technological change: An empirical exploration." *The Quarterly Journal of Economics*, 118(4):1279-1333, 2003.

[14] D. Acemoglu and P. Restrepo. "Robots and jobs: Evidence from US labor markets." *Journal of Political Economy*, 128(6):2188-2244, 2020.

[15] S. Noy and W. Zhang. "Experimental evidence on the productivity effects of generative artificial intelligence." *Science*, 381(6654):187-192, 2023.

[16] E. Brynjolfsson, D. Li, and L. Raymond. "Generative AI at work." *The Quarterly Journal of Economics*, 140(2):889-942, 2025.

[17] T. Eloundou, S. Manning, P. Mishkin, and D. Rock. “GPTs are GPTs: Labor market impact potential of large language models.” *Science*, 384(6702):1306-1308, 2024.

[18] D. Acemoglu. “The simple macroeconomics of AI.” *Economic Policy*, 40(121):13-58, 2025.

[19] A. Humlum and E. Vestergaard. “Still waters, rapid currents: Early labor market transformation under generative AI.” NBER Working Paper 33777, May 2025, revised March 2026. Previously circulated as “Large language models, small labor market effects.”

[20] G. Henseke. “From exposure to adoption: Generative AI in European workplaces.” arXiv:2604.18849 [econ.GN], 2026.

[21] A. C. Cameron, J. B. Gelbach, and D. L. Miller. “Bootstrap-based improvements for inference with clustered errors.” *The Review of Economics and Statistics*, 90(3):414-427, 2008.

[22] J. G. MacKinnon and M. D. Webb. “Wild bootstrap inference for wildly different cluster sizes.” *Journal of Applied Econometrics*, 32(2):233-254, 2017.

[23] Ò. Jordà. “Estimation and inference of impulse responses by local projections.” *American Economic Review*, 95(1):161-182, 2005.

[24] E. W. Felten, M. Raj, and R. Seamans. “Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses.” *Strategic Management Journal*, 42(12):2195-2217, 2021.

[25] P. Gmyrek, J. Berg, and D. Bescond. “Generative AI and jobs: A global analysis of potential effects on job quantity and quality.” ILO Working Paper 96, International Labour Organization, 2023.

[26] S. J. Davis and J. Haltiwanger. “Gross job creation, gross job destruction, and employment reallocation.” *The Quarterly Journal of Economics*, 107(3):819-863, 1992.

[27] Code, data processing, outputs and the full technical log of this project. Provided with this submission as anonymised supplementary material; the public repository will be linked in the camera-ready version.

[28] *OpenAI*. “Introducing ChatGPT,” 30 November 2022. [Online]. Available: `https://openai.com/index/chatgpt/`

[29] *OECD.AI*. “LinkedIn data: AI talent migration and relative AI hiring methodology.” [Online]. Available: `https://oecd.ai/en/linkedin`

[30] *OECD*. “Venture capital investments in artificial intelligence through 2025,” 2026. [Online]. Available: `https://www.oecd.org/en/publications/venture-capital-investments-in-artificial-intelligence-through-2025_a13752f5-en/full-report.html`

[31] *Eurostat*. “ICT specialists in employment,” data extracted April 2026. [Online]. Available: `https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/47162.pdf`

[32] *Eurostat*. “Use of artificial intelligence in enterprises,” 2025 data. [Online]. Available: `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises`

[33] S. Albanesi, A. Dias da Silva, J. F. Jimeno, A. Lamo, and A. Wabitsch. “New technologies and jobs in Europe.” *ECB Working Paper Series*, No. 2831, 2023. [Online]. Available: `https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2831~fabeeb6849.en.pdf`

[34] S. Singh, Y. Nan, A. Wang, D. D’Souza, S. Kapoor, et al. “The leaderboard illusion.” *arXiv preprint* arXiv:2504.20879, 2025. [Online]. Available: `https://arxiv.org/abs/2504.20879`

[35] LMArena. “Arena human preference 100k.” Dataset of 106,134 anonymous pairwise model comparisons, June to August 2024. [Online]. Available: `https://huggingface.co/datasets/lmarena-ai/arena-human-preference-100k`

[36] I. A. Canay, J. P. Romano, and A. M. Shaikh. "Randomization tests under an approximate symmetry assumption." *Econometrica*, 85(3):1013-1030, 2017.

[37] A. Young. "Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results." *The Quarterly Journal of Economics*, 134(2):557-598, 2019.

[38] J. G. MacKinnon, M. Ø. Nielsen, and M. D. Webb. "Cluster-robust inference: A guide to empirical practice." *Journal of Econometrics*, 232(2):272-299, 2023.

[39] J. Roth. "Pretest with caution: Event-study estimates after testing for parallel trends." *American Economic Review: Insights*, 4(3):305-322, 2022.

[40] R. Adão, M. Kolesár, and E. Morales. "Shift-share designs: Theory and inference." *The Quarterly Journal of Economics*, 134(4):1949-2010, 2019.

[41] P. Goldsmith-Pinkham, I. Sorkin, and H. Swift. "Bartik instruments: What, when, why, and how." *American Economic Review*, 110(8):2586-2624, 2020.

