# Corporate AI Adoption and Firm-level Outcomes

## Evidence from Earnings Call Disclosures

Jonathan Rice[1] Giulia Maria Guerrini[2]

[1]Central Bank of Ireland [2]European Central Bank

11th Bank of Finland and European Systemic Risk Board Joint Conference on AI and Systemic Risk
4th June 2026

We thank Frederik Andersen and Lisa Chardon-Denizot for excellent research assistance, and Markus Schwedeler (NL Analytics) for the provision of the earnings-call data.

# What happens to firms when they adopt AI?

- Task-level field experiments find large GenAI productivity gains in writing, customer support, coding and consulting.
- Firm-level outcomes need not move the same way: adjustment costs, organisational frictions, lagged output responses, and intangible complementarities.
- The launch of ChatGPT in late 2022 changed both the *capability frontier* (more labour-substituting tasks accessible) and the *pace of adoption* – so the firm-level pattern may have shifted with it.

## The empirical question

Do task-level gains show up in firm accounts – and has the answer changed since ChatGPT?

# From task gains to firm outcomes

The unresolved question is whether local GenAI gains show up in firm accounts.

**What is well established**

**Task RCTs find large, local gains**

+15% in customer support; writing 40% faster and 18% higher quality; coding field RCTs +26% completed tasks; consulting gains only inside the AI frontier.

Brynjolfsson, Li & Raymond (2025); Noy & Zhang (2023); Cui et al. (2025); Dell'Acqua et al. (2023)

**Earlier firm studies are expansionary**

AI investment or adoption raises sales, employment or labour productivity, often through product innovation or capital deepening.

Babina et al. (2024); Aghion et al. (2025); Aldasoro et al. (2026)

**What remains unsettled**

**High use, muted realised effects**

Recent survey and worker evidence points to high AI use but muted realised effects so far, especially on employment, hours and measured productivity.

Yotzov et al. (2026); Humlum & Vestergaard (2025)

**Why accounts may lag**

GPT technologies may move slowly through accounts because data, software, workflow redesign and training are complementary intangible investments.

Brynjolfsson, Rock & Syverson (2021); McElheran et al. (2025)

## This paper

We measure sustained AI deployment directly from earnings calls and link it to matched quarterly firm outcomes for 13,331 global listed firms, 2010Q1–2025Q4.

Measurement builds on earnings-call text methods in Hassan et al. (2019) and Eisfeldt, Schubert & Zhang (2023).

# What we do, and what we find

## What we do

- Build a firm-quarter panel of corporate AI deployment from earnings call transcripts, 2010Q1–2025Q4. Two-stage classifier separates substantive deployment from generic mentions; manually validated.
- Merge with quarterly financials (S&P Capital IQ).
- Propensity-score matched design; static TWFE and Sun–Abraham (2021) IW event study.

## Headline finding – the pattern changed after ChatGPT

- **Pre-2022Q4:** adopters' *revenue* and *revenue per employee* fall (each by $\sim$1.5 pp/yr) while *employment* is statistically unchanged. Consistent with firms investing in AI and absorbing a near-term output cost without cutting headcount – the early phase of a productivity J-curve.
- **Post-2022Q4:** *employment* growth slows by $\sim$0.7 pp/yr ($p = 0.02$); the revenue gap closes; *revenue per employee* flips positive (marginal). Consistent with GenAI capabilities becoming more labour-substituting and the first signs of productivity gains.

# Two data sources, merged at the firm level

**Earnings call transcripts**
(NL Analytics)

- 340,361 calls; 15,000+ unique firms, 90+ countries
- Management presentation + analyst Q&A
- Identifiers: company name, CIQ ID, ISIN, ticker, date

**Quarterly financials**
(S&P Capital IQ)

- Revenue, employees, assets, EBITDA, margins, debt
- YoY log growth in employment, revenue, rev/employee
- Match rate to call data $> 99\%$

## Resulting panel

13,331 firms / 292,598 firm-quarters / 2010Q1–2025Q4
2,756 ever-sustained adopters | 9,021 never-treated

# Measuring AI adoption from earnings-call text

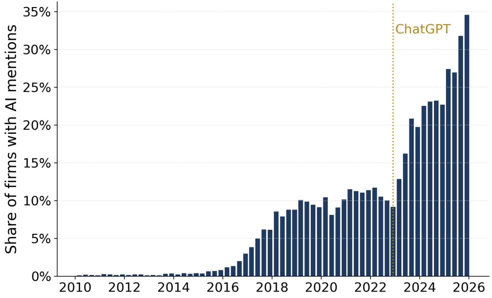
Two-stage classifier on every sentence of every call: keyword filter, then LLM with confidence score. Aggregated to firm-quarter.

1. **Boolean screen** — Search calls for AI-investment language; collect candidate sentences.
2. **Manual labels** — Hand-check a random sample with surrounding sentences for context.
3. **Prompt tests** — Six OpenAI models $\times$ three prompts on the manual sample.
4. **Final classifier** — GPT-4.1-mini with the final confidence-scored prompt.
5. **Full corpus run** — Classify corpus; aggregate to firm-quarter; build sustained measure.

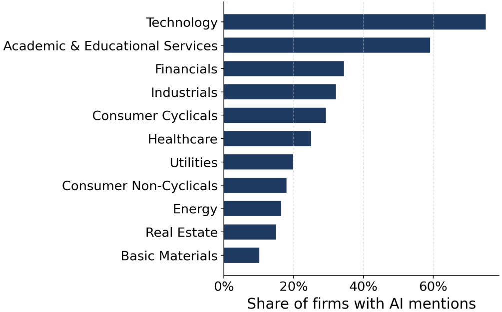
**Validation against 500 hand-labelled sentences:** overall accuracy 87.2%; at the extremes ($\leq 29$ and $\geq 90$ confidence) agreement is 94–100%. **Sustained adoption:** treatment switches on after two consecutive AI-positive quarters; absorbing thereafter.

Appendix: keyword list, production prompt, full validation chart, pipeline diagram, worked examples.

# Adoption: rare before 2018, broad-based today



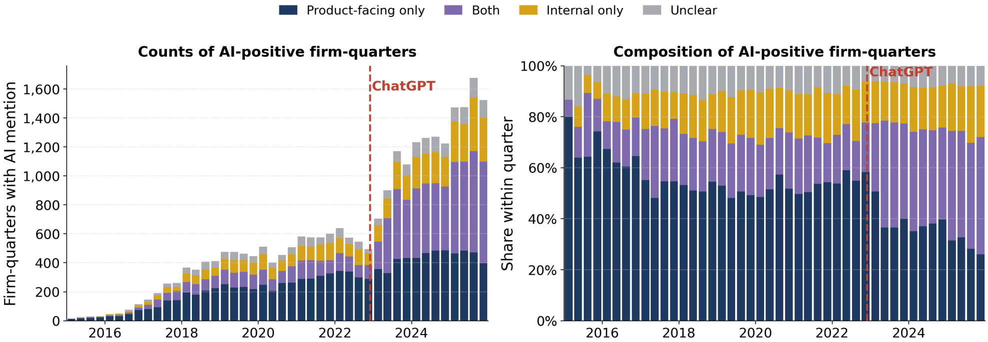

Quarterly share of listed firms with AI mentions.




Share with AI mentions in 2025Q4, by economic sector.

Mentions rare before 2018; they accelerate sharply after ChatGPT (Nov 2022) and by end-2025 reach $\sim$35% of listed firms. Technology leads at $\sim$75%, but every sector now sits above 10% – AI has moved beyond the technology sector.

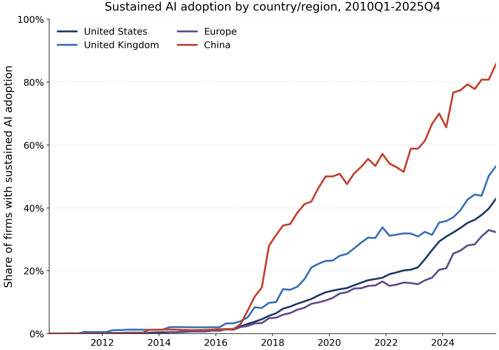
# Use type: from product-facing to internal, post-ChatGPT



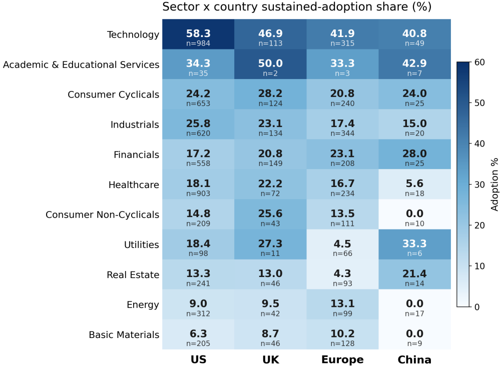

Use type from a separate LLM pass on each AI-positive call. *Product-facing* = customer-facing features / new AI products; *Internal* = back-office, ops, productivity tools; *Both* = the call discusses both; *Unclear* = ambiguous.

**Volume:** AI-positive firm-quarters grow from ∼15 in 2015Q1 to ∼1,500 in 2025Q4. **Mix:** 2015 was dominated by product-facing firms (80%); by end-2025 the modal category is *Both* (46%) and *Internal* use has risen to ∼20%. The post-ChatGPT shift into internal applications is consistent with the labour-substitution channel we estimate later.

# Adoption is uneven across countries



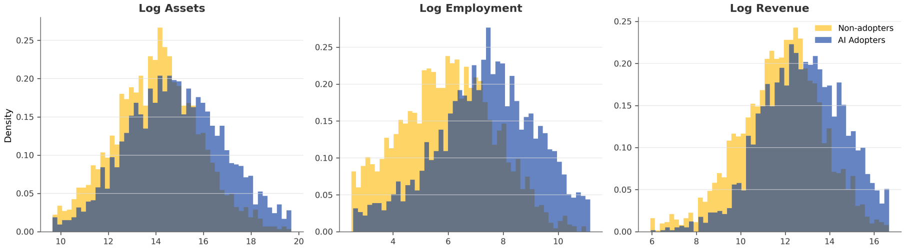

Share of firms with sustained AI adoption, by country / region.




Unique-firm sustained-adoption share (%), sector × country / region.

China leads sustained adoption (∼86% by end-2025), well above the UK (∼53%), US (∼43%) and Europe (∼32%). The heatmap shows Technology leading everywhere (US 58%, UK 47%, Europe 42%, China 41%); Academic & Educational Services is the highest non-Tech cell (UK 50%, China 43%, US 34%). Europe = continental EU + EFTA, ex-UK.

# Adopters differ from non-adopters *before* they adopt

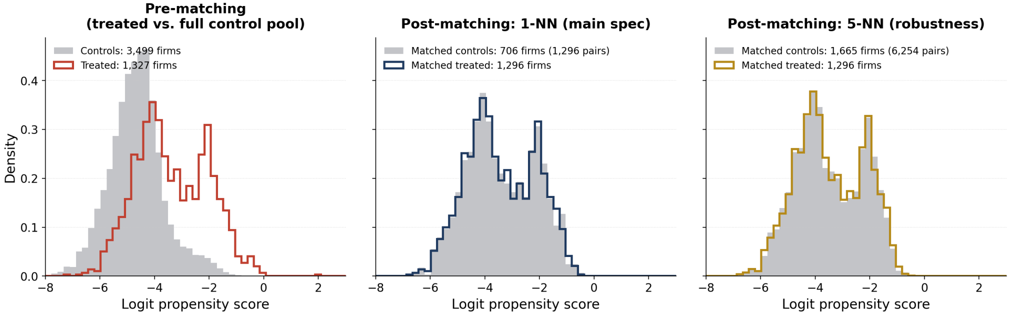



Pre-treatment distributions of log assets, log employment and log revenue. Adopters in red, non-adopters in grey.

Adopters are systematically larger and faster-growing than non-adopters. Standardised mean differences exceed 0.5 on most size measures and reach 0.52 on log employment (0.22 on log revenue, 0.16 on log assets).

$\Rightarrow$ An unmatched DiD would conflate the treatment effect with this pre-existing size/growth advantage. We address it with propensity-score matching.

# Matching restores covariate balance and pscore overlap




Propensity-score distributions, logit scale. Pre-matching: treated firms vs. the full control pool. Post-matching: 1-NN (main) and 5-NN (robustness) matched panels.

**Before matching:** treated firms have a fat right tail (high predicted adoption probability); the full control pool does not. **After 1-NN matching:** the matched-control distribution aligns with treated across the support; max SMD = 0.091, mean = 0.040 across 15 covariates – all below the 0.10 threshold. 5-NN gives similar balance with a larger donor pool.

# Empirical strategy

Match each sustained adopter to a never-treated control observed at the same calendar quarter, in the same broad sector, with the closest pre-treatment profile.

- **Event quarter ($G_i$).** First quarter in which the treated firm has two consecutive AI-positive calls. Matched control assigned the same $G_i$, so pre- and post-windows align.
- **Pre-treatment covariates ($t = G_i - 8..G_i - 1$):** log revenue, log assets, log employment, EBITDA / operating / gross margins, SG&A share, debt / assets, current ratio, 4q and 8q mean and trend in employment and revenue growth.
- **Matching:** 1-NN (main) and 5-NN (robustness) PSM on the linearised score within sector, caliper 0.25 SD.

**Static specification:**

$$Y_{ipt} = \alpha_i + \gamma_t + \theta\, \mathbf{1}\{\text{Treated}_{ip}\}\, \mathbf{1}\{t \geq G_p\} + \varepsilon_{ipt}$$

**Sun–Abraham (2021) event study:**

$$Y_{ipt} = \alpha_i + \gamma_t + \sum_{k \in [-6,-2] \cup [0,12],\, k \neq -1} \beta_k\, \mathbf{1}\{\text{Treated}_{ip}\}\, \mathbf{1}\{t - G_p = k\} + \varepsilon_{ipt}$$

Firm and quarter FE; SE clustered by firm.

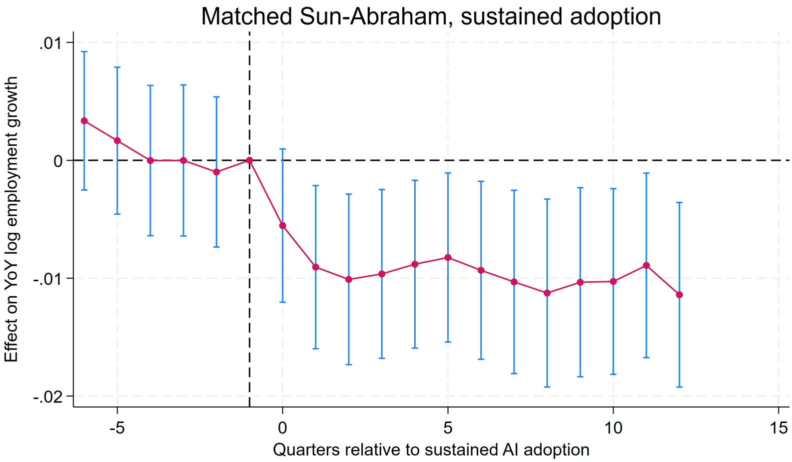
# Main matched results

Static post-vs-pre and Sun–Abraham IW post-period mean ($k = 0$–12). 1-NN matched panel (main spec) and 5-NN (robustness). Firm and quarter FE; SE clustered by firm.

| | Static matched | | Sun–Abraham IW | |
|---|---|---|---|---|
| Outcome | 1-NN (main) | 5-NN | 1-NN (main) | 5-NN |
| Employment (log YoY) | $-0.0143^{***}$ | $-0.0138^{***}$ | $-0.0095^{***}$ | $-0.0091^{***}$ |
| | (0.0038) | (0.0034) | (0.0029) | (0.0028) |
| Revenue (log YoY) | $-0.0204^{***}$ | $-0.0217^{***}$ | $-0.0114^{**}$ | $-0.0121^{***}$ |
| | (0.0054) | (0.0053) | (0.0047) | (0.0046) |
| Rev / employee | −0.0062 | −0.0080 | −0.0017 | −0.0028 |
| | (0.0052) | (0.0051) | (0.0044) | (0.0044) |
| EBITDA margin | −0.0231 | −0.0191 | 0.0063 | 0.0076 |
| | (0.0147) | (0.0140) | (0.0099) | (0.0098) |
| Debt-to-assets | 0.0056 | 0.0070 | 0.0069 | $0.0079^{*}$ |
| | (0.0060) | (0.0056) | (0.0043) | (0.0041) |
| Matched treated firms | 1,296 | 1,296 | 1,296 | 1,296 |

**Read:** sustained AI adoption is followed by lower employment and revenue growth across all four columns; rev / employee flat on average. Sun–Abraham IW shrinks magnitudes vs. static TWFE but signs and significance are unchanged. These are averages across the 2018–2026 window – the next slides show the SA event-time profile and how the pattern shifts before vs. after ChatGPT.