[42] A. Georgieff and R. Hyee. "Artificial intelligence and employment: New cross-country evidence." *OECD Social, Employment and Migration Working Papers*, No. 265, OECD Publishing, Paris, 2021. [Online]. Available: `https://doi.org/10.1787/c2c1d276-en`

[43] T. Babina, A. Fedyk, A. He, and J. Hodson. "Artificial intelligence, firm growth, and product innovation." *Journal of Financial Economics*, vol. 151, 103745, 2024.

[44] M. Hampole, D. Papanikolaou, L. D. W. Schmidt, and B. Seegmiller. "Artificial intelligence and the labor market." NBER Working Paper No. 33509, 2025.

[45] E. Brynjolfsson, B. Chandar, and R. Chen. "Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence." Stanford Digital Economy Lab Working Paper, 2025.

[46] OECD.AI Policy Observatory. "AI talent concentration by country." [Online]. Available: `https://oecd.ai/en/data?selectedArea=ai-jobs-and-skills&selectedVisualization=ai-talent-concentration-by-country`

[47] *Eurostat*. "Job vacancy statistics by NACE Rev. 2 activity, annual," table `jvs_a_rate_r2`. [Online]. Available: `https://ec.europa.eu/eurostat/databrowser/view/jvs_a_rate_r2/default/table?lang=en`

[48] P. Gmyrek, J. Berg, K. Kamiński, F. Konopczyński, A. Ładna, B. Nafradi, K. Rosłaniec, and M. Troszyński. "Generative AI and jobs: A refined global index of occupational exposure." *ILO Working Paper* No. 140, International Labour Organization, Geneva, 2025.

[49] D. Guarascio and J. Reljic. "AI and employment in Europe." *Economics Letters*, vol. 247, 112183, 2025.

[50] L. Lebastard and D. Sondermann. "Artificial intelligence: friend or foe for hiring in Europe today?" *ECB Blog*, 4 March 2026. [Online]. Available: `https://www.ecb.europa.eu/press/blog/date/2026/html/ecb.blog20260304~d9e34fc95f.en.html`

[51] I. Aldasoro, L. Gambacorta, R. Pal, D. Revoltella, C. Weiss, and M. Wolski. "AI adoption, productivity and employment: Evidence from European firms." *BIS Working Papers*, No. 1325, Bank for International Settlements, 2026.

[52] M. Cazzaniga, F. Jaumotte, L. Li, G. Melina, A. J. Panton, C. Pizzinelli, E. J. Rockall, and M. M. Tavares. "Gen-AI: Artificial intelligence and the future of work." *IMF Staff Discussion Note* SDN/2024/001, International Monetary Fund, 2024.

[53] L. Nurski and N. Ruer. "Exposure to generative artificial intelligence in the European labour market." *Bruegel Working Paper* 06/2024, Bruegel, Brussels, 2024.

[54] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. I. Jordan, J. E. Gonzalez, and I. Stoica. "Chatbot Arena: An open platform for evaluating LLMs by human preference." In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235:8359-8388, 2024.

[55] A. V. Carter, K. T. Schnepel, and D. G. Steigerwald. "Asymptotic behavior of a t-test robust to cluster heterogeneity." *The Review of Economics and Statistics*, 99(4):698-709, 2017.

[56] A. Agrawal, J. Gans, and A. Goldfarb, editors. *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press for the National Bureau of Economic Research, Chicago, 2019.

[57] J. Kleinberg and M. Raghavan. "Algorithmic monoculture and social welfare." *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

[58] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. "Performative prediction." In *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:7599-7609, 2020.

[59] M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. "Strategic classification." In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111-122, 2016.

[60] K. Pearson. "Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurement of organs." *Proceedings of the Royal Society of London*, 60:489-498, 1897.

[61] R. A. Kronmal. "Spurious correlation and the fallacy of the ratio standard revisited." *Journal of the Royal Statistical Society, Series A*, 156(3):379-392, 1993.

[62] E. Kuh and J. R. Meyer. "Correlation and regression estimates when the data are ratios." *Econometrica*, 23(4):400-416, 1955.

[63] D. A. Belsley. "Specification with deflated variables and specious spurious correlation." *Econometrica*, 40(5):923-927, 1972.

[64] J. Aitchison. "The statistical analysis of compositional data." *Journal of the Royal Statistical Society, Series B*, 44(2):139-177, 1982.

[65] *Eurostat*. "Employment by sex, age and occupation (2-digit ISCO-08)," table `lfsa_egai2d`. [Online]. Available: `https://ec.europa.eu/eurostat/databrowser/view/lfsa_egai2d/default/table?lang=en`

A Dashboard descriptives and occupation results

This appendix reports the arithmetic behind the descriptive statements and the full occupation results. These estimates document concentration, co-movement, and reallocation; none identifies the numerator-specific employment effect defined in Section 3.2.

Venture-capital concentration. Table 3 shows why the headline European aggregate is sensitive to its geographic base. The country field in this extract denotes the investor side of the flow. The United Kingdom by IT-infrastructure cell alone accounts for approximately 54% of the 2025 all-country total. Moreover, the provider's EU27 aggregate is approximately 2.8 times the sum of the member-state rows. We cannot reconcile those two series, so we report both rather than silently selecting one.

Capital, hiring, and employment. In the 2018-2025 descriptive panel, lagged log AI venture capital is positively associated with the following year's relative-AI-hiring index when country and year effects are included (coefficient 3.60, SE 1.73, $p = 0.048$). The association is absent without those effects (-0.22, $p = 0.48$), and lagged own-industry venture capital does not predict industry employment growth (-0.388, $p = 0.299$). These specifications therefore establish co-movement, not a capital-to-employment mechanism.

Table 3: European AI venture capital by geographic base (USD millions).

Geographic base	2021 total	2024 total	2025 total	IT share, 2025
All country rows	41,074	23,654	62,015	65.2%
Excluding United Kingdom	18,225	13,170	18,740	37.1%
EU27 member-state rows	14,837	10,611	13,835	27.8%
Provider's EU27 aggregate row	42,039	29,996	39,505	27.2%

Table 4: Occupation-exposure interaction estimates.

Exposure specification	$\hat{\theta}$	p_{CRV1}	p_{WCB}	G^*	N
Published ILO index, weighted crosswalk	-0.618	0.093	0.195	4.71	1,458
Published ILO index, unweighted crosswalk	-0.657	0.080	0.163	4.66	1,458
Published ILO index, WLS by lagged employment	-0.336	0.376	0.440	3.05	1,458
AIOE-anchored rank proxy	-0.892	0.030	0.043	4.76	1,458
ILO-anchored rank proxy	-0.717	0.082	0.194	4.23	1,458
Cognitive versus manual	-1.691	0.026	0.063	5.81	1,296

On the 33 countries observed at both endpoints, employment shares between 2018 and 2025 change by $+0.960$ percentage points for IT infrastructure and hosting, $+0.904$ for health care, drugs, and biotechnology, and -1.479 for the residual category. Pooling a changing country set would instead give $+0.845$, $+0.450$, and -0.774 points, respectively. We use the balanced figures because the United Kingdom is 11.7% of pooled 2018 employment and is absent in 2025.

On the industry panel, penalised and boosted learners evaluated by leave-one-country-out cross-validation do not improve on least squares (RMSE 0.226 for the lasso and 0.239 for gradient boosting against 0.224 for OLS and 0.278 for the mean baseline); the lasso penalty is tuned within each training fold and the boosting rounds by country-grouped inner cross-validation.

Occupation interaction. The estimating equation is

$$\Delta \log E_{iot} = \theta \left(R_{it} X_o \right) + \alpha_{it} + \mu_{io} + \lambda_{ot} + \varepsilon_{iot},$$

where R_{it} is standardised relative AI hiring and X_o is occupation exposure. The fixed effects are country-year, country-occupation, and occupation-year; standard errors are clustered by country. Continuous exposure measures are standardised. The cognitive/manual contrast is binary. Table 4 distinguishes five exposure encodings from the third row, which is a lagged-employment-weighted regression using the same published ILO exposure measure.

The wild-cluster-bootstrap values in Table 4 use 99,999 Rademacher draws with the null imposed and seed 42; their Monte Carlo standard errors range from 0.0006 to 0.0016. The published ILO estimate is sensitive to the 14 of 1,458 cells with $|\Delta \log E| > 20$ log points: trimming those cells moves the coefficient from -0.62 to -0.33 ($p_{\mathrm{WCB}} = 0.34$). Winsorising at ± 20 gives -0.49 ($p_{\mathrm{WCB}} = 0.26$). For the author-coded AIOE rank, the corresponding trimmed and winsorised estimates are -0.42 and -0.66, with bootstrap p-values 0.065 and 0.043. Excluding the COVID years, the largest economy, or 2021 preserves the sign of the published-index interaction but gives CRV1 p-values 0.047, 0.080, and 0.103, versus bootstrap p-values 0.191, 0.174, and 0.201. Estonia contributes 37.9% of the unrestricted squared-score mass in the primary occupation specification.

A separate post-2022 design estimates -0.405 per SD for the interaction of the published ILO index with the post indicator ($p = 0.194$, $N = 1{,}502$). Interacting the published ILO exposure measure with national AI-hiring intensity gives -0.040 ($p = 0.824$, $N = 1{,}286$). These estimates do not provide independent evidence of displacement.

B Auxiliary tests and inference diagnostics

Remaining preprocessing checks. Table 5 completes the preprocessing rows summarised in Table 2. The sign is stable, while inference remains sensitive to the reference distribution. All wild-cluster-bootstrap values use 4,999 null-imposed Rademacher draws.

Table 5: Additional preprocessing checks for the Panel A hiring coefficient.

Specification	Estimate	95% CRV1 interval	p_{CRV1}	p_{WCB}	N
Winsorise employment growth	-0.453	$[-0.927,\ 0.021]$	0.060	0.039	144
Winsorise hiring index	-0.562	$[-1.130,\ 0.006]$	0.052	0.048	144
Asinh migration	-0.481	$[-1.003,\ 0.042]$	0.070	0.050	144
Control for survey break	-0.476	$[-0.976,\ 0.024]$	0.061	0.037	144

Table 6: Denominator-independent measures and attainable power.