# Event study: clean pre-trends, persistent post-adoption decline




Sun–Abraham IW coefficients on log employment growth (YoY), matched panel. 95% CIs, firm-clustered SEs.

Pre-adoption leads small and individually insignificant (joint Wald $p = 0.475$). Effect drops to $\sim -1$ pp within a year of adoption and persists through $k = 12$.

# Before vs after ChatGPT: outcomes split by calendar window

Re-run the Sun–Abraham IW estimator on the *same 1-NN matched panel*, but separately on firm-quarter observations from before vs. after 2022Q4 (the quarter ChatGPT was released).

| Outcome (yr-on-yr) | Pre-genAI window (outcome $tq <$ 2022Q4) | Post-genAI window (outcome $tq \geq$ 2022Q4) |
|---|---|---|
| Employment growth | −0.0034 (0.0042) | −0.0069** (0.0030) |
| Revenue growth | −0.0175** (0.0070) | +0.0073 (0.0084) |
| Rev. per employee | −0.0144** (0.0064) | +0.0159* (0.0087) |
| EBITDA margin | −0.0043 (0.0126) | +0.0217* (0.0117) |
| Debt / assets | +0.0003 (0.0063) | +0.0002 (0.0032) |
| Matched treated firms | 1,296 | 1,296 |
| Firm-quarter obs. | 82,442 | 28,855 |

Each column's post-period mean ($k = 0$–$12$) is identified from the (cohort, event-time) cells whose calendar quarter falls inside that window; joint-test SEs via `lincom`. Pre-trend $p$-values per outcome and the precise statement of which cohorts identify each column are in Appendix A11.

# Findings: a J-curve before GenAI; labour substitution after

**Pre-2022Q4: investment phase, no headcount cuts**

- Adopters absorb a revenue and productivity hit ($\sim$1.5 pp/yr each); employment flat.
- Fits the productivity J-curve (Brynjolfsson, Rock & Syverson, 2021): GPTs need data, workflows and intangible complements before gains show in accounts.
- Near-term cost in pursuit of long-run returns – not labour-shedding.

**Post-2022Q4: labour substitution emerges**

- Employment growth slows $\sim$0.7 pp/yr; revenue gap closes; rev/employee flips positive (marginal).
- Coincides with GenAI capabilities expanding into more labour-substituting tasks and with adopters shifting toward internal use.
- First signs of productivity gains alongside slower labour growth – consistent with GenAI substituting for labour and with the J-curve upswing beginning.

# Systemic risk implications

**From firm-level evidence to systemic risk**

**Labour market and income channels**

- Adoption is becoming labour-substituting – employment growth slows while productivity gains begin to emerge.
- A widening wage-productivity gap can weaken household debt-servicing capacity (mortgages, consumer credit) and tilt income toward capital, compressing aggregate demand.
- Persistent divergence risks greater reliance on leverage and debt-financed consumption.

**Concentration and fiscal channels**

- Large firms absorb AI adjustment without measurable headcount cuts; mid-sized firms cut labour – a pattern that can accelerate market concentration.
- Greater concentration raises the correlation of corporate exposures across bank and non-bank portfolios, reducing effective diversification.
- Fiscal pressures may build if governments expand income support, retraining or transfer programmes in response to labour displacement.

---

*Systemic risk may emerge less from productivity growth itself, and more from a transition characterised by increasingly concentrated income, profits and financial exposures.*

# Conclusion

- A new earnings-call measure of corporate AI deployment for global listed firms, 2010Q1–2025Q4.
- Matched-design results: the firm-level pattern is very different before and after ChatGPT.
- Before ChatGPT: adopters absorb a revenue and productivity dip without cutting labour – an early J-curve pattern.
- After ChatGPT: employment growth slows, the revenue gap closes, and productivity starts to turn – labour-substitution channel beginning.
- Whether the productivity upswing strengthens is a question the panel can keep answering as it extends.

# Thank you

Jonathan Rice – jonathan.rice@centralbank.ie
Giulia Maria Guerrini – giulia.guerrini@ecb.europa.eu

Comments welcome. Dataset & code at ricejonathan.com.

# Appendix

Measurement details, additional robustness, identification notes

## A1. Boolean keyword list (Stage 1)

| Category | Terms |
|---|---|
| General AI terminology | ai, artificial intelligence, generative-ai, genai, gen-ai, machine learning, deep learning, natural language processing, nlp, computer vision, augmented intelligence, aigc |
| Foundation & generative models | generative model*, foundation model*, large language model*, llm*, diffusion model*, transformer model* |
| Specific LLM families | ChatGPT, GPT, Claude, Gemini, PaLM, Bard, Llama, Falcon, Mistral, Mixtral, Jurassic, Jamba, Gemma, Grok, ERNIE Bot, Qwen, Tongyi Qianwen, Yi, Phi, Command R/R+, PPLX, DeepSeek |
| Image generation | DALL-E, Stable Diffusion, Midjourney, Firefly, Imagen, Ideogram |
| Video generation | Runway, Pika, Sora, Veo |
| Application terms | text-to-image, text-to-video, code completion, retrieval-augmented generation, chatbot*, copilot, co-pilot, prompt engineering |
| Technical vocabulary | neural network*, reinforcement learning, supervised learning, unsupervised learning, backpropagation, hyperparameter*, hallucination*, supercomput*, quantum comput* |

“ * ” denotes a wildcard. Filter applied at the sentence level on lower-cased text; flagged sentences are then passed with their preceding and following sentence as context to the LLM.

# A2. Production prompt sent to GPT-4.1-mini (Stage 4)

**Role.** You are an expert in analysing earnings call transcripts for references to artificial intelligence (AI).

**Context.** The text comes from an earnings call and can be either (a) a management statement, or (b) an analyst question directed at the company. When company representatives speak, “we” and “our” refer to the company. When analysts speak, “you” and the company name refer to the company.

**Task.** Classify the text as:

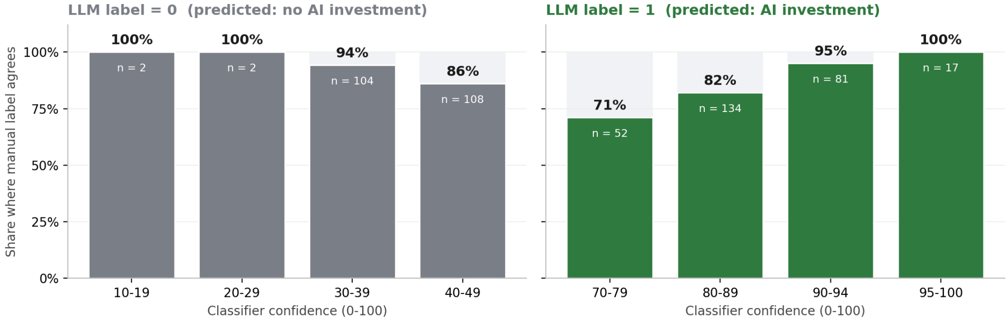
- **1 (AI employed):** the company is *currently* deploying, using, implementing, integrating, piloting, or investing in AI or AI-powered tools in its operations, products, or strategy. Proprietary tools or internal applications count. Concrete hiring for AI projects counts.
- **0 (AI not employed):** general commentary about AI without evidence the company itself is deploying or investing in AI; or future plans that will not materialise within three months.

**Confidence score.** Return an integer from 0 to 100 representing how strongly the snippet suggests evidence of AI employment. Confidence must be consistent with the label: $> 50$ if label $= 1$, $< 50$ if label $= 0$.

**Output.** Strict JSON: `{"label": 0 or 1, "confidence": 0-100}`.

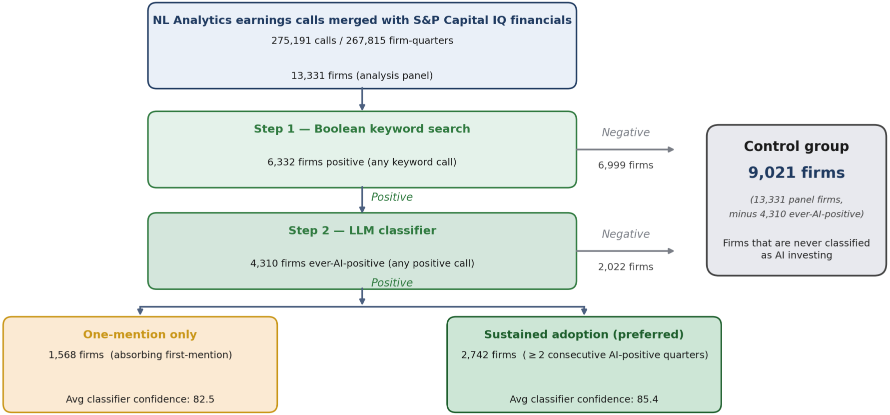
# A3. Validation: 500 manually labelled sentences

Accuracy per confidence cohort: share of LLM predictions in each bin that match the hand-coded label. Negatives (predicted label = 0) on the left; positives (predicted label = 1) on the right. Bin size $n$ shown inside each bar.




At the extremes ($\leq 29$ and $\geq 90$) the LLM and the manual labels agree 94–100% of the time. Disagreements concentrate just above the threshold (70–79 bin, $n = 52$, 71%) and just below it (40–49 bin, $n = 108$, 86%). Overall accuracy 87.2%.

# A4. From raw earnings calls to AI-adopting firms




Two-stage classifier (boolean keyword → LLM). Treatment output split into one-mention and sustained adoption; classifier-confidence summary is the firm-level mean confidence across each firm's AI-positive quarters.

# A5. What counts and what doesn't

## Counted as adoption

- *"Recently, we deployed new AI-driven linehaul models, designed to improve freight flows across our network."*

  XPO Inc. – Industrials

- *"A gen AI based SEWA bot pilot was launched for customer service team query resolution."*

  Max Financial Services

- *"He is driving AI adoption across our R&D organisation [...] measuring how many of our engineers are using AI tools."*

  Hinge Health

## Not counted

- *"AI is top of mind for everyone today."*

  GEE Group – generic

- *"Welcome to the SES AI first-quarter 2025 earnings call."*

  Firm name contains "AI"

- *"The rapid proliferation of artificial intelligence is an extremely attractive opportunity for us."*

  Infineon – aspirational

---

Use type (product-facing, internal, both, unclear) is identified by a separate LLM pass on positives. Treatment is **absorbing** once sustained.

# A6. Treatment definition: sustained adoption

Two variables from the firm-quarter classifier output.

- **Absorbing first-mention.** Treatment switches on at the firm's first AI-positive quarter; remains on thereafter.
- **Sustained (preferred).** Treatment switches on the first quarter in which the firm has *two consecutive* AI-positive quarters; remains on thereafter.

**Why sustained?**

- Filters one-off mentions driven by analyst questions or one-quarter strategic comments.
- Picks up firms for whom AI is a recurring management agenda.
- Cleaner pre-trend diagnostics than absorbing first-mention.

2,756 firms become ever-sustained; first sustained dates fall on or after 2015Q3.

## A7. 1-NN diagnostics: covariate balance and event-time pre-trend

1-NN matched panel (1,296 pairs). Left: standardised mean differences for each of the 15 pre-treatment covariates, before and after matching. Right: Sun–Abraham IW employment coefficient at each event time.

**Standardised mean differences (1-NN)**

| Covariate | SMD pre | SMD post |
|---|---|---|
| log Employment | +0.508 | +0.010 |
| log Revenue | +0.208 | −0.007 |
| Debt / assets | −0.166 | **−0.092** |
| log Assets | +0.145 | −0.039 |
| Employment growth, 4q-8q trend | −0.063 | **+0.062** |
| Operating margin | +0.061 | **+0.077** |
| EBITDA margin | +0.044 | **+0.074** |
| Gross margin | +0.044 | −0.010 |
| *All 15 covariates:* | | |
| Mean $\lvert$SMD$\rvert$ | 0.095 | 0.040 |
| Max $\lvert$SMD$\rvert$ | 0.508 | 0.092 |

**SA IW per event-time (employment)**

| Event time $k$ | Coef. | SE | $p$ |
|---|---|---|---|
| *Pre-period leads* | | | |
| −6 | +0.0033 | (0.0030) | 0.263 |
| −5 | +0.0017 | (0.0032) | 0.601 |
| −4 | −0.0000 | (0.0033) | 0.995 |
| −3 | −0.0000 | (0.0033) | 0.996 |
| −2 | −0.0010 | (0.0032) | 0.761 |
| −1 (reference) | 0.0000 | — | — |
| *Post-period summary* | | | |
| Mean coefficient | −0.0095 | | |
| Range across $k = 0$–12 | −0.0114 to −0.0055 | | |
| *Joint test* | | | |
| $F$-test on all 5 leads | $F(5, 1\,290) = 0.93$, $p = 0.475$ | | |

Post-matching SMDs $\geq 0.05$ shown in bold (none above 0.10). All five leads $k \in [-6, -2]$ individually insignificant ($p \geq 0.26$); joint Wald test $p = 0.475$.

# A8. Heterogeneity by asset tercile (SA IW)

Sun–Abraham IW post-period mean ($k = 0$–$12$), 1-NN matched panel, treated firms partitioned by asset tercile (within the matched sample).

| Outcome | Small | Mid | Large |
|---|---|---|---|
| Employment | $-0.0072$ | $\mathbf{-0.0187}^{***}$ | $-0.0048$ |
| Revenue | $-0.0015$ | $-0.0183^{**}$ | $-0.0120^{**}$ |
| Rev/emp | $+0.0045$ | $+0.0014$ | $-0.0071$ |
| Debt/assets | $+0.0008$ | $+0.0293^{***}$ | $-0.0060$ |
| Pretrend $p$ (emp) | 0.901 | 0.239 | 0.760 |
| $N$ treated firms | 281 | 480 | 535 |

Visible employment effect is largest in mid-asset firms; large firms appear to absorb the adjustment without measurable headcount change. Revenue effect is concentrated in mid and large firms.

# A9. Robustness: excluding US tech and tighter confidence thresholds

Sun–Abraham IW post-period mean ($k = 0$–$12$), 1-NN matched panel.

**Excluding US tech**

| Outcome | All | US tech | Excl. US tech |
|---|---|---|---|
| Employment | −0.0095*** | −0.0135** | −0.0076** |
| Revenue | −0.0114** | −0.0239*** | −0.0069 |
| Rev/emp | −0.0017 | −0.0115 | +0.0015 |
| Pretrend $p$ | 0.475 | 0.963 | 0.126 |

**Tighter classifier-confidence thresholds (static)**

| Outcome | Preferred | 80+ conf. | 90+ conf. |
|---|---|---|---|
| Employment | −0.0143*** | −0.0149*** | −0.0274*** |
| Revenue | −0.0204*** | −0.0293*** | −0.0400*** |
| Rev/emp | −0.0062 | −0.0150** | −0.0118 |
| Matched pairs | 1,296 | 944 | 633 |

**Read.** Employment decline survives once US tech firms are dropped (effect roughly halves but remains significant); revenue decline is concentrated in US tech / scale. Tightening the classifier confidence threshold makes both effects *larger* – consistent with measurement noise attenuating the baseline.