Measure (per SD)	Estimate	SE	p	N	Power	MDE_{80}
Enterprise AI use, common sample	-0.443	0.545	0.425	96	0.13	1.596
Talent concentration, annual change	$+0.017$	0.141	0.908	221	0.90	0.411
Talent concentration, growth rate	-0.153	0.184	0.414	221	0.70	0.537
Talent concentration, level	-0.347	0.556	0.539	221	0.13	1.622

Measures whose denominator is not total hiring. Table 6 reports the denominator-independent probes. Power is calculated against an effect of magnitude 0.474 per SD, using each row's standard error and a two-sided $t(G-1)$ test. The enterprise-AI-use comparison is not informative: on its 96-observation common sample, relative AI hiring is itself null ($p = 0.655$), and the enterprise-use regression has only 13% power against the Panel A estimate. The annual change in talent concentration is the informative comparison: it has 90% power and is close to zero. The three encodings are reported together to avoid choosing the only well-powered one after seeing the estimates.

Adding Eurostat's annual job-vacancy rate does not discriminate between the competing mechanisms. On the Panel A vacancy subsample the hiring coefficient moves from -0.359 ($p = 0.081$, $N = 131$, 22 countries) to -0.366 ($p = 0.085$). On the Panel B vacancy subsample it moves from -0.0300 ($p = 0.057$, $N = 120$, 24 countries) to -0.0287 ($p = 0.054$).

Panel B and the 2021 break. Panel B retains the negative sign in the full 2021-2025 sample but does not reproduce it after removing the survey-redesign year. Coefficients in Table 7 are in raw index points. The full-sample bootstrap p-value is 0.0557 from 99,999 null-imposed Rademacher draws.

Dynamics. Local projections estimate cumulative associations of -0.471, -0.361, -0.615, and -0.367 at horizons zero through three. The corresponding CRV1 p-values are 0.063, 0.252, 0.041, and 0.161, while the sample falls from 144 to 72 observations. The one-year lead is -0.528 with a 95% interval $[-1.386, 0.330]$. The non-monotone path, shrinking samples, and imprecise lead do not support a dynamic causal interpretation and do not exclude reverse causality.

Hierarchical model and flexible learners. A Bayesian hierarchical version of the baseline gives a posterior median of -0.43 per SD for the hiring coefficient with 90% interval $[-0.69, -0.19]$ ($\widehat{R} \leq 1.005$, no divergent transitions), matching the least-squares estimate. A random forest of 500 trees evaluated by leave-one-country-out cross-validation attains RMSE 1.83 against 1.87 for the linear specification and 1.95 for predicting the mean, so flexible learners recover no nonlinear signal the linear model misses.

Influence concentration. Let q_g be country g's score contribution for the hiring coefficient and let $s_g = q_g^2 / \sum_h q_h^2$. We report

$$G^* = \frac{1}{\sum_g s_g^2}$$

as an inverse-Herfindahl diagnostic. It should not be presented as a general degrees-of-freedom theorem. In this fixed-effects design, the equicorrelated-cluster construction of Carter, Schnepel and Steigerwald [55] is unavailable: the within-cluster vector of ones it is built on is annihilated by the country effects. The value reported here is therefore a squared-score plug-in. For the baseline, unrestricted scores give $G^* = 2.29$, with Estonia carrying 61.1% of squared-score mass; null-restricted scores give 3.22, with Türkiye carrying 47.4%. Weighting countries by the within-cluster

Table 7: Panel B sensitivity to the 2021 Labour Force Survey break.

Sample	Estimate	Clustered SE	p	N	Countries
Full, 2021-2025	-0.0274	0.0138	0.0588	130	26
Excluding 2021	$+0.0016$	0.0219	0.9427	104	26
Excluding break-flagged observations	$+0.0029$	0.0217	0.8961	102	26

Table 8: Leave-one-country-out influence for the Panel A hiring coefficient.

Estimation sample	Estimate	CRV1 SE	p	N
Full sample	-0.474	0.241	0.061	144
Drop Türkiye	-0.350	0.191	0.080	138
Drop Latvia	-0.497	0.344	0.163	138
Drop Estonia	-0.727	0.233	0.005	138
Drop Türkiye and Latvia	-0.312	0.254	0.232	132

sum of squares of the residualised hiring regressor instead gives 5.52. These alternatives show that the exact scalar is estimator-dependent even though all diagnose severe concentration. Table 8 reports the corresponding leave-one-country-out estimates.

The four reported reference distributions answer different questions and rest on different assumptions. The conventional CRV1 $t(23)$ reference gives $p = 0.0609$; the null-imposed wild-cluster bootstrap gives $p = 0.0366$; and studentized randomization inference over 20,000 permutations of whole-country hiring trajectories gives $p = 0.0450$. The latter requires country trajectories to be exchangeable conditional on the fixed-effect structure. Referring the same statistic to $t(G^* - 1)$ gives $p = 0.253$ using unrestricted scores and $p = 0.175$ using restricted scores, but those values are sensitivity diagnostics rather than established finite-sample tests. Using the within-cluster-regressor weighting gives $p = 0.112$.

Design-matched calibration. Table 9 holds the observed design matrix fixed, imposes a zero hiring effect, and draws errors independently within countries. The first row allows country-specific variances calibrated from the restricted residuals; the second uses a common variance. This is a favourable calibration, not a model of arbitrary serial correlation. It therefore cannot validate any procedure for a general country panel.