# A10. 5-NN matching diagnostics

5-NN enlarges the donor pool while preserving balance; headline coefficients survive under SA IW with very similar magnitudes.

| | 1-NN (main) | 5-NN (robustness) |
|---|---|---|
| *Balance* | | |
| Matched treated firms | 1,296 | 1,296 |
| Mean $\lvert$SMD$\rvert$ (15 covs) | 0.040 | 0.044 |
| Max $\lvert$SMD$\rvert$ (15 covs) | 0.091 | 0.107 |
| *SA IW pre-trend* (joint Wald, $k \in [-6, -2]$) | | |
| Employment $p$ | 0.475 | 0.457 |
| Revenue $p$ | 0.390 | 0.435 |
| *SA IW post-period mean* ($k = 0$–$12$) | | |
| Employment (log YoY) | $-0.0095^{***}$ | $-0.0091^{***}$ |
| Revenue (log YoY) | $-0.0114^{**}$ | $-0.0121^{***}$ |

Both specs pass the conventional pre-trend bar. Max SMD on 5-NN sits just above the 0.10 rule-of-thumb because the extra controls come from slightly less-similar firms; mean SMD remains well within tolerance.

# A11. The calendar-window split, in detail

Same 1-NN matched-pair pool and same set of cohorts $G$ in both windows. Each column's post-period mean is identified only from (cohort, event-time) cells whose calendar quarter falls inside that window.

**Which cohorts contribute where**

- *Very early adopters* ($G <$ 2018Q4): all $k = 0..12$ lags fall pre-2022Q4 $\Rightarrow$ **pre-window only**.
- *Mid cohorts* (2018Q4 $\leq G <$ 2022Q4): *split* – early lags to pre-window, late lags to post-window. E.g. a 2020Q1 adopter has $k = 0..10$ in pre, $k = 11..12$ in post.
- *Late adopters* ($G \geq$ 2022Q4): all $k = 0..12$ lags fall in 2022Q4+ $\Rightarrow$ **post-window only**. A 2023Q4 first-time adopter contributes only pre-treatment leads ($k = -6, -5$) to the pre-window regression and its full $k = 0..12$ to the post-window.

**Cleaner than splitting firms by cohort.** Every matched pair is in both regressions, so balance and identification are unchanged; pre-trend test runs separately within each window.

**Magnitudes.** Windowed IW estimates re-weight cohorts; the informative comparison is post vs. pre, not vs. the headline.

**Pre-trend $p_{\text{min}}$**
($k \in [-6, -2]$, by window)

| Outcome | Pre | Post |
|---|---|---|
| Employment growth | 0.368 | 0.111 |
| Revenue growth | 0.208 | 0.200 |
| Rev. per employee | 0.213 | 0.095 |
| EBITDA margin | 0.200 | **0.030** |
| Debt / assets | **0.010** | 0.236 |

Bold marks $p < 0.05$ – pre-trend flag. Headline employment / revenue / rev-per-emp results pass cleanly in both windows.

## A12. Caveats

- **Disclosure $\neq$ adoption.** Earnings-call mentions are a managerial signal; classification error attenuates the effect. Confidence-threshold robustness suggests measurement noise is attenuating.
- **Selection on unobservables.** Matching addresses observable selection. Unobserved restructuring plans coinciding with AI disclosure could still confound.
- **Macro confounders, post-2022.** Rate hikes, the technology-sector correction, broader cost discipline. The US-tech exclusion check helps but does not fully resolve this.
- **Listed firms only.** Cannot speak to small private firms or aggregate effects via reallocation.

# Corporate AI Adoption and Employment: Evidence from Earnings Call Disclosures*

Jonathan Rice† Giulia Maria Guerrini‡

March 2026

**Abstract**

We construct a novel firm-quarter panel linking AI adoption disclosures in earnings calls to financial outcomes for over 13,000 global public firms from 2010 through end-2025, using an LLM-based classifier to distinguish genuine AI investment announcements from generic commentary. To address the strong positive selection of AI adopters on pre-treatment characteristics, we employ a cohort-specific propensity score matching design that groups treated firms by adoption timing and matches each to observably comparable controls using rolling pre-treatment windows. On the matched sample, AI adoption is associated with approximately one percentage point lower annual employment growth, an effect that builds gradually over the three years following adoption. Revenue per employee shows no significant improvement in the baseline, though suggestive gains emerge among firms with more persistent AI adoption. The employment effect is substantially larger among firms adopting during the generative AI wave beginning in late 2022 than among earlier adopters. Naïve comparisons that ignore selection find no employment effect, highlighting how much the matching design matters for identification.

**Keywords:** Artificial intelligence, corporate investment, employment, labour productivity, difference-in-differences

**JEL Classification:** G31, J23, O33, M51

*The views expressed in this paper are those of the authors and do not necessarily reflect the views of the European Systemic Risk Board, the European Central Bank, or any of their affiliated institutions. We thank Frederik Andersen for excellent research assistance and [acknowledgments placeholder] for helpful comments and suggestions. All remaining errors are our own.

†European Systemic Risk Board.

‡European Central Bank.

# 1 Introduction

How does corporate investment in artificial intelligence affect employment? Despite its importance, systematic firm-level evidence remains scarce. Task-based models of automation predict that AI should reduce labour demand for automatable activities while potentially creating new tasks that reinstate workers (Acemoglu and Restrepo, 2018). A growing body of field experiments documents large task-level productivity gains from generative AI tools, with improvements of 15–40% in settings ranging from customer support to software development (Brynjolfsson et al., 2025b; Noy and Zhang, 2023; Peng et al., 2023). Whether these task-level gains translate into measurable firm-level outcomes, and if so whether they manifest as employment growth, employment displacement, or productivity improvement, is an open empirical question.

We address this by constructing a novel firm-quarter panel that links AI adoption disclosures in earnings calls to comprehensive financial outcomes for over 13,000 global public firms from 2010 through end-2025. We use an LLM-based classifier to identify firms making substantive AI investments from earnings call text, distinguishing genuine adoption from generic industry commentary. The classifier also identifies the type of AI use, distinguishing customer-facing products from internal operations or both. Linking these adoption signals to quarterly financial data from S&P Capital IQ yields a panel of approximately 293,000 firm-quarter observations covering 4,300 AI-adopting firms and 9,000 non-adopters across more than 90 countries.

Our main finding is that AI adoption is associated with lower employment growth among adopting firms relative to observably comparable controls. On the matched sample, employment growth declines by approximately 1.1 percentage points per year ($p = 0.003$). The effect is robust across alternative matching specifications and is stronger when restricting to high-confidence AI classifications or to sustained treatment requiring consecutive mentions ($-1.9$ pp, $p = 0.0001$). Event-study estimates show no pre-treatment divergence in employment trajectories, followed by a decline that becomes larger and statistically significant at longer post-adoption horizons. Revenue per employee, a labour productivity proxy, shows no significant improvement in the baseline specification ($+0.9$ pp, $p = 0.24$), though it becomes marginally significant under sustained treatment ($+1.6$ pp, $p = 0.088$).

When we split the sample into firms that first disclose AI adoption before 2023 and those that do so during the generative AI wave, the employment effect is substantially larger among the later cohort ($-1.7$ pp, $p = 0.001$) compared with earlier adopters ($-0.5$ pp, insignificant). This pattern is consistent with generative AI capabilities in language, code generation, and content creation being more directly substitutable for white-collar labour than earlier corporate AI applications.

Selection is central to the analysis. AI-adopting firms are systematically larger, more profitable, and faster-growing before adoption, with pre-treatment standardised mean differences exceeding one standard deviation on some covariates. A naïve comparison of adopters and non-adopters therefore conflates the treatment effect with the superior baseline characteristics of adopting firms. In our data, the simple unmatched difference-in-differences estimate for employment growth is close to zero and statistically insignificant. The negative employment effect emerges only after cohort-specific propensity score matching removes the observable selection channel, underscoring how much selection matters when studying technology adoption effects.

We make three contributions. First, we provide new evidence on firm-level AI effects using a dataset that extends through the generative AI era. Most prior studies examine periods ending before 2020; our panel runs through end-2025, covering the rapid diffusion of generative AI capabilities. Second, we demonstrate the importance of addressing selection when estimating technology adoption effects, a point with implications beyond AI. Third, we contribute a novel, publicly available, and continuously updatable measurement of firm-level AI adoption based on earnings call transcripts, covering a broader cross-section of firms and countries than existing alternatives.[1]

Section 2 reviews related literature. Section 3 describes the data. Section 4 develops the empirical strategy. Section 5 presents the main results. Section 6 examines generative AI cohorts. Section 7 investigates heterogeneity. Section 8 reports robustness checks. Section 9 discusses implications and limitations. Section 10 concludes.

# 2 Related Literature

Our paper relates to several strands of literature on automation, AI, and firm performance.

## 2.1 Firm-Level Effects of AI Investment

The most comprehensive firm-level evidence on AI investment comes from Babina et al. (2024), who construct a novel AI investment measure from 535 million employee resumes. Using an instrumental variable strategy based on firms' proximity to university AI talent, they find that AI-investing firms experience higher growth in sales, employment, and market valuations over 2010–2018. Crucially, this growth operates through product innovation

[1] Replication codes and the dataset will be made available on the corresponding author's website: https://ricejonathan.com.

(trademarks and product patents increase) rather than through cost reduction or process efficiency. AI investments show no association with increases in sales per worker or revenue TFP, suggesting that AI's early economic impact was expansionary, not efficiency-enhancing. A companion paper documents that AI-investing firms transition toward more educated workforces with greater STEM specialisation, while flattening organisational hierarchies (Babina et al., 2023). Our results complement this work by extending the sample through the generative AI era. The contrast between their positive employment effects and our negative ones is consistent with a structural shift in AI's labour market impact as capabilities have moved from narrow applications toward general-purpose language and reasoning tools.

For European evidence, Calvino et al. (2022) use direct survey data on AI adoption from France's statistical agency and a difference-in-differences design. AI-adopting French firms see increases in both employment and sales over 2017–2020, with productivity gains appearing to outweigh displacement. Effects vary importantly by AI use type, with some applications expanding employment while others, particularly administrative automation, contract it. A 2026 European Investment Bank working paper finds that AI adoption increases labour productivity by approximately 4% in European firms, driven by capital deepening rather than job losses, though the authors emphasise that complementary investments in software, data, and training are essential for realising these gains (European Investment Bank, 2026). Zolas et al. (2024) link machine-learning-classified AI patents to U.S. Census microdata and find AI-patenting firms show roughly 25% faster employment growth and 40% faster revenue growth versus matched controls. These studies, which largely cover the pre-generative-AI period, form the backdrop against which our post-2022 results should be interpreted.

A recent NBER working paper by Yotzov et al. (2025) uses firm-level survey data from the German Bundesbank and the Australian Bureau of Statistics to document current AI adoption patterns and their correlates. They find that AI-adopting firms are larger, more productive, and more innovative, but that causal effects on employment remain difficult to isolate from selection, a challenge our propensity score matching design directly addresses.

## 2.2 Productivity Effects and the J-Curve

Firm-level productivity gains from AI appear real but smaller than task-level experiments suggest, and they emerge with lags. The most important conceptual framework is the "Productivity J-Curve" developed by Brynjolfsson et al. (2021). They argue that general-purpose technologies like AI require large complementary intangible investments (business process redesign, new data infrastructure, workforce training) that depress measured productivity early while building toward later gains. Adjusting for intangible capital yields TFP levels 15.9% higher than official measures. Behrens et al. (2024) provide the first large-scale micro-level evidence of J-curve patterns, documenting short-run productivity losses from production disruptions during AI adoption, followed by medium-term improvements conditional on organisational readiness.

McElheran et al. (2025) provide the most comprehensive survey-based picture of U.S. AI adoption using the Census Bureau's Annual Business Survey covering approximately 850,000 firms. Fewer than 6% of firms used any AI technology in 2019, though the employment-weighted share was substantially higher at 18%, with adoption concentrated among large, young, technologically sophisticated firms. Our finding that productivity growth shows no clear improvement is consistent with the early phase of the J-curve, though we cannot yet determine whether an upswing will follow. The dynamic evidence (the employment effect becoming more negative over time rather than reversing) is at minimum consistent with a prolonged adjustment period.

## 2.3 Generative AI and Task-Level Productivity

Field experiments on generative AI consistently document substantial task-level productivity improvements, with a distinctive pattern in which lower-skilled workers benefit most, compressing the productivity distribution. Brynjolfsson et al. (2025b) study the staggered rollout of a GPT-based assistant to 5,172 customer support agents at a Fortune 500 firm and find that average productivity rose 15%, with less experienced workers improving by roughly 30%. Noy and Zhang (2023) find ChatGPT reduced task completion time by 40% and raised output quality by 18% in a pre-registered RCT with 453 professionals. Dell'Acqua et al. (2023) conducted an RCT with 758 BCG consultants and found that for tasks inside AI's capability frontier, quality improved 40% and speed increased 25%; for tasks outside the frontier, consultants using AI were 19 percentage points less likely to produce correct solutions. In software development, Peng et al. (2023) find a combined 26% increase in completed tasks across field experiments at Microsoft, Accenture, and a Fortune 100 firm.

Two recent papers temper expectations about aggregate effects. Emanuel et al. (2025) conducted a six-month RCT with 7,137 knowledge workers given Microsoft 365 Copilot and found that adopters spent 2 fewer hours on email per week but detected no shifts in task quantity or composition, suggesting that productivity gains require organisational coordination rather than individual tool access alone. Hougaard et al. (2024) track Danish workers in generative-AI-exposed occupations and find no meaningful employment or earnings effects, reinforcing that task-level gains do not automatically translate into market-level outcomes.

Most directly relevant, Brynjolfsson et al. (2025a) document that unemployment among 20- to 30-year-olds in AI-exposed tech occupations has risen by nearly 3 percentage points since early 2025, providing early labour market evidence consistent with our firm-level displacement findings.

## 2.4 Measuring AI from Corporate Disclosures

The methodological foundation for measuring firm-level AI exposure from earnings calls rests on Hassan et al. (2019), who pioneered using computational linguistics on earnings call transcripts to construct firm-level political risk measures. This approach was extended to climate exposure by Sautner et al. (2023), who used ML-based keyword discovery on earnings calls from 10,000+ firms across 34 countries.

The most directly relevant measurement paper is Eisfeldt et al. (2023), who construct firm-level generative AI workforce exposure by having ChatGPT evaluate task-level substitutability, then merge occupation-level scores with Revelio Labs firm workforce data. They validate the measure against AI discussion in 2023 earnings calls, finding a strong positive relationship. A portfolio long high-exposure firms earned approximately 5% excess returns in the two weeks following ChatGPT's release, directly validating that textual AI discussion in earnings calls correlates with fundamental AI exposure. Bloom et al. (2024) provide a comprehensive multi-source measurement framework tracking novel technology phrases simultaneously across patents, job postings, and earnings calls, demonstrating that each source captures distinct aspects of technology diffusion. Our measurement approach is closest to that of Eisfeldt et al. (2023) in using earnings call text, but differs in that we classify adoption directly from transcript content using an LLM rather than inferring exposure from occupational composition.

## 2.5 The Industrial Robots Benchmark

The automation literature provides a natural benchmark for interpreting AI effects. Acemoglu and Restrepo (2020) established the canonical framework using IFR robot data and a shift-share IV design, finding that one additional robot per thousand workers reduces the U.S. employment-to-population ratio by approximately 0.2 percentage points and wages by approximately 0.42%, with effects concentrated in manufacturing and routine manual occupations. Graetz and Michaels (2018) find robots contributed 0.36 percentage points to annual labour productivity growth across 17 countries without significantly reducing total employment, though low-skilled workers' employment share declined. Webb (2020) provides the key conceptual distinction between AI and robots. Using NLP on patent text matched to O*NET task descriptions, he shows AI targets high-skilled cognitive tasks while robots target manual tasks and software targets routine cognitive tasks. This skill-targeting difference means AI's distributional effects may compress rather than widen wage inequality, and its employment effects may fall on different workers than previous waves of automation. Our firm-level results are consistent with this distinction. The displacement we document is concentrated outside the technology sector and does not appear in manufacturing-heavy segments, suggesting a white-collar rather than blue-collar labour impact.

# 3 Data

Our analysis panel combines earnings call transcripts, from which we classify AI adoption using a large language model, with quarterly financial data from S&P Capital IQ. The resulting panel covers 13,331 firms and approximately 293,000 firm-quarter observations from 2010Q1 through 2025Q4.

## 3.1 Earnings Call Transcript Data

Our primary source of AI adoption information is quarterly earnings call transcripts provided by NL Analytics. Earnings calls are a natural setting for identifying corporate AI investment because they occur at regular quarterly intervals, cover a broad cross-section of publicly listed firms worldwide, and capture managerial statements about strategic investments at the point of executive decision-making. Unlike patent filings, which measure invention rather than deployment, or job postings, which capture hiring intent but not organisational outcomes, earnings calls reflect what firms are actually doing with AI and how management frames these investments to analysts and shareholders.

The raw transcript corpus covers approximately 340,000 earnings calls from over 15,000 unique firms across more than 90 countries. Each transcript includes the full text of the management presentation and analyst question-and-answer session, together with firm identifiers (company name, CIQ identifier, ISIN, ticker) and the call date.

## 3.2 AI Adoption Classification

We classify AI adoption from earnings call transcripts using a three-stage pipeline comprising keyword pre-filtering, human validation, and large language model (LLM) classification.

**Stage 1: Keyword pre-filtering.** We compile a list of 176 AI-related terms spanning general terminology ("artificial intelligence", "machine learning", "deep learning"), specific model families ("GPT", "LLaMA", "Claude"), application categories ("natural language processing", "computer vision", "chatbot"), and technical vocabulary ("neural network", "transformer model", "reinforcement learning"). We apply this keyword list as a boolean filter at the sentence level across all 340,361 earnings calls. A sentence is flagged if it contains at least one keyword match. This yields 279,857 flagged sentences across 53,172 distinct calls from 9,551 of the 13,331 firms in our panel. The full keyword list is reported in Appendix Table 14. Each flagged sentence is stored together with the immediately preceding and following sentences to provide the LLM with local context.

**Stage 2: Human validation and prompt design.** To develop and validate the LLM classifier, two of the authors and a research assistant independently labelled a random sample of 500 flagged sentences as either genuine AI adoption (label = 1) or not (label = 0). Inter-rater agreement was assessed by comparing the labels; disagreements were resolved by joint review, producing a consensus "final manual label" for all 500 sentences. Of the 500 sentences, 262 (52.4%) were labelled as genuine AI adoption.

We then designed three candidate classification prompts and tested each against the manual labels across six OpenAI models (GPT-4o-mini, GPT-4o, GPT-4.1-mini, GPT-4.1, GPT-5-mini, and GPT-5.2). Prompt 1 asks for both a binary label and a confidence score, with explicit guidance on classifying only current AI deployment (not vague future plans). Prompt 2 is a simplified version that omits the confidence score. Prompt 3 adds explicit negative-class examples to guide the model on borderline cases. Full prompt texts are in Appendix C.

Table 1 reports accuracy metrics for each model–prompt combination. Across all three prompts, GPT-4.1-mini achieves the highest F1 score (0.883 on Prompt 1). We also tested a two-pass approach in which sentences classified with low confidence on the first pass are re-evaluated by a stricter follow-up prompt; this marginally improved precision but reduced recall. We selected GPT-4.1-mini with Prompt 1 (single pass) for the production run on the basis of its F1 score and cost-accuracy tradeoff. The classifier labels a sentence as AI-positive only if the firm is *currently* deploying, implementing, or investing in AI; general commentary and announced plans beyond a three-month horizon are classified as negative.

**Stage 3: LLM classification.** We apply the selected model and prompt to all 279,857 flagged sentences. Each sentence is sent to the LLM together with its surrounding context. The classifier returns a binary label (0 or 1) and an integer confidence score (0–100). A separate second-stage prompt classifies the AI use type for positive sentences as *product-facing*, *internal*, *both*, or *unclear*. Of the 252,065 successfully classified segments, 146,974

(58.3%) receive a positive label. Figure 1 shows that among sentences classified as AI-positive, the precision (share that are true positives) rises with the model's own confidence score, reaching over 90% for sentences rated above 80.

**Aggregation to firm-quarter level.** Segment-level predictions are aggregated to the call level, where a call is classified as AI-positive if any segment receives a positive label. Call-level results are then collapsed to the firm-quarter level using a maximum operator, so a firm-quarter is treated as AI-positive if any call that quarter contains a positive classification. We do not impose a confidence threshold at the treatment-definition stage. Instead, we retain the maximum confidence score per firm-quarter and use it in robustness analysis (restricting to high-confidence classifications of 80 or above).

Our treatment definition is absorbing. Once a firm is first classified as an AI adopter, it remains treated in all subsequent quarters for the purposes of the difference-in-differences analysis. This reflects the view that AI adoption represents a persistent shift in firm strategy. In total, the pipeline yields 4,310 ever-treated firms and 9,021 firms that are never classified as AI adopters during our sample period.

Table 1: LLM Classification Accuracy by Model and Prompt

| Model | Prompt | Accuracy (%) | Precision | Recall | F1 |
|---|---|---|---|---|---|
| GPT-4o-mini | Prompt 3 | **87.0** | 0.866 | 0.889 | **0.878** |
| GPT-4.1-mini | Prompt 3 | 86.6 | 0.868 | 0.878 | 0.873 |
| GPT-4.1-mini | Prompt 1 | 86.0 | 0.831 | 0.920 | 0.873 |
| GPT-5.2 | Prompt 1 | 86.4 | 0.837 | 0.920 | 0.876 |
| GPT-4o | Prompt 3 | 86.0 | 0.900 | 0.824 | 0.861 |
| GPT-4.1 | Prompt 1 | 84.0 | 0.825 | 0.882 | 0.852 |
| GPT-5-mini | Prompt 3 | 81.4 | 0.744 | 0.985 | 0.847 |

*Notes:* Each model–prompt combination is evaluated on the same 500 manually labelled sentences. Accuracy is the share of correct predictions. Precision is the share of predicted positives that are true positives. Recall is the share of true positives correctly identified. Only the best-performing prompt per model is shown; the full 6×3 grid is in Appendix Table 12. GPT-4.1-mini with Prompt 1 (row 3) was selected for the production run on the basis of its F1 score, the confidence score it provides, and its cost-accuracy tradeoff. All models use temperature = 0.

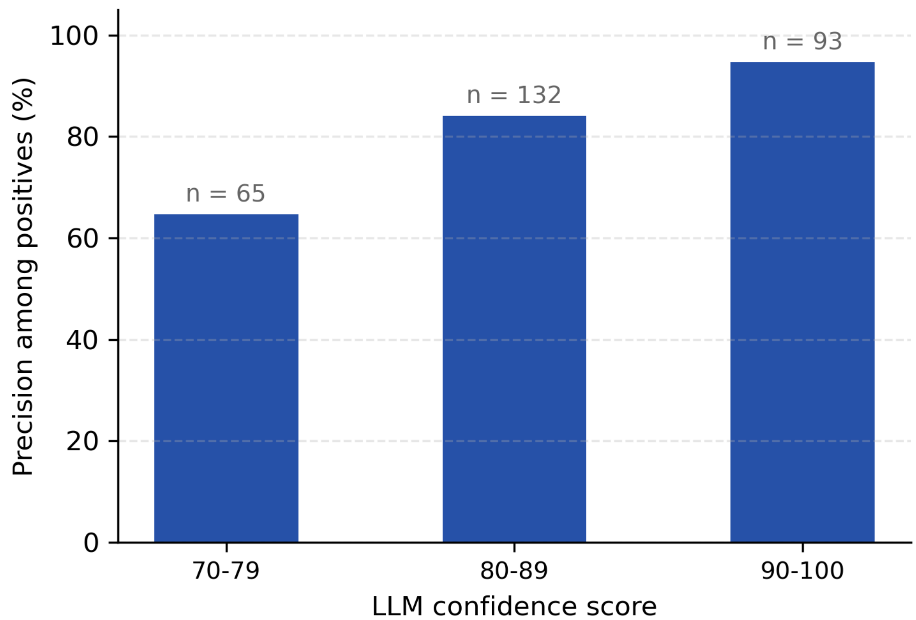



Figure 1: Precision of Positive Classifications by Confidence Level

*Notes:* This figure plots the precision (share of true positives among predicted positives) of GPT-4.1-mini on the 500 manually labelled sentences, restricted to sentences the model classified as AI-positive (label = 1) and grouped by the model's confidence score. Higher confidence scores are associated with higher precision. Numbers above each bar indicate the count of positive predictions in that confidence bucket.

### 3.3 Financial Data

Quarterly financial data come from S&P Capital IQ. For each firm-quarter, we observe total revenue, total employees, total assets, EBITDA, operating income, cost of goods sold, selling and administrative expenses, research and development expenditure, total debt, cash and short-term investments, capital expenditure, and equity. We also observe market-based variables including market capitalisation and closing stock price, though we note below that these contain significant data quality limitations.

From the raw financial variables we construct several derived measures. Our three primary outcomes are year-over-year log growth rates, namely employment growth ($\Delta_4 \ln \text{Emp}_{it}$), revenue growth ($\Delta_4 \ln \text{Rev}_{it}$), and revenue-per-employee growth ($\Delta_4 \ln(\text{Rev}/\text{Emp})_{it}$), where $\Delta_4$ denotes the four-quarter log difference. These three outcomes are linked by an accounting identity in which revenue-per-employee growth equals revenue growth minus employment growth. We exploit this identity in our analysis by estimating all three outcomes on a common sample, ensuring that our coefficients are internally consistent. We also construct standard financial ratios (EBITDA margin, operating margin, gross margin, debt-to-assets, and current ratio) which serve both as matching covariates and as secondary outcomes.

All financial variables are denominated in US dollars. We filter the sample to firms with at least 10 reported employees, as smaller headcounts in Capital IQ frequently reflect data gaps rather than genuine firm size.

## 3.4 Panel Construction

The analysis panel is formed by merging the AI treatment panel with the financial panel at the firm-quarter level, using company identifiers provided by Capital IQ. The match rate exceeds 99% of financial-panel firms. The merged panel contains 292,598 firm-quarter observations for 13,331 unique firms over 64 calendar quarters (2010Q1–2025Q4). Of these, 4,310 firms are ever classified as AI adopters and 9,021 are never treated. Table 2 reports summary statistics for the full panel, split by treatment status.

Three features of the panel are worth noting. First, the panel is unbalanced, as not all firms report in every quarter and some firms enter or exit the sample over time. Second, the treatment rate rises sharply over the sample period, from near zero before 2015 to over 30% of active firms by late 2025 (see Section 3.5 below). Third, treatment is not randomly assigned. AI-adopting firms are systematically larger, more profitable, and faster-growing than non-adopters even before they begin discussing AI. This selection pattern is central to our empirical strategy and is documented in detail in Section 4.2.

Table 2: Panel Summary Statistics

| | All Firms | AI Adopters | Non-Adopters |
|---|---|---|---|
| Firms | 13,331 | 4,310 | 9,021 |
| Firm-quarters | 292,598 | 144,684 | 147,914 |
| Countries | 93 | 68 | 89 |
| Sectors | 13 | 12 | 13 |
| Median log revenue | 12.45 | 13.00 | 11.98 |
| Median log employment | 6.88 | 7.55 | 6.16 |
| Median EBITDA margin | 0.147 | 0.150 | 0.144 |

*Notes:* AI adopters are firms classified as making substantive AI investments in at least one quarter. Log revenue and log employment are natural logarithms of USD revenue and headcount.

### 3.5 Descriptive Facts

Five patterns in the data are worth highlighting. First, AI adoption in earnings calls rises sharply over the sample period. Figure 2 plots the share of firms disclosing AI adoption each quarter. AI mentions are rare before 2015, grow steadily from 2017, and accelerate dramatically after the release of ChatGPT in late 2022. By the end of 2025, over 30% of sample firms discuss substantive AI activity on their earnings calls. Figure 3 shows the corresponding cumulative count of ever-treated firms, which follows an S-shaped diffusion curve with a clear inflection point around 2023.

Second, the technology sector leads adoption, but AI diffusion is broad-based by 2023. Figure 4 plots sector-level adoption rates over time. While the technology sector exceeds a 20% adoption rate as early as 2019, healthcare, industrials, and consumer sectors all show rapid convergence during the generative AI era. By 2025, no major sector has an adoption rate below 15%.

Third, the sample is genuinely global. Figure 5 displays firm counts by country of headquarters for the 20 most-represented countries. The United States accounts for the largest share, but the sample includes substantial representation from Japan, the United Kingdom, Canada, India, China, Germany, Australia, and other economies. This geographic breadth distinguishes our panel from datasets limited to a single country.

Fourth, the composition of AI use types evolves over time. Figure 6 decomposes AI-positive firm-quarters by use type. In early years, product-facing AI and unclear classifications dominate. During the generative AI wave, internal operations and combined product-internal deployments grow fastest, consistent with firms increasingly deploying AI for process automation and efficiency alongside customer-facing applications.

Fifth, AI-adopting firms differ systematically from non-adopters before treatment. As documented in detail in Section 4.2, treated firms are substantially larger, more profitable, and faster-growing in the pre-treatment period, motivating our propensity score matching design.

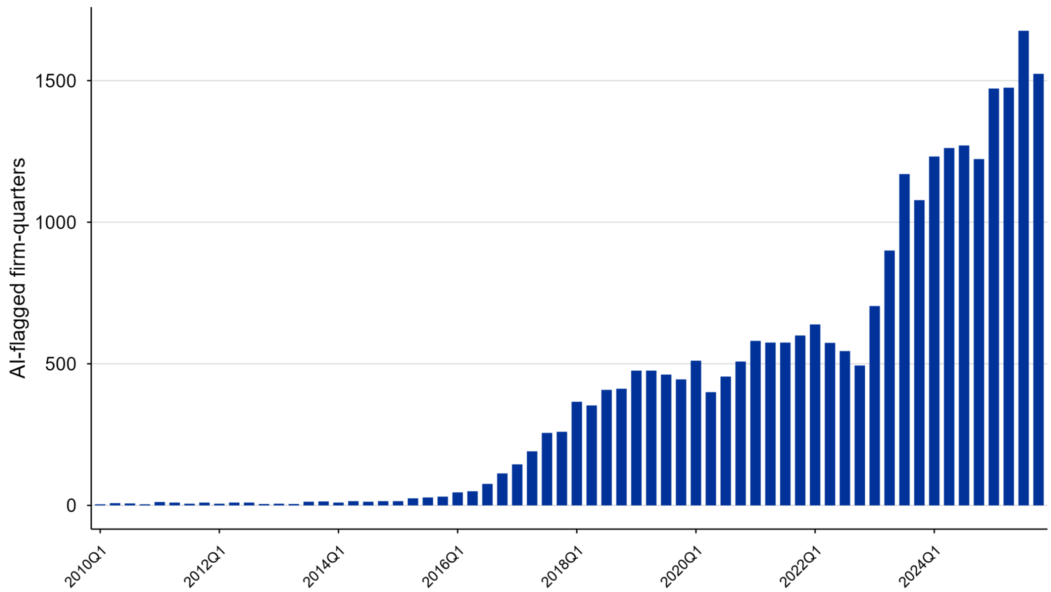



Figure 2: AI Adoption in Earnings Calls Over Time

*Notes:* This figure plots the percentage of firms disclosing substantive AI adoption in quarterly earnings calls. The dashed vertical line marks the release of ChatGPT in November 2022.