Among the 353 heteroskedastic replications with score concentration at least as severe as observed ($G^* \leq 2.29$), rejection rates are 0.011 for CRV1, 0.031 for the wild bootstrap, and zero for $t(G^* - 1)$. The plug-in itself is noisy: under this calibration its median is 3.38 and its 5th-95th percentile range is 1.81-7.21, while the observed 2.29 lies near the 17th percentile. The simulations therefore show that no single one of the four p-values should be treated as dispositive.

At the baseline CRV1 standard error of 0.2405, the closed-form 80%-power detection floor is 2.13 per SD at $G^* = 2.29$, 1.19 at the restricted-score value 3.22, 0.86 under the within-cluster-regressor weighting 5.52, and 0.70 if all 24 countries were equally effective. Exact inversion of the noncentral t gives 2.38 and approximately 0.87 at the two ends of the diagnostic range. These calculations describe the current allocation of identifying variation. Additional observations concentrated in the same influential countries need not improve the floor; adding countries or more balanced within-country variation can.

Simulation behind Figure 2. The numerical illustration fixes the true numerator-specific effect at zero, sets $\gamma = 0.6$, and draws 200,000 observations at each point of the grid. It uses

$$R = (\gamma - 1)\Delta \log H + u, \qquad g = 0.8\Delta \log H - 0.5s + \varepsilon,$$

with independent standard-normal $\Delta \log H$ and s, Gaussian ε with standard deviation 0.3, and $u = v + \rho s$ residualised on $\Delta \log H$. The grid is $\rho \in [-2.5, 2.5]$. The index-only coefficient ranges from -0.35 to $+0.07$ even though the true effect is zero; after observing the denominator, the coefficient on u ranges from -0.18 to $+0.18$. Because changing ρ changes both $\mathrm{Cov}(u, \eta)$ and $\mathrm{Var}(u)$, the plotted curves are nonlinear. The exercise is an existence illustration, not a calibrated claim about the empirical frequency or magnitude of either bias.

Table 9: Rejection rates at a nominal 5% level (2,000 replications).

Error process	CRV1 $t(23)$	Wild bootstrap	$t(G^* - 1)$	Randomization
Country-heteroskedastic	0.051	0.054	0.014	0.041
Homoskedastic	0.067	0.046	0.032	0.041

Table 10: Analysis samples.

Block	Years	Countries	N	Construction
Panel A country-year	2019-2024	24	144	Frozen Panel A hiring and migration extracts; primary country panel.
Panel B country-year	2021-2025	26	130	Refreshed hiring and migration extracts; includes 2025.
Occupation interaction	2019-2024	27	1,458	Nine ISCO-08 major groups; country-year, country-occupation, and occupation-year effects.
Post-2022 occupation design	2017-2025	27	1,502	Published ILO index; observations touching a Eurostat "b" break removed.
Descriptive country/industry panels	2018-2025	varies	varies	OECD.AI activity and venture-capital series joined to Eurostat employment.

Event-study and shift-share diagnostics. The annual joint pre-trend test rejects for the AI-hiring level under employment exposure ($p = 0.010$), the change in AI hiring under employment exposure ($p = 0.015$), and the AI-hiring level under digital-core exposure ($p = 0.024$). It does not reject for total employment growth under employment exposure ($p = 0.100$), whose post-2022 coefficients are $+0.436$, $+0.014$, and $+0.313$ and individually insignificant. A separate monthly AI-hiring event study rejects its joint pre-trend test ($p = 0.017$). There are four annual designs and one separate monthly design.

The shift-share first stage is $+0.377$ ($p = 0.025$, $R^2 = 0.023$) in a bivariate regression and -0.790 ($p = 0.031$) after adding the reported controls. The reduced-form coefficients are -11.276 for AI-hiring growth ($p = 0.051$) and $+1.154$ for employment growth ($p = 0.267$); the resulting IV coefficient of AI hiring on employment growth is -0.043 ($p = 0.905$). The first-stage sign reversal and rejected pre-trends preclude a causal reading.

C Specification curve for the primary coefficient

Figure 3 evaluates the defined specification set used in the supplement: 192 Panel A combinations of controls, sample restrictions, employment weights, winsorisation, and country-specific trends, together with eight Panel B combinations of the 2021 restriction, vacancy control, and employment weights. Panel B coefficients are rescaled to percentage points per sample SD of the hiring index. All specifications include country and year effects and use country-clustered CRV1 intervals with $t(G - 1)$ critical values.

Of the 200 point estimates, 198 are negative. Their median is -0.533 and their range is $[-1.513, +0.097]$; the declared Panel A primary estimate (-0.474) is at the 62nd percentile. These summaries describe sign and magnitude sensitivity only. The specifications are highly dependent, were not a set of independently preregistered tests, and the display supplies no familywise or selective-inference correction. Accordingly, counts of rows with $p < 0.05$ are intentionally not used as evidence.

D Sources, samples, and additional figures

Sample construction and versioning. Table 10 records the samples used by the reported empirical blocks. Panel A is rebuilt from frozen raw inputs; Panel B is independently joined from the later root-level extracts. OECD.AI omits covered country-years with zero recorded venture-capital deals, so the scripts fill those cells with zero; a country absent from the extract entirely remains missing. Türkiye is the only such case in Panel A and has no venture-capital value in any of its six years.