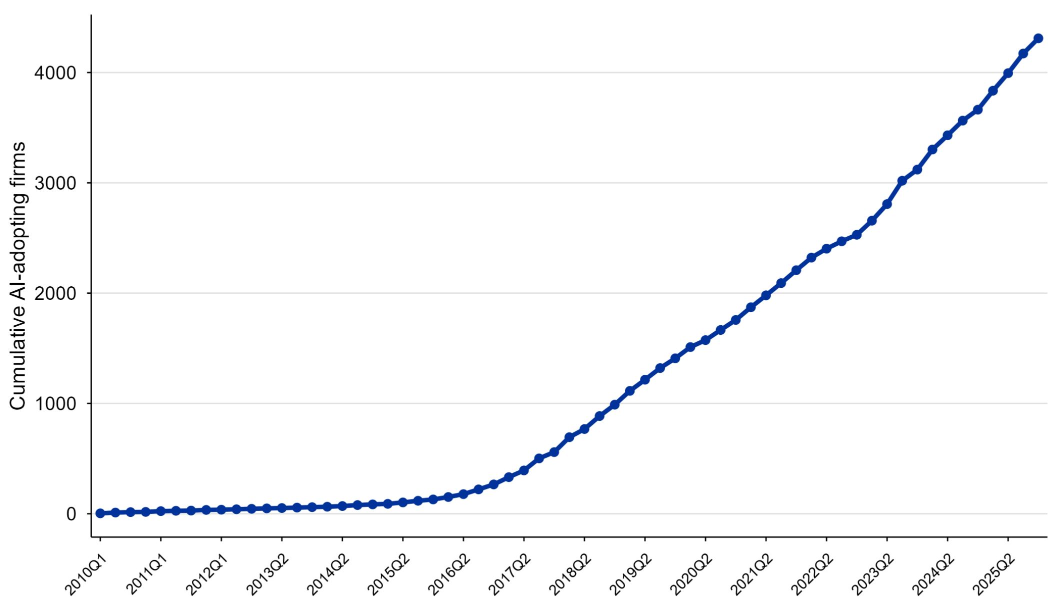



Figure 3: Cumulative AI Adopters

*Notes:* This figure plots the cumulative number of unique firms classified as AI adopters (first substantive AI disclosure in an earnings call).

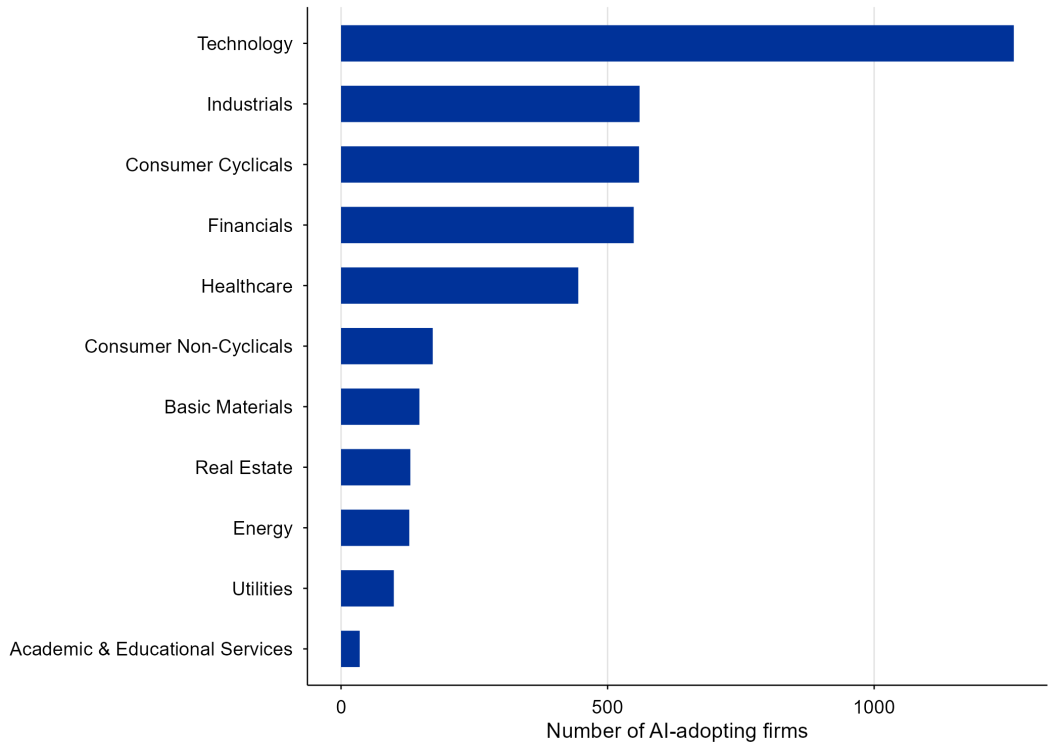



Figure 4: AI Adoption Rate by Economic Sector

*Notes:* This figure plots the share of firms disclosing AI adoption by economic sector, for the eight sectors with the highest adoption rates in the most recent quarter.

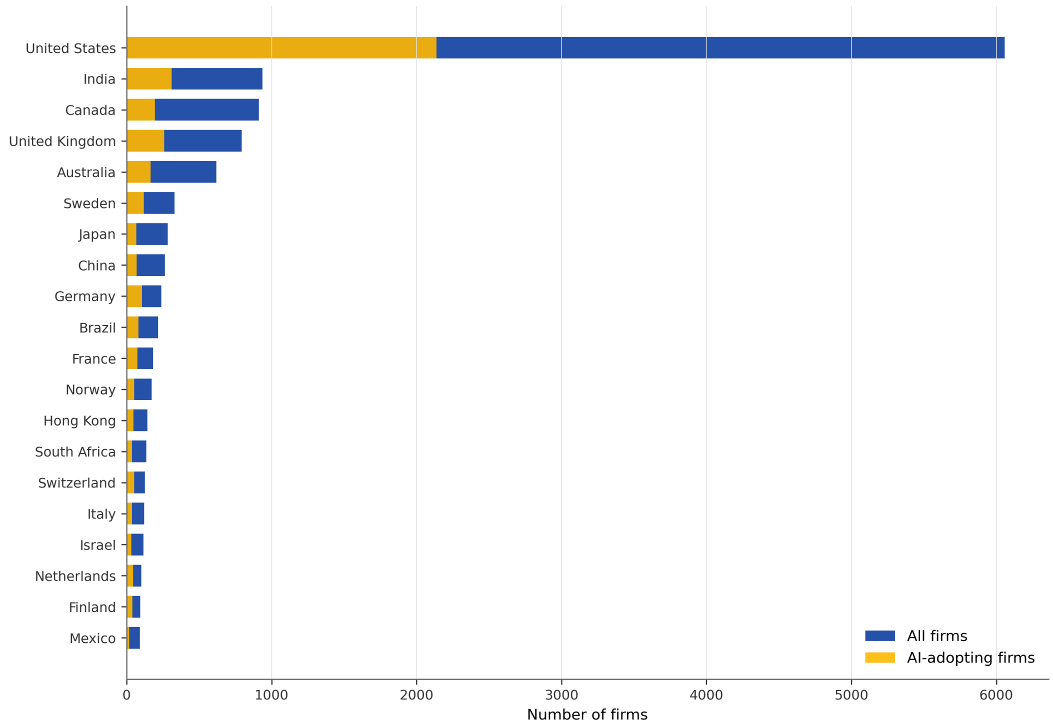



Figure 5: Geographic Coverage of Sample

*Notes:* This figure shows the number of firms by country of headquarters for the 20 most-represented countries. Blue bars show all firms; gold bars show ever-treated (AI-adopting) firms.

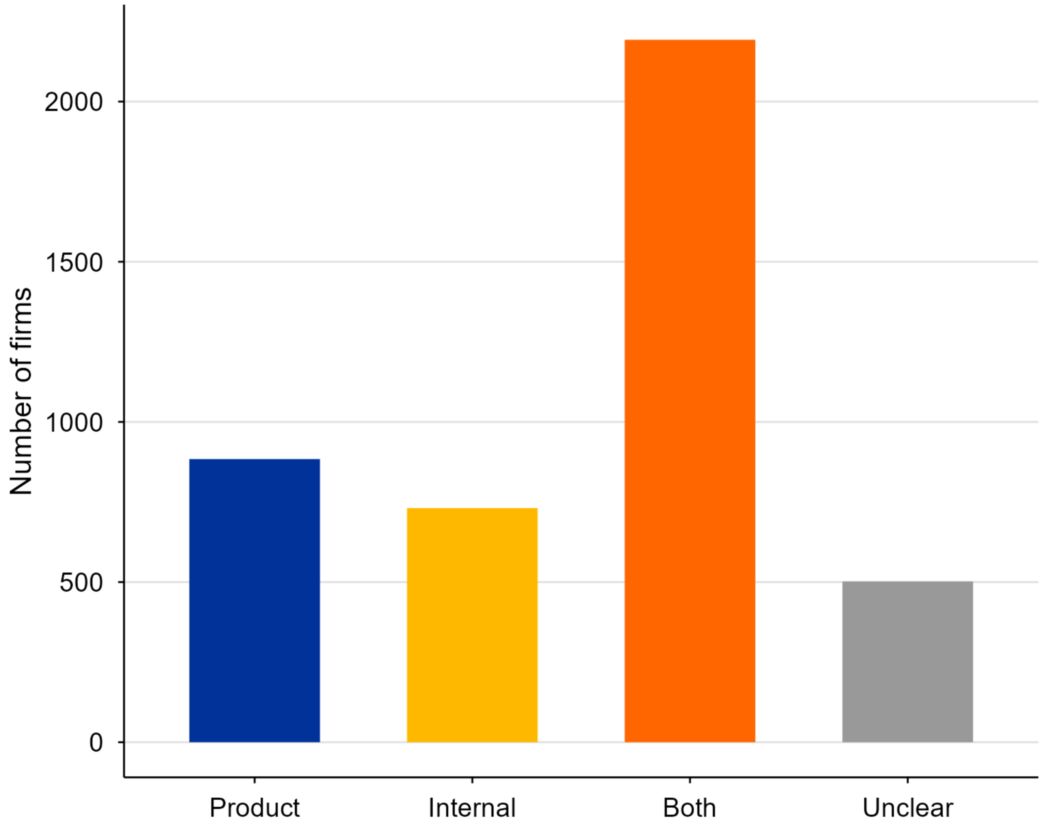



Figure 6: Composition of AI Use Types Over Time

*Notes:* This figure shows the composition of AI adoption disclosures by use type: product-facing (customer applications), internal operations (process automation), both, and unclear.

# 4 Empirical Strategy

Our identification strategy proceeds in four stages. We begin with a standard full-sample difference-in-differences specification as a benchmark. We then document the selection problem (the systematic differences between AI adopters and non-adopters that contaminate the benchmark estimates) and motivate cohort-specific propensity score matching as a remedy. Third, we describe the dynamic event-study specification used to trace out the time path of treatment effects. Fourth, we introduce the cohort-robust estimator of Callaway and Sant'Anna (2021) to address concerns about heterogeneous treatment effects under staggered adoption.

## 4.1 Full-Sample Difference-in-Differences

Our baseline specification is a two-way fixed effects (TWFE) difference-in-differences model:

$$Y_{it} = \alpha_i + \gamma_t + \beta \cdot D_{it} + \varepsilon_{it} \tag{1}$$

where $Y_{it}$ is the outcome for firm $i$ in quarter $t$, $\alpha_i$ is a firm fixed effect, $\gamma_t$ is a quarter fixed effect, and $D_{it}$ is an indicator equal to one if firm $i$ has been classified as an AI adopter by quarter $t$ and zero otherwise. Because treatment is absorbing, $D_{it}$ switches from zero to one at the quarter of first adoption and remains one thereafter. Standard errors are clustered at the firm level throughout.

We estimate Equation (1) for three primary outcomes, namely year-over-year log employment growth ($\Delta_4 \ln \mathrm{Emp}_{it}$), year-over-year log revenue growth ($\Delta_4 \ln \mathrm{Rev}_{it}$), and year-over-year log revenue-per-employee growth ($\Delta_4 \ln(\mathrm{Rev}/\mathrm{Emp})_{it}$). As noted in Section 3.3, these three outcomes are linked by an accounting identity, $\Delta_4 \ln \mathrm{Rev}_{it} = \Delta_4 \ln \mathrm{Emp}_{it} + \Delta_4 \ln(\mathrm{Rev}/\mathrm{Emp})_{it}$. We estimate all three on a common sample (the set of firm-quarters where all three outcomes are simultaneously observed) so that the estimated coefficients satisfy $\hat{\beta}_{\mathrm{rev}} = \hat{\beta}_{\mathrm{emp}} + \hat{\beta}_{\mathrm{rpe}}$ exactly. This ensures internal consistency and allows the reader to trace how any employment effect decomposes into revenue and productivity channels. We do not winsorise the outcome variables; the unwinsorised estimates are our preferred specification, and we show in Section 8 that results are not sensitive to outlier treatment.

The identifying assumption of the TWFE specification is that of parallel trends. In the absence of AI adoption, treated and control firms would have followed the same trajectory in outcomes. This assumption is plausible if adoption timing is uncorrelated with time-varying shocks that also affect employment and revenue growth. However, as we document next, this assumption is threatened by the systematic observable differences between adopters and non-adopters.

## 4.2 The Selection Problem

Firms that adopt AI are not drawn randomly from the population. Figure 7 compares the distributions of key pre-treatment characteristics, averaged over the pre-adoption period, for firms that eventually adopt AI versus those that do not. AI adopters are substantially larger (higher log revenue and log employment), more profitable (higher EBITDA margins), and faster-growing (higher pre-treatment employment growth). Standardised mean differences between adopters and non-adopters exceed 0.5 standard deviations on most size measures and reach 1.0 on the propensity score itself.

This positive selection has a clear implication for the TWFE estimator. If adopters were already on superior growth trajectories before treatment, the TWFE estimate of $\beta$ in Equation (1) conflates any causal treatment effect with the pre-existing growth advantage of adopting firms. In the case of employment, this means the TWFE estimator will overstate any positive effect, or equivalently attenuate any negative effect. The direction of the bias is unambiguous, since adopters are positively selected on growth and the naïve estimate is therefore biased upward relative to the true treatment effect.

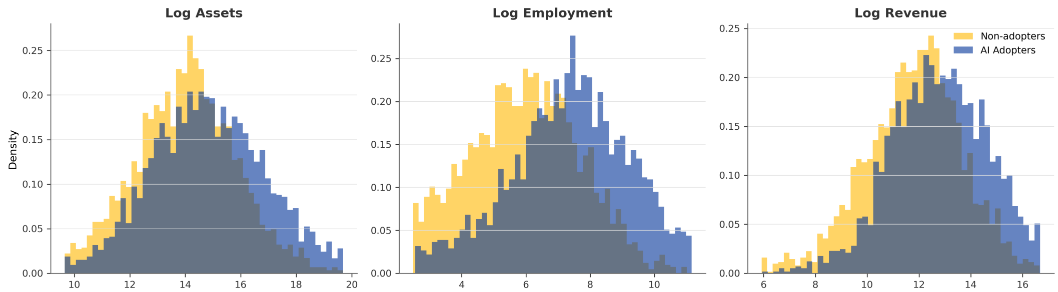



Figure 7: Pre-Treatment Characteristics: Treated vs. Control

*Notes:* This figure compares the distributions of key pre-treatment characteristics for AI-adopting firms versus non-adopters. The systematic rightward shift for adopters on size and profitability measures motivates the cohort-specific propensity score matching design.

## 4.3 Cohort-Specific Propensity Score Matching

To address selection on observables, we use a cohort-specific propensity score matching (PSM) design that accounts for the staggered timing of AI adoption. A fixed pre-treatment averaging window such as 2017–2019 becomes increasingly stale for later adopters. A firm first disclosing AI in 2024 or 2025 would be matched on covariates measured five or more years before treatment, during which its characteristics may have changed substantially. Our rolling-window design keeps matching covariates temporally relevant for every cohort.

The procedure has five steps.

First, we group treated firms into half-year cohort bins by the quarter in which they are first classified as AI adopters. For each cohort bin, we compute firm-level averages of matching covariates over the eight quarters immediately preceding the bin's median treatment date. The covariates include log revenue, log assets, log employment, year-over-year revenue growth, year-over-year employment growth, EBITDA margin, operating margin, gross margin, SG&A share of revenue, debt-to-assets, current ratio, and economic sector (with a "missing sector" category for firms without sector data). For control firms, we compute the same covariate means over the same calendar quarters, ensuring that treated and control firms are compared on characteristics measured at the same point in time. We exclude firms first treated before 2020Q1, ensuring that at least eight pre-treatment quarters are available for the rolling window for all retained cohorts.

Second, within each cohort bin, we estimate a logistic regression of treatment status on the matching covariates and sector indicators, yielding a propensity score for each firm.

Third, we match each treated firm to the nearest control firm on the linearised propensity score, using one-to-one nearest-neighbour matching without replacement, with a caliper of 0.25 standard deviations of the logit propensity score. The estimand is the average treatment effect on the treated (ATT). Each matched control inherits the treatment date of its matched treated firm, so that both face the same macroeconomic conditions at every event-time horizon.

Fourth, controls are matched without replacement *across* cohort bins, so a control firm used in one cohort cannot be reused in another. This prevents the same control from appearing multiple times in the matched panel, which would introduce mechanical correlation across cohort-specific estimates.

Fifth, we verify balance. Figure 8 displays the absolute standardised mean difference (SMD) for each covariate before and after matching, computed as a weighted average across cohort bins. Before matching, SMDs on the key size variables (log employment, log revenue, log assets) range from 0.36 to 0.78. After matching, the three size SMDs fall below 0.09, and the maximum SMD across all covariates is 0.14 (on SG&A share). The first-stage logistic regressions achieve median pseudo-$R^2$ of 0.26 across cohort bins (range 0.17–0.36), confirming that the propensity score has meaningful discriminatory power. The caliper of 0.25 standard deviations excludes the most poorly matched treated firms: 77% of eligible treated firms find a match within the caliper, yielding 1,333 matched pairs. Figure 9 confirms substantial overlap in the propensity score distributions of treated and control firms.

This procedure yields 1,333 matched pairs. Having constructed the matched sample, we re-estimate Equation (1) on the matched panel. The matched estimate of $\beta$ is our preferred measure of the treatment effect. We restrict to the common sample where all three outcomes (employment growth, revenue growth, revenue-per-employee growth) are simultaneously observed, yielding an effective sample of 2,467 firms (1,302 treated) and 77,018 firm-quarter observations. As robustness checks, we also estimate specifications restricting to non-technology firms, to high-confidence classifications (confidence $\geq 80$), and to sustained treatment requiring consecutive AI mentions.

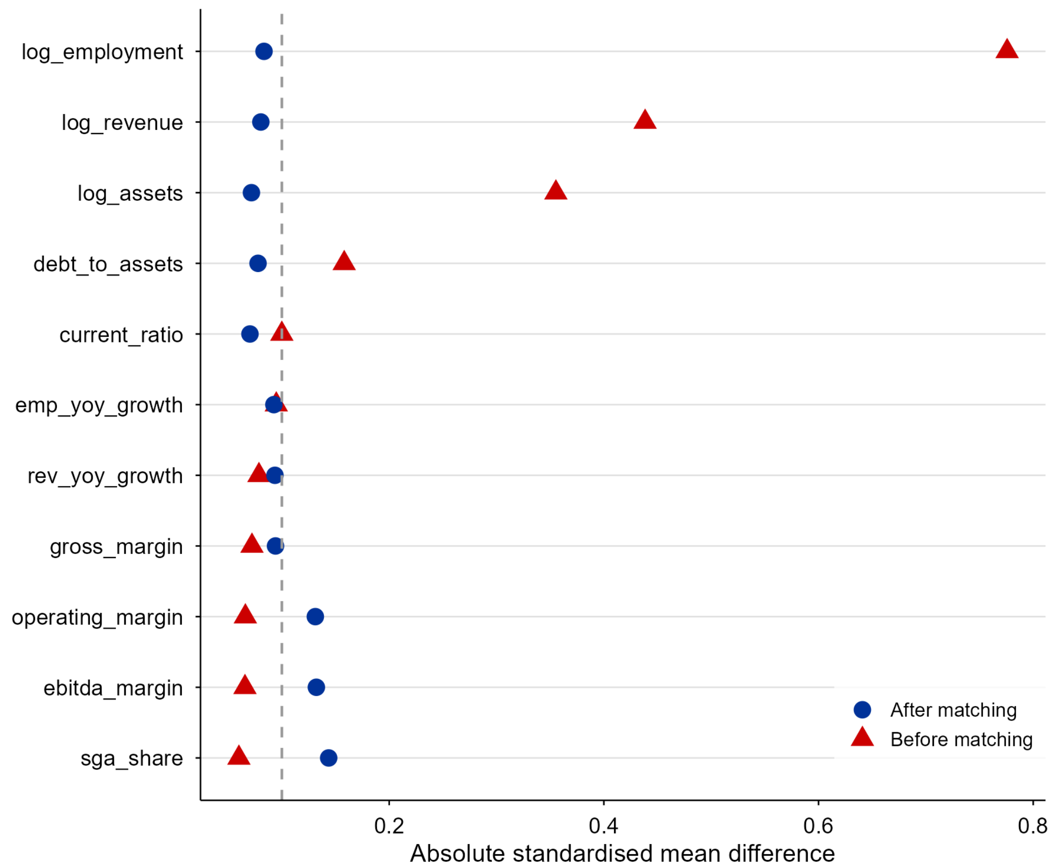



Figure 8: Covariate Balance Before and After Matching

*Notes:* This figure displays the absolute standardised mean difference for each matching covariate before (red triangles) and after (blue circles) cohort-specific propensity score matching. Balance statistics are computed as weighted averages across the 12 half-year cohort bins, weighted by the number of treated firms in each bin. The dashed vertical line marks the 0.10 threshold.

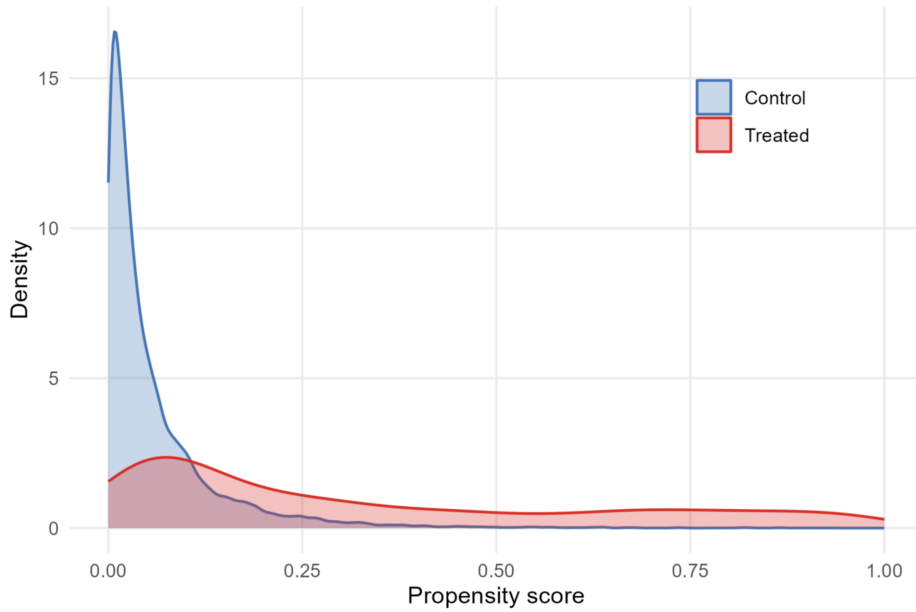



Figure 9: Propensity Score Distributions: Treated vs. Control

*Notes:* Kernel density estimates of the estimated propensity score for treated (red) and control (blue) firms, pooled across all cohort bins. The substantial overlap in the common support region confirms that the matching procedure draws controls from a comparable region of the covariate space.

### 4.4 Dynamic Specification

The static estimate from Equation (1) captures the average post-treatment effect across all horizons. To examine the time path of treatment effects and to test the parallel trends assumption, we estimate a dynamic event-study specification using the interaction-weighted estimator of Sun and Abraham (2021):

$$Y_{it} = \alpha_i + \gamma_t + \sum_{k \neq -1} \mu_k \cdot \mathbf{1}[t - g_i = k] + \varepsilon_{it} \tag{2}$$

where $g_i$ is the quarter in which firm $i$ is first treated and $k$ indexes event time relative to adoption. The reference period is $k = -1$ (the quarter immediately before adoption). The Sun–Abraham estimator properly accounts for treatment effect heterogeneity across adoption cohorts, avoiding the negative weighting problems that can arise in standard TWFE event studies with staggered treatment (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020).

We estimate Equation (2) over event times $k \in [-8, +12]$, corresponding to two years before through three years after adoption. The pre-treatment coefficients ($k \in [-8, -2]$)

serve as a test of the parallel trends assumption. If treated firms were already diverging before adoption, these coefficients would be non-zero. The post-treatment coefficients trace out the dynamic response.

We report event-study results for both the full unmatched sample and the propensity-score-matched sample. The full-sample specification has more statistical power due to the larger number of firms, while the matched specification addresses observable selection. As we show in Section 5.3, the two specifications produce qualitatively similar patterns, with the full-sample event study providing the sharper test of the timing of effects.

### 4.5 Cohort-Robust Estimation

Because AI adoption is staggered, with adoption cohorts spanning 2020 through 2025, the standard TWFE estimator may produce biased estimates if treatment effects are heterogeneous across cohorts (de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2021; Callaway and Sant'Anna, 2021). As a complement to the Sun–Abraham interaction-weighted event study used above, we estimate the group-time ATT of Callaway and Sant'Anna (2021). This estimator computes separate treatment effects for each adoption cohort at each calendar period, using only never-treated firms as the control group. The cohort-specific effects are then aggregated into an overall ATT and a dynamic event-study representation. We run the Callaway–Sant'Anna estimator on the cohort-matched panel, ensuring that it inherits the selection correction from our propensity score matching.

## 5 Results

This section presents results in three stages. We first report the full-sample benchmark, which finds no significant effects. We then show that propensity score matching reveals a negative employment effect that was masked by selection. Finally, we present dynamic event-study evidence on the timing of the effect.

### 5.1 Full-Sample Benchmark

Table 3 (left panel) reports the full-sample TWFE estimates from Equation (1). On the full sample of all firms, the coefficient on employment growth is $-0.002$ ($p = 0.34$), small, negative, and statistically indistinguishable from zero. Revenue growth and productivity growth are similarly null.

These null results are misleading. As documented in Section 4.2, AI adopters are positively selected on pre-treatment characteristics, so the full-sample estimate conflates a negative treatment effect with the superior baseline trajectories of adopting firms. The result is an attenuated coefficient.

## 5.2 Matched Results

Table 3 (right panel) reports the matched DiD estimates. The results differ sharply from the full-sample benchmark. On the cohort-matched common sample (2,467 firms, 77,018 observations), employment growth declines by 1.1 percentage points per year ($p = 0.003$). Revenue growth is essentially zero ($-0.2$ pp, $p = 0.81$), and revenue-per-employee growth shows a positive but insignificant point estimate ($+0.9$ pp, $p = 0.24$).

Figure 10 compares the full-sample and matched estimates. The matched employment coefficient ($-0.011$) is roughly five times the full-sample estimate ($-0.002$). Firms that adopt AI were growing employment faster to begin with; once this advantage is removed through matching, the underlying displacement becomes visible.

Revenue per employee growth is positive but insignificant on the matched sample ($+0.009$, $p = 0.24$), providing no clear evidence of productivity improvement accompanying the employment decline, though as we show in Section 8, the RPE effect becomes marginally significant under a sustained-treatment definition. Revenue growth is close to zero and insignificant. Appendix Table 10 also shows a modest increase in debt-to-assets among matched non-technology firms, consistent with some firms financing AI-related adjustment or restructuring with slightly higher leverage. Overall, the pattern is one of employment contraction without compensating revenue-per-employee improvement in the baseline specification.

The effect is stable across alternative specifications (Table 8). Restricting to non-technology firms yields $-1.0$ pp ($p = 0.010$). Restricting treatment to high-confidence classifications (confidence score $\geq 80$) produces a slightly larger estimate of $-1.3$ pp ($p = 0.0005$), consistent with measurement error in the baseline treatment definition attenuating the effect toward zero.

Table 3: Main Results: Full-Sample vs. Matched Difference-in-Differences

| | Full-Sample DiD | | | PSM Matched DiD | | |
|---|---|---|---|---|---|---|
| | Emp | Rev | RPE | Emp | Rev | RPE |
| Treated × Post | −0.002 | −0.004 | −0.001 | −0.0107*** | −0.0018 | +0.0089 |
| | (0.002) | (0.004) | (0.004) | (0.0036) | (0.0074) | (0.0075) |
| $N$ | 200,694 | 200,694 | 200,694 | 77,018 | 77,018 | 77,018 |
| Firms | 8,231 | 8,231 | 8,231 | 2,467 | 2,467 | 2,467 |

*Notes:* This table reports difference-in-differences estimates of the effect of AI adoption on year-over-year employment growth, revenue growth, and log revenue per employee. Left panel: full unmatched sample (TWFE, absorbing treatment). Right panel: cohort-specific propensity-score-matched sample (all firms). All specifications include firm and quarter fixed effects. Standard errors clustered by firm in parentheses. All three outcomes are estimated on a common sample (firm-quarters where all three variables are observed), so observation counts are identical across outcomes within each panel. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

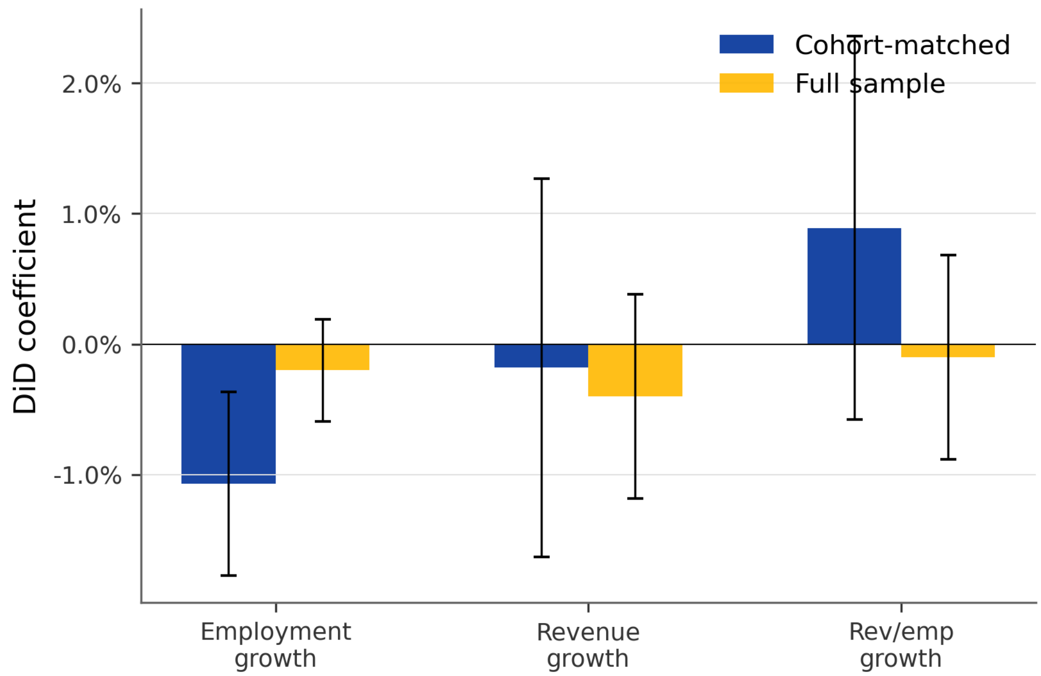



Figure 10: Employment Growth Effect: Full-Sample vs. Matched

*Notes:* This figure plots the point estimates and 95% confidence intervals for the three outcome variables from both the full-sample and cohort-matched DiD specifications on all firms. Matching amplifies the employment effect approximately fivefold relative to the full-sample benchmark.