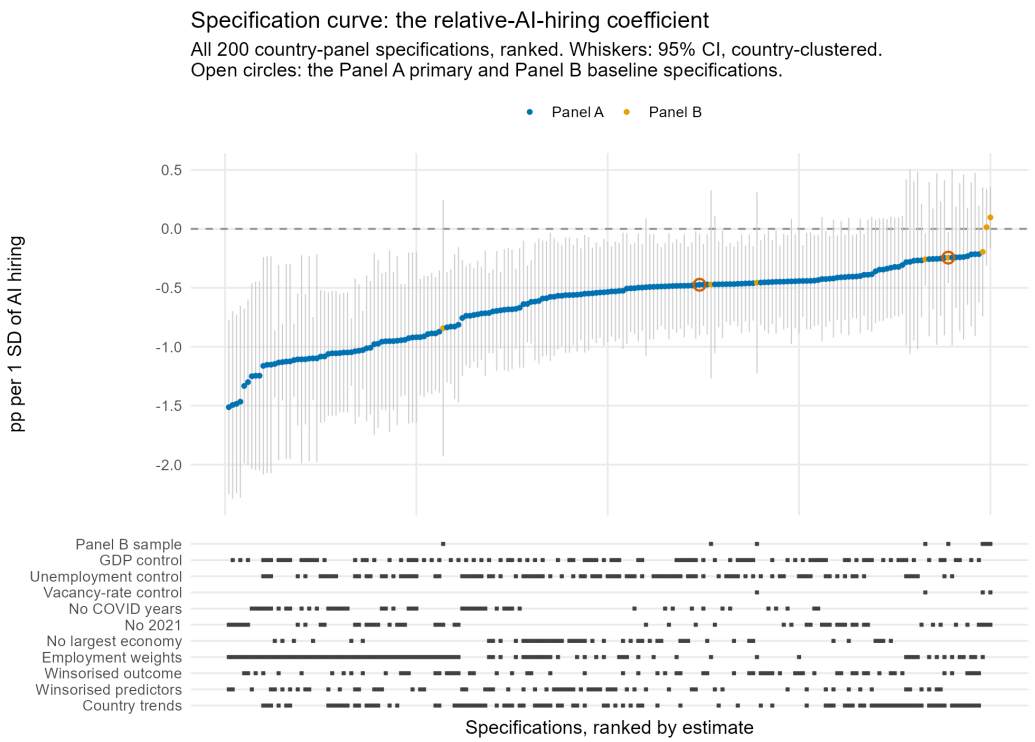

Figure 3: Relative-AI-hiring coefficients across 200 country-panel specifications. Whiskers are 95% country-clustered CRV1 intervals using $t(G-1)$ critical values. Open circles mark the Panel A primary and Panel B baseline specifications.

The two annualized relative-AI-hiring files are not interchangeable. On 208 overlapping country-years, the Panel A and refreshed extracts correlate 0.447; 200 Panel A cells are positive versus 119 refreshed cells. The former has mean 15.24 and standard deviation 13.22, while the latter has mean 1.67 and standard deviation 7.72. The exact files are `data/panel_a/input/ai_hiring_index.csv` (truncated SHA-256 `65594a5d8ffff942`) and `data/ai_hiring_index.csv` (`54df61bf6e9e3719`). We found no public version history or methodological notice explaining the revision. The package's `data/MANIFEST.csv` records file hashes, providers, and redistribution status; Panel A's frozen inputs were exported on 13 August 2026 and are mapped to their source objects in `data/panel_a/PROVENANCE.md`.

The Arena battle log is not redistributed because it declares no licence. `scripts/fetch_arena_battles.R` downloads it, and the manifest records the truncated SHA-256 `d4fb8013b0306624`. All randomized procedures set seed 42; the baseline country-panel and occupation bootstraps use 99,999 draws, and the randomization-inference calculation uses 20,000 permutations.

Additional figures. The intervals in Figure 4 and Figure 7 use the estimation package's default small-sample correction, as those figures belong to the descriptive pipeline. The primary country-panel tables, Figure 3, and Figure 6 use the explicit CRV1 convention described in the main text. Stating the convention here avoids treating the package default and the paper's CRV1 calculation as the same standard error.