### 5.3 Dynamic Evidence

The static estimates in Table 3 pool all post-treatment periods into a single coefficient. Figure 11 presents the dynamic event-study estimates from Equation (2), allowing the employment effect to vary by quarters since adoption. We report two complementary specifications, the full unmatched sample (which maximises statistical power) and the propensity-score-matched sample (which addresses selection).

We highlight three features. First, the pre-treatment coefficients ($k = -8$ through $k = -2$) are small and individually insignificant in the full-sample specification. A joint Wald test fails to reject the null that all seven pre-treatment coefficients are simultaneously zero (full-sample all-firms $p = 0.92$), providing no evidence that treated firms were on diverging employment trajectories before adoption. On the matched all-firms sample (Table 4), one pre-treatment coefficient is individually significant and the joint test rejects at 5% ($p = 0.022$), though this reflects isolated noise in one pre-treatment period rather than a systematic trend.

Second, the employment effect grows more negative after adoption. In the first year post-adoption (quarters 0–4), the average effect is small. By the second and third years (quarters 5–12), the effect becomes larger and reaches $-1.6$ to $-1.8$ percentage points at the longest horizons. One interpretive caveat applies to the early post-treatment coefficients. Because the outcome is a four-quarter log difference, each observation at event times $k = 0$ through $k = 3$ mechanically combines post-treatment and pre-treatment employment levels. The gradual build-up in this window therefore partly reflects the overlapping construction of the outcome variable. The continued deepening of the effect beyond $k = 4$, where the outcome is fully post-treatment at both ends, is not subject to this concern and provides cleaner evidence of an effect that grows over time.

Third, the long-horizon effects are statistically significant. On both the full all-firms sample and the cohort-matched all-firms sample, all four individual coefficients at event times 9 through 12 are significant at the 5% level. Table 4 summarises the average coefficient and the number of individually significant estimates within each event-time window.

The timing is consistent with a causal channel but does not definitively establish one. Employment growth diverges from matched controls only *after* AI adoption is first disclosed, not before. This is hard to reconcile with simple reverse causality. We cannot, however, rule out unobserved strategic decisions that coincide with AI adoption and independently affect employment.

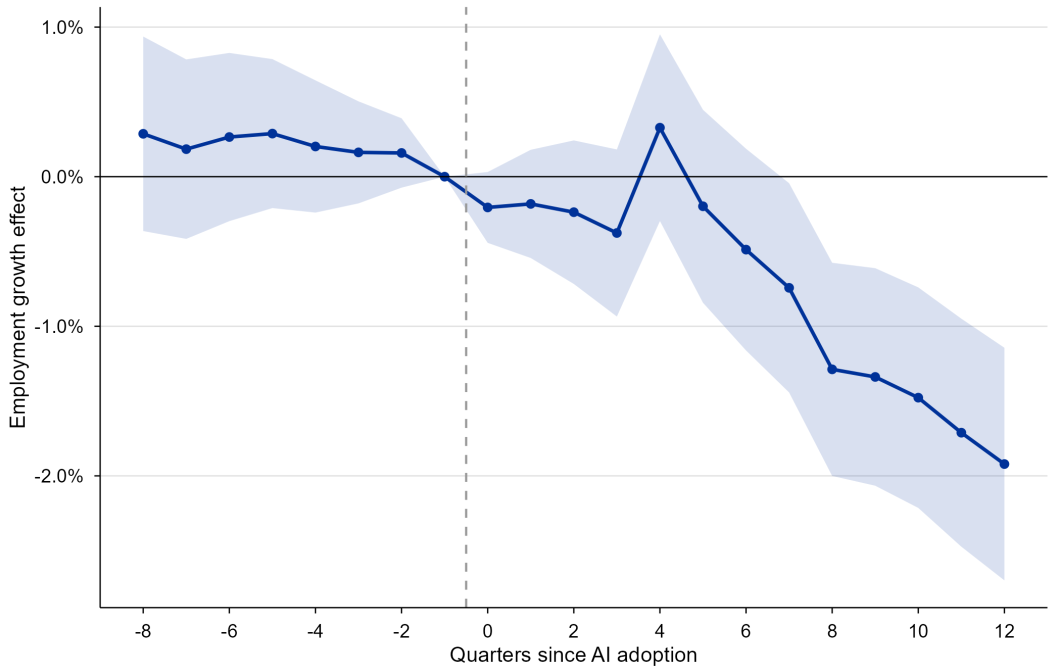



Figure 11: Dynamic Employment Effects After AI Adoption

*Notes:* Sun–Abraham interaction-weighted event-study coefficients for year-over-year employment growth, estimated on the full unmatched sample. The omitted reference period is $k = -1$. Shaded area represents 95% confidence intervals based on firm-clustered standard errors. The dashed vertical line marks the adoption quarter.

Table 4: Dynamic Employment Windows

| | Full-Sample (all) | | PSM Matched (all) | |
|---|---|---|---|---|
| Window | Avg. effect | Sig. at 5% | Avg. effect | Sig. at 5% |
| Pre (−8 to −2) | +0.001 | 0/7 | +0.006 | 1/7 |
| Joint Wald $p$ | 0.918 | | 0.022 | |
| 0–4 quarters | −0.002 | 0/5 | −0.007 | 4/5 |
| 5–8 quarters | −0.007 | 1/4 | −0.012 | 1/4 |
| 9–12 quarters | −0.016 | 4/4 | −0.021 | 4/4 |

*Notes:* This table reports average Sun–Abraham event-study coefficients for employment growth within each event-time window, along with the number of individually significant coefficients (at 5%) within each window. The joint Wald $p$-value tests the null that all seven pre-treatment coefficients ($k = -8$ through $k = -2$) are simultaneously zero. Pre-treatment window: $t \in [-8, -2]$. Specifications include firm and quarter fixed effects with firm-clustered standard errors.

## 5.4 Cohort-Robust Estimates

Table 5 compares the baseline TWFE-on-matched estimates with the Callaway–Sant'Anna cohort-robust ATT. The cohort-robust ATT for employment growth on the full matched sample is −1.3 percentage points ($p = 0.064$), comparable in magnitude to the TWFE matched estimate of −1.1 pp ($p = 0.003$) though less precisely estimated due to the smaller effective sample that results from the balanced-panel requirement within the `did` estimator. Revenue per employee and revenue growth are both statistically insignificant under either estimator.

The fact that the cohort-robust estimate is at least as large as the TWFE estimate is reassuring. It indicates that the "negative weighting" problem documented in the staggered DiD literature is not inflating our headline result. If anything, TWFE slightly attenuates the employment effect by averaging across heterogeneous cohort-specific effects.

Table 5: Estimator Comparison: TWFE vs. Callaway–Sant'Anna (Matched Panel)

| | | TWFE Matched | | Callaway–Sant'Anna | |
|---|---|---|---|---|---|
| Outcome | Sample | Coeff. | $p$ | ATT | $p$ |
| Emp growth | All | $-0.011^{***}$ | 0.003 | $-0.013^{*}$ | 0.064 |
| Rev/Emp growth | All | +0.009 | 0.239 | +0.018 | 0.549 |
| Rev growth | All | −0.002 | 0.810 | +0.005 | 0.865 |

*Notes:* Left panel: TWFE with absorbing treatment, firm and quarter fixed effects, on the cohort-matched common sample (all firms). Right panel: Callaway and Sant'Anna (2021) group-time ATT aggregated to a simple overall ATT, run on the same cohort-matched panel with never-treated firms as the control group. Standard errors clustered at the firm level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

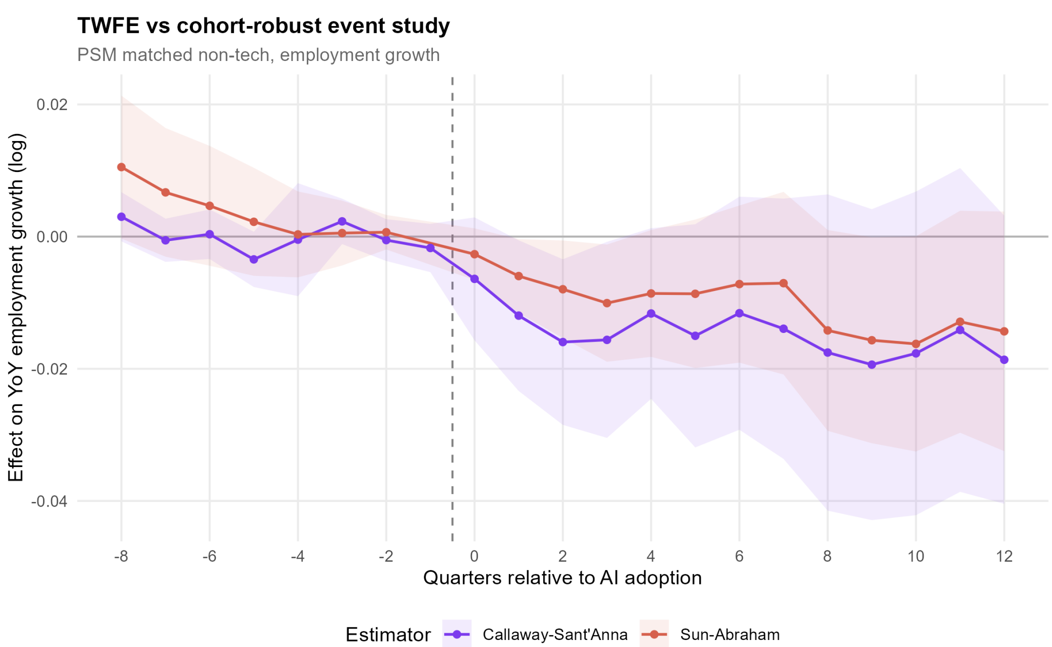



Figure 12: Event Study: Sun–Abraham vs. Callaway–Sant'Anna (Matched Sample)
*Notes:* Dynamic treatment effects on employment growth from the Sun–Abraham interaction-weighted estimator (red) and the Callaway and Sant'Anna (2021) group-time ATT aggregated to event time (purple), both estimated on the cohort-matched sample (all firms). Shaded bands show 95% confidence intervals.

# 6 The Generative AI Acceleration

The results so far pool all adoption cohorts. Here we ask whether employment displacement differs between firms that adopted AI before and after the generative AI wave. Pre-generative-AI adopters are firms whose first AI classification occurs before 2023Q1; post-generative-AI adopters are those first classified in 2023Q1 or later, following the public release of ChatGPT in November 2022. Each cohort is compared against the same never-treated controls within the matched sample.

Table 6 presents the results. Among all matched firms, the pre-generative-AI cohort shows an employment growth effect of $-0.5$ percentage points ($p = 0.331$), negative but statistically insignificant. The post-generative-AI cohort shows an effect of $-1.7$ percentage points ($p = 0.001$), more than three times the magnitude and clearly significant.

The larger displacement for the later cohort is consistent with a qualitative shift in what AI can do. Pre-2023 corporate AI was dominated by computer vision, recommendation systems, and predictive analytics, all narrow tools requiring substantial custom engineering. Generative AI operates on language, code, and content, domains that overlap broadly with white-collar knowledge work, and deploys with lower barriers. Both facts could explain why employment effects are larger for this cohort.

Two caveats apply. First, this is a cohort comparison, not a sharp discontinuity. The 2023Q1 cutoff is somewhat arbitrary, and firms adopting just before or after it may not differ meaningfully in AI capabilities. Second, the post-2022 period coincides with rising interest rates, a broad technology-sector correction, and increased cost discipline across industries, all of which independently affect employment. We cannot fully separate the generative AI channel from these concurrent shocks, so the cohort result is best read as suggestive evidence of intensification rather than a definitive causal estimate of the generative AI effect.

Table 6: Pre-GenAI vs. Post-GenAI Adoption Cohorts

| | Emp growth | $p$-value | $N$ | Firms |
|---|---|---|---|---|
| *All matched firms* | | | | |
| Pre-GenAI (< 2023Q1) | −0.0047 | 0.331 | 55,790 | 1,717 |
| Post-GenAI (≥ 2023Q1) | $-0.0172^{***}$ | 0.001 | 57,403 | 1,915 |

*Notes:* This table reports cohort-matched DiD estimates for employment growth separately for firms whose first AI disclosure occurs before 2023Q1 (pre-GenAI) versus 2023Q1 or later (post-GenAI). Each cohort is compared against the same pool of never-treated control firms within the cohort-matched sample. All specifications include firm and quarter fixed effects. Standard errors clustered by firm. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

# 7 Heterogeneity by AI Use Type

We decompose the main employment result by AI use type. Table 7 presents the results.

**Use-type heterogeneity.** Our classifier categorises AI adoption as product-facing, internal, both, or unclear. On the cohort-matched sample, we interact the treatment indicator with use-type dummies in a single regression (Table 7, Panel A). Firms deploying AI for both product and internal purposes show the largest and most precisely estimated employment effect (−1.3 pp, $p = 0.016$), followed by product-facing adopters (−1.3 pp, $p = 0.061$). Internal-only adopters show a smaller and insignificant effect (−0.9 pp, $p = 0.14$). As a complementary exercise, we also estimate separate subsample DiDs restricting the treated group to a single use type (Panel B). All three categories show negative and marginally significant employment declines, with product-only at −1.2 pp ($p = 0.066$), always-internal at −1.1 pp ($p = 0.080$), and both-type at −1.0 pp ($p = 0.061$).

**Interpretation.** The use-type decomposition is suggestive rather than definitive. The individual use-type coefficients are not statistically distinguishable from each other. All three use types show negative employment effects of broadly similar magnitude (−0.9 to −1.3 pp in the interaction model), suggesting that the displacement is not driven by a single mode of AI deployment. The subsample DiDs confirm this pattern. The strongest result is for dual-purpose adopters (both product and internal), consistent with firms undertaking broader AI programmes experiencing larger workforce adjustments.

Table 7: Heterogeneity by AI Use Type

| Specification | Emp growth | SE | Treated firms |
|---|---|---|---|
| *Panel A: Use-type interactions (matched, all firms)* | | | |
| Product-facing | $-0.0125^{*}$ | (0.0067) | 298 |
| Internal only | $-0.0091$ | (0.0062) | 346 |
| Both | $-0.0127^{**}$ | (0.0053) | 582 |
| Unclear | $+0.0105$ | (0.0097) | 76 |
| *Panel B: Subsample DiD by use type (all firms)* | | | |
| Product-only firms | $-0.0124^{*}$ | (0.0067) | 298 |
| Always-internal firms | $-0.0110^{*}$ | (0.0063) | 346 |
| Both firms | $-0.0101^{*}$ | (0.0054) | 582 |

*Notes:* Panel A reports coefficients from a single regression interacting the treatment indicator with use-type dummies on the cohort-matched sample. Panel B reports separate subsample DiDs restricting treated firms to a single use type. All specifications include firm and quarter fixed effects. Standard errors clustered by firm. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

# 8 Robustness

Table 8 reports the employment growth coefficient across a range of alternative specifications. The key finding is that the negative effect is stable in sign, magnitude, and significance across all variants we consider.

*Matching specifications.* The primary cohort-matched estimate on all firms is $-1.1$ pp ($p = 0.003$). Restricting to non-technology firms yields $-1.0$ pp ($p = 0.010$). Restricting treatment to high-confidence classifications (confidence $\geq 80$) produces a slightly larger estimate of $-1.3$ pp ($p = 0.0005$), consistent with measurement error in the baseline treatment definition attenuating the effect toward zero. The stability of the coefficient across these specifications indicates that the result is not driven by a particular matching choice or sample restriction.

*Sector split.* When we split the matched sample by sector, non-technology firms show a significant employment decline of $-1.0$ pp ($p = 0.010$), while technology firms show a negative but insignificant coefficient ($-1.1$ pp, $p = 0.486$). The technology subsample contains only enough firms to yield 5,097 observations on the common sample, severely limiting statistical power; this is an imprecise null rather than a precise zero.