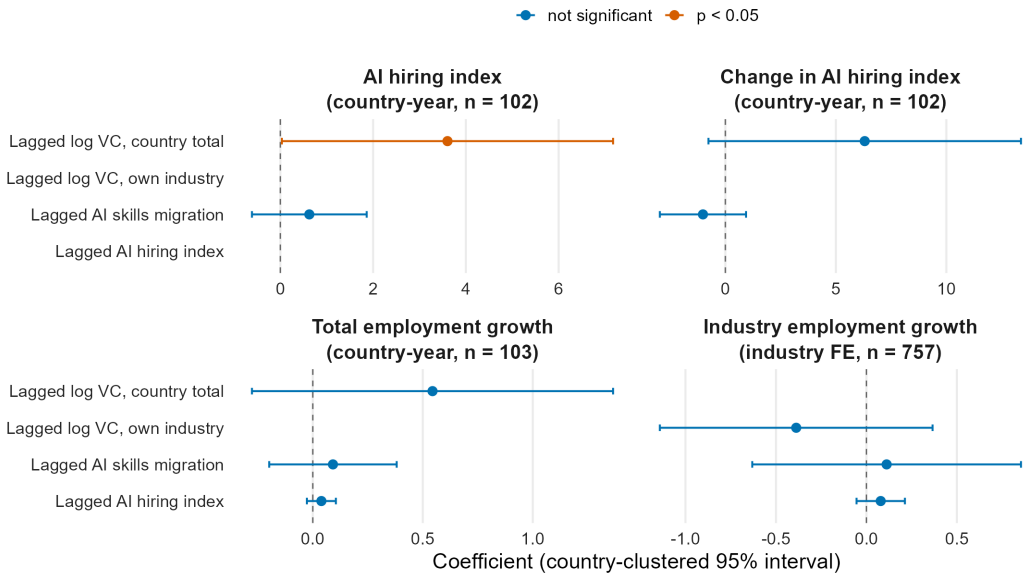

Figure 4: Four fixed-effects co-movement regressions for 2022-2025. The first three are country-year models; the fourth is a country-industry-year model with industry effects. Intervals are 95% country-clustered intervals using the estimation package's default small-sample correction. Only lagged country-total venture capital in the hiring-level regression excludes zero.

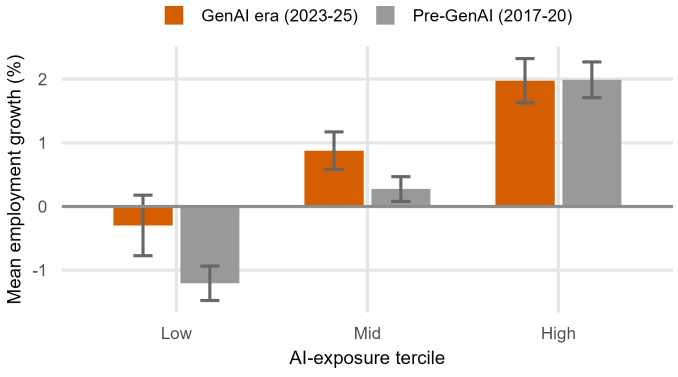

Figure 5: Mean occupation employment growth by tercile of the published ILO exposure measure before generative AI (2017-2020) and in 2023-2025. Bars are one standard error across country means. Growth is higher in both eras for the most exposed tercile, and the era-over-era increase is largest for the least exposed tercile.

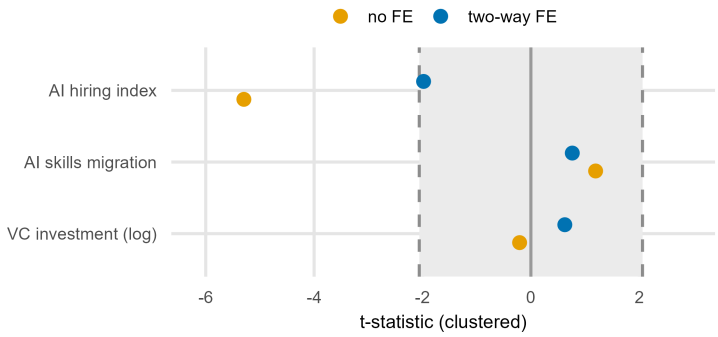

Figure 6: Panel B (26 countries, 2021-2025, $N = 130$): clustered t-statistics for the three baseline predictors with and without country and year effects. Dashed lines mark two-sided 5% critical values based on $t(25)$. The full-sample hiring coefficient is near, but does not cross, the threshold under two-way effects.

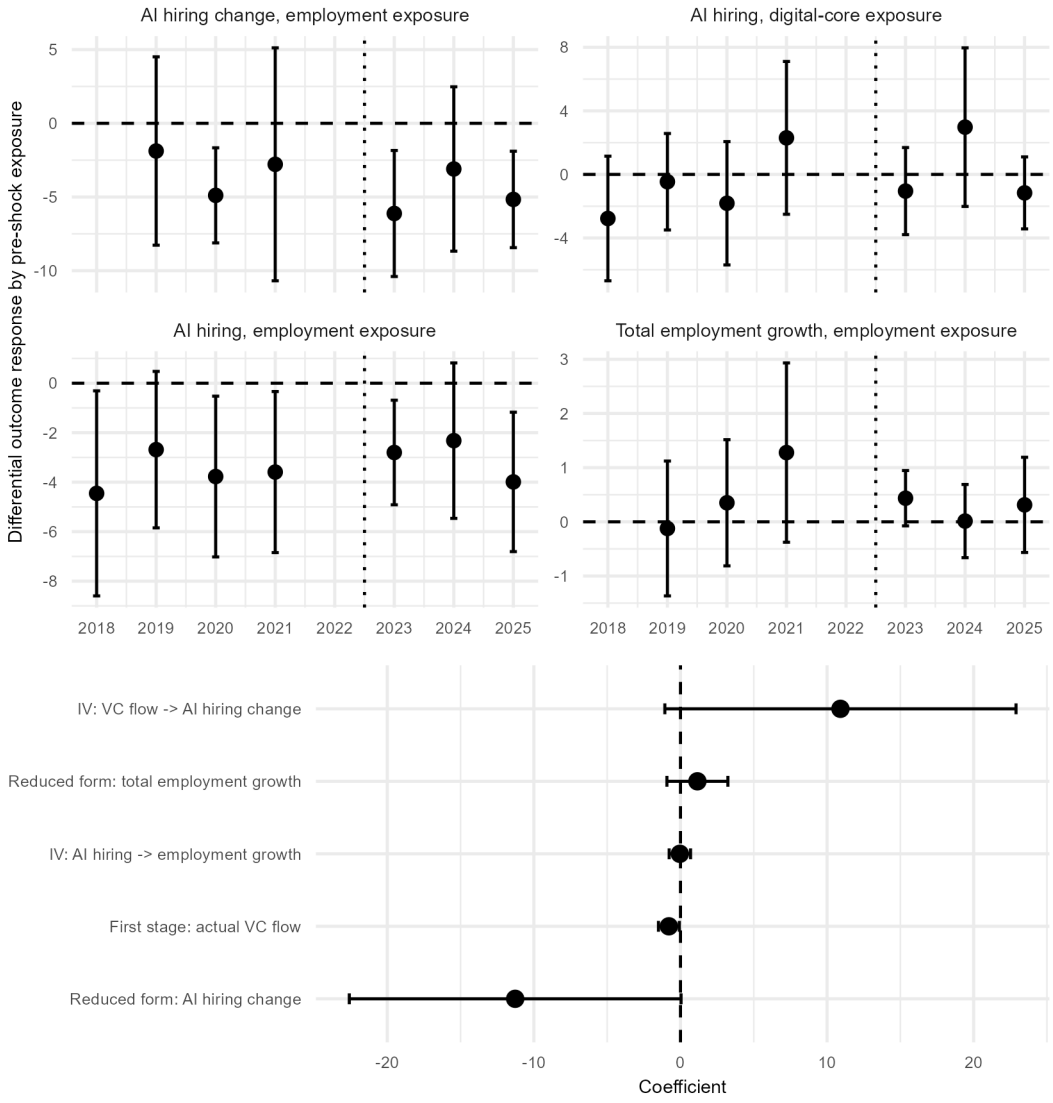

Figure 7: Identification diagnostics. Top: annual event studies interacting year with 2018 employment exposure, with 2022 omitted. Three of four annual specifications reject the joint pre-trend null. Bottom: shift-share first stage, reduced forms, and IV estimates; the instrument uses 2018 main-industry employment shares interacted with leave-one-out European AI venture-capital main-industry shocks, and the first-stage coefficient changes sign when controls enter. Intervals use the estimation package's default country-clustered small-sample correction. These failures motivate a descriptive, not causal, interpretation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the three formal results (Section 3.2), the descriptive findings (Section 4.1) and the employment association (Sections 4.2 and 4.3), and say up front that the paper measures the instrument rather than estimating AI's employment effect.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations section covers what the series measure, the linear-projection and idealised-index assumptions, and the fact that our estimand is a direct effect, so no total effect is identified even after disclosure.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All three results are in Section 3.2, with the linear projection and the outcome equation (1) stated immediately above Proposition 1; Propositions 1 and 2 are proved there; Proposition 3 follows in one line from the definition of the minimum detectable effect under a $t(G^* - 1)$ reference distribution, and Appendix B reports the exact noncentral-t version.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3.1 and Table 1 define every sample and source, and the text states the specifications, clustering, degrees-of-freedom convention and every seed and draw count. The replication package regenerates every reported number from frozen inputs, the one exception being the Arena battle log, which a shipped script downloads and `MANIFEST.csv` fingerprints.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymised replication package ships all scripts, frozen inputs, outputs and a README giving the exact commands and environment, and `CLAIMS.md` maps each claim in the paper to the file and column behind it.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Estimating equations, fixed effects, clustering and all four inference procedures are in Sections 4.1 to 4.3; the leave-one-country-out benchmark states its split and tree count in Appendix B, and Appendix A how the penalty and boosting rounds were tuned.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Every estimate carries country-clustered standard errors and intervals, and the primary specification is reported under four reference distributions fixed in advance, which place it between $p = 0.037$ and $p = 0.25$. Figure captions state what each interval represents.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All analyses are small-panel regressions, resampling and one small random forest on a single consumer CPU with no GPU; the README gives measured wall-clock, the longest script about eighteen minutes. The pipeline was rebuilt in full several times, so the project used more compute than the paper reports.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research analyses published aggregate statistics, involves no human subjects or personal data, and preserves anonymity in the submission and supplement.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Discussion cautions against displacement readings the evidence cannot support. The sharpest risk is self-inflicted: the introduction quotes the association the paper argues cannot be read causally, and Proposition 1 bars that reading at every p-value we report.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No model weights are released and nothing scraped or crowdsourced is redistributed, so nothing here requires gated or controlled release.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Every input is cited in Table 1 and the bibliography, and `data/MANIFEST.csv` records provenance and licence per file: Eurostat under CC BY 4.0, OECD.AI extracts not relicensed, and the Arena log, which declares no licence, fetched rather than shipped. The released code is MIT.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new assets are the analysis code, released under MIT, and two author-constructed inputs: the policy score, whose construction rule and components are in `build_panel.R`, and the occupation rank proxies, described in the README. No asset derives from identifiable people, so no consent was required.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: We recruited, instructed and compensated no one. The Arena battle log is third-party crowdsourced data, already anonymised and publicly released, which we analyse rather than collect.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research was conducted, so no institutional review was required and no participant risk arises.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM assistance was limited to writing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.