*Treatment definition.* We also consider a stricter definition of AI adoption that requires firms to mention AI in two consecutive earnings calls before treatment onset. This "sustained" indicator is then absorbing from that quarter onward. If the baseline treatment were picking up stray or inconsequential AI references, tightening the definition should attenuate the coefficient toward zero. Instead, the estimated employment decline rises to $-1.9$ pp ($p = 0.0001$) for all matched firms and $-1.6$ pp ($p = 0.002$) for non-technology firms, roughly 80 per cent larger than the baseline. This pattern is consistent with the employment effect being driven by persistent AI investment rather than one-off mentions. Revenue per employee under sustained treatment shows a marginally significant increase of $+1.6$ pp ($p = 0.088$), providing suggestive evidence that productivity gains may be emerging among firms with more committed AI adoption.

*Outlier treatment.* Our main results are reported without winsorisation. Winsorised versions (clipping at the 1st and 99th percentiles) produce slightly smaller but still significant estimates. The unwinsorised coefficient being larger confirms that outliers, if anything, strengthen the baseline effect rather than creating it.

*Secondary outcomes.* EBITDA margin, operating margin, and gross margin are all statistically insignificant across matched samples. Debt-to-assets shows a modest increase among non-technology matched firms, suggestive of firms taking on somewhat more leverage while implementing AI-related investment or restructuring. We treat this as a secondary pattern rather than a central result, but it is directionally consistent with transition costs around adoption.

Table 8: Robustness of the Employment Growth Effect

| Specification | Emp growth | SE | $p$-value | $N$ |
|---|---|---|---|---|
| *Matching specifications* | | | | |
| Primary cohort-matched, all firms | $-0.0107^{***}$ | 0.0036 | 0.003 | 77,018 |
| Non-tech only | $-0.0098^{**}$ | 0.0038 | 0.010 | 70,674 |
| High-confidence (≥80), all firms | $-0.0127^{***}$ | 0.0037 | 0.0005 | 77,018 |
| Tech only | −0.0111 | 0.0159 | 0.486 | 5,097 |
| *Treatment definition* | | | | |
| Sustained (consecutive), all firms | $-0.0191^{***}$ | 0.0049 | 0.0001 | 77,018 |
| Sustained (consecutive), non-tech | $-0.0162^{***}$ | 0.0052 | 0.002 | 70,674 |

*Notes:* Each row reports the cohort-matched DiD coefficient on employment growth under an alternative specification. The primary specification uses cohort-specific matching on all firms. The "sustained" rows redefine treatment onset as the first quarter in which a firm mentions AI in two consecutive earnings calls, rather than the first single mention; the indicator is absorbing from that point onward. All specifications include firm and quarter fixed effects. Standard errors clustered by firm. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

# 9 Discussion

This section discusses interpretation, mechanisms, and limitations.

**The revenue-per-employee puzzle.** The central puzzle is the absence of clear revenue-per-employee improvement alongside employment displacement. Standard task-based automation models predict that labour-replacing technologies should raise output per worker, at least in the medium run (Acemoglu and Restrepo, 2018). In our baseline specification, revenue-per-employee growth shows a positive but insignificant point estimate (+0.9 pp, $p = 0.24$). However, the puzzle is partially resolved under the sustained-treatment definition (requiring consecutive AI mentions), where RPE growth rises to +1.6 pp and becomes marginally significant ($p = 0.088$). This suggests that productivity gains may be emerging among firms with more committed AI adoption, while diluted in the full sample by firms with less intensive investment. We stress that revenue per employee is a labour productivity *proxy*, not a total factor productivity measure, and it conflates genuine efficiency gains with compositional shifts, pricing power, and exchange-rate effects for multinational firms.

Country-by-quarter fixed effects absorb aggregate price and exchange-rate movements, but firm-specific confounds remain. One interpretation, following Brynjolfsson et al. (2021), is that we are observing the early downward phase of a productivity J-curve, in which firms are incurring the adjustment costs of AI adoption (reorganisation, retraining, process redesign) without yet realising the productivity gains. The sustained-treatment result is consistent with a partial transition, in that firms that persist in AI adoption long enough to mention it in two consecutive calls may be further along the J-curve. Our dynamic evidence supports this view, as the employment effect is still growing more negative at the longest horizons we observe, suggesting the adjustment process is ongoing rather than complete. Whether productivity gains become more broadly significant is a question that future data will help resolve, as our panel continues to extend.

**Contraction versus efficiency.** An alternative interpretation is that AI adoption, at this stage, is not primarily about efficiency. The parallel decline in employment and revenue growth is more consistent with a restructuring narrative in which firms that adopt AI are simultaneously downsizing and reorganising, with both inputs and outputs contracting during the transition. The modest rise in debt-to-assets among matched non-technology firms is also consistent with this interpretation, suggesting that some firms may be funding AI adoption or related reorganisation with additional leverage rather than already realising clear operating gains. This could reflect several mechanisms. Firms facing competitive pressure or declining demand may adopt AI as part of a broader cost-cutting strategy, in which case AI is a response to, rather than a cause of, contraction. Alternatively, the implementation of AI systems may disrupt existing workflows sufficiently to reduce output temporarily, even as headcount is cut. Disentangling these mechanisms would require data on investment spending and organisational structure that we do not observe.

## 10 Conclusion

This paper constructs a novel panel of corporate AI adoption for over 13,000 global public firms and examines its effects on employment and productivity through end-2025. We show that naïve comparisons of AI adopters and non-adopters are uninformative because adopters are positively selected on pre-treatment characteristics. Using a cohort-specific propensity score matching design with rolling pre-treatment windows, AI adoption is associated with lower employment growth of approximately 1.1 percentage points per year. The effect is nearly twice as large ($-1.9$ pp) under a sustained-treatment definition requiring consecutive AI mentions, and is significantly larger among firms adopting during the generative AI wave ($-1.7$ pp) compared with earlier cohorts ($-0.5$ pp, insignificant), though we cannot fully separate this from concurrent macroeconomic forces. Revenue-per-employee growth shows no significant improvement in the baseline, but becomes marginally significant under sustained treatment (+1.6 pp, $p = 0.088$), providing suggestive evidence that productivity gains may be emerging among firms with more committed AI adoption.

These findings contribute to an emerging picture of AI's firm-level impact that is more nuanced than either the optimistic narrative (AI boosts productivity and creates jobs) or the pessimistic one (AI eliminates jobs but raises efficiency). In our data, job creation has not materialised at the firm level, and efficiency gains appear only tentatively among the most persistently adopting firms. Whether this reflects a transitory adjustment phase, a J-curve that will eventually turn upward, or a more durable pattern remains to be seen. As corporate AI investment continues to accelerate, the panel we construct can be updated to track these dynamics in real time.

We make the AI adoption panel publicly available to support further research. The dataset will be made available on the corresponding author's website and updated quarterly as new earnings call transcripts become available.

# References


**Acemoglu, Daron and Pascual Restrepo**, "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review*, 2018, *108* (6), 1488–1542.

**_ and _** , "Robots and Jobs: Evidence from US Labor Markets," *Journal of Political Economy*, 2020, *128* (6), 2188–2244.

**Babina, Tania, Anastassia Fedyk, Alex Xi He, and James Hodson**, "Firm Investments in Artificial Intelligence Technologies and Changes in Workforce Composition," in "Technology, Productivity, and Economic Growth," NBER, 2023.

**_ , _ , _ , and _** , "Artificial Intelligence, Firm Growth, and Product Innovation," *Journal of Financial Economics*, 2024, *151*, 103745.

**Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum**, "Artificial Intelligence and Firm-Level Productivity," *Journal of Economic Geography*, 2024. Forthcoming.

**Bloom, Nicholas, Tarek Alexander Hassan, Aakash Kalyani, Josh Lerner, and Ahmed Tahoun**, "The Diffusion of New Technologies," *NBER Working Paper*, 2024, (28999).

**Brynjolfsson, Erik, Abhiman Chandar, and Jiayi Chen**, "Canaries in the Coal Mine? Six Facts about the Recent Decline in Tech Employment," *Stanford Digital Economy Lab Working Paper*, 2025.

**_ , Daniel Rock, and Chad Syverson**, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics*, 2021, *13* (1), 333–372.

**_ , Danielle Li, and Lindsey R. Raymond**, "Generative AI at Work," *Quarterly Journal of Economics*, 2025, *140* (2), 889–938.

**Callaway, Brantly and Pedro H. C. Sant'Anna**, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

**Calvino, Flavio, Luca Samek, Mariagrazia Squicciarini, and Cecilia Morris**, "Closing the Italian Digital Gap: The Role of Skills, Intangibles and Policies," *OECD Science, Technology and Industry Policy Papers*, 2022.

**de Chaisemartin, Clément and Xavier D'Haultfœuille**, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

**Dell'Acqua, Fabrizio, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani**, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," *Harvard Business School Working Paper*, 2023, (24-013).

**Eisfeldt, Andrea L., Gregor Schubert, and Miao Ben Zhang**, "Generative AI and Firm Values," *NBER Working Paper*, 2023, (31222).

**Emanuel, Natalia, Emma Harrington, and Amanda Pallais**, "The Effects of AI on Workers: Evidence from a Field Experiment with Microsoft 365 Copilot," *NBER Working Paper*, 2025.

**European Investment Bank**, "AI Adoption, Productivity and Employment: Evidence from European Firms," *EIB Working Papers*, 2026, (2026/02).

**Goodman-Bacon, Andrew**, "Difference-in-Differences with Variation in Treatment Timing," *Econometrica*, 2021, *89* (5), 2291–2336.

**Graetz, Georg and Guy Michaels**, "Robots at Work," *Review of Economics and Statistics*, 2018, *100* (5), 753–768.

**Hassan, Tarek Alexander, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun**, "Firm-Level Political Risk: Measurement and Effects," *Quarterly Journal of Economics*, 2019, *134* (4), 2135–2202.

**Hougaard, Jes Leth et al.**, "The Short-Term Effects of Generative AI on the Labor Market," *University of Copenhagen Working Paper*, 2024.

**McElheran, Kristina, J. Frank Li, Erik Brynjolfsson, Zachary Kroff, Emin Dinlersoz, Lucia Foster, and Nikolas Zolas**, "AI Adoption in America: Who, What, and Where," *Journal of Economics & Management Strategy*, 2025. Forthcoming.

**Noy, Shakked and Whitney Zhang**, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 2023, *381* (6654), 187–192.

**Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer**, "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot," *Communications of the ACM*, 2023.

**Sautner, Zacharias, Laurence van Lent, Grigory Vilkov, and Ruishen Zhang**, "Firm-Level Climate Change Exposure," *Journal of Finance*, 2023, *78* (3), 1449–1498.

**Sun, Liyang and Sarah Abraham**, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

**Webb, Michael**, "The Impact of Artificial Intelligence on the Labor Market," *Stanford University Working Paper*, 2020.

**Yotzov, Ivan, Jose Maria Barrero et al.**, "Firm Data on AI," *NBER Working Paper*, 2025, (34836).

**Zolas, Nikolas, Nathan Goldschlag, Ron Jarmin, Paula Stephan, Jason Owen-Smith, Rebecca F. Rosen, Bruce Allen, Bruce A. Weinberg, and Julia I. Lane**, "Advanced Technologies Adoption and Use by U.S. Firms: Evidence from the Annual Business Survey," *NBER Working Paper*, 2024, (28290).

# Appendices

## A Additional Tables

Table 9: Propensity Score Matching Balance

| Covariate | SMD Before | SMD After |
|---|---|---|
| Log employment | 0.776 | 0.083 |
| Log revenue | 0.438 | 0.080 |
| Log assets | 0.355 | 0.072 |
| Debt-to-assets | 0.158 | 0.078 |
| Current ratio | 0.100 | 0.070 |
| Employment growth | 0.095 | 0.093 |
| Revenue growth | 0.079 | 0.094 |
| Gross margin | 0.072 | 0.094 |
| Operating margin | 0.066 | 0.131 |
| EBITDA margin | 0.066 | 0.132 |
| SG&A share | 0.060 | 0.144 |

*Notes:* This table reports standardised mean differences for each matching covariate before and after cohort-specific propensity score matching. Balance statistics are computed by pooling across cohort bins.

Table 10: Secondary Financial Outcomes

| Outcome | All matched | Non-tech | Tech |
|---|---|---|---|
| EBITDA margin | −0.97 (1.73) | −1.03 (1.86) | 1.03 (0.99) |
| Operating margin | −0.92 (1.74) | −0.98 (1.87) | 1.16 (1.12) |
| Gross margin | −0.34 (0.91) | −0.35 (0.98) | 0.02 (0.02) |
| Debt-to-assets | 0.010 (0.007) | 0.013* (0.007) | −0.032* (0.018) |

*Notes:* Matched DiD estimates for secondary financial outcomes on the cohort-matched sample. Columns correspond to all matched firms, non-technology firms, and technology firms. All specifications include firm and quarter FEs. Standard errors clustered by firm. Coefficients on margins are in percentage-point units.

Table 11: Distribution of AI Use Types Among Matched Treated Firms

| Use Type | Matched Treated Firms |
|---|---|
| Product only | 298 |
| Internal only | 346 |
| Both | 582 |
| Unclear | 76 |
| *Total treated* | *1,302* |

*Notes:* Firm counts on the common sample (all three outcomes non-missing). Use types are classified from the cumulative set of treated quarters. "Product" = firm's AI disclosures are exclusively product-facing. "Internal" = disclosures are internal operations only. "Both" = at least one product and one internal disclosure. "Unclear" = no clear product or internal classification.

# B Additional Figures

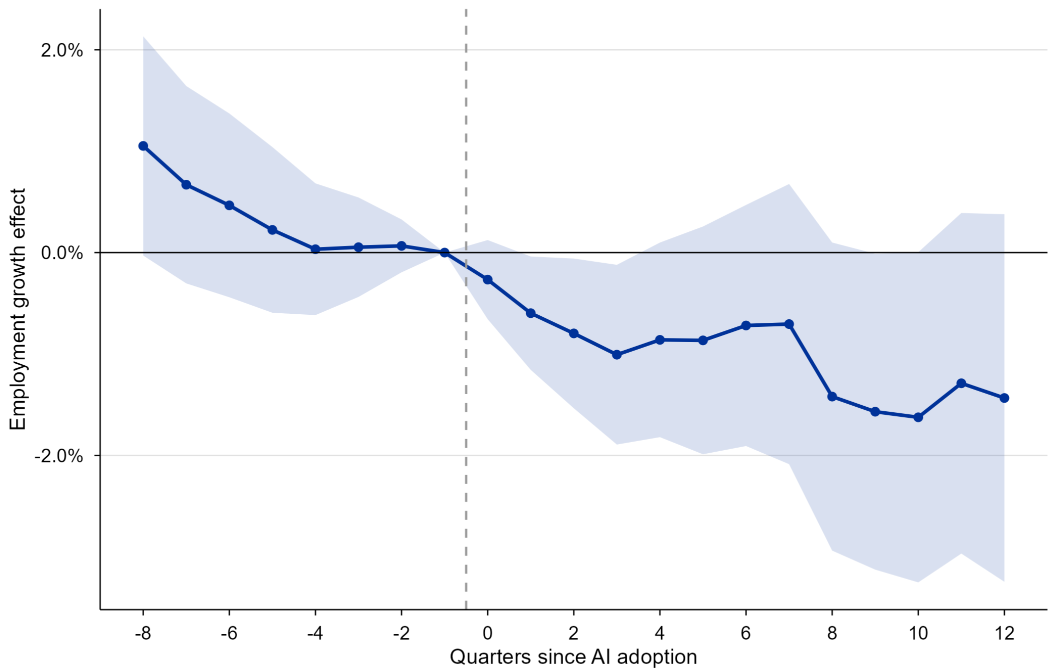



Figure 13: Dynamic Employment Event Study (PSM Matched Sample)

*Notes:* Sun–Abraham event-study coefficients estimated on the cohort-matched sample (all firms). Point estimates at long horizons are comparable to the full-sample event study (Figure 11) but confidence intervals are wider, reflecting the smaller matched sample.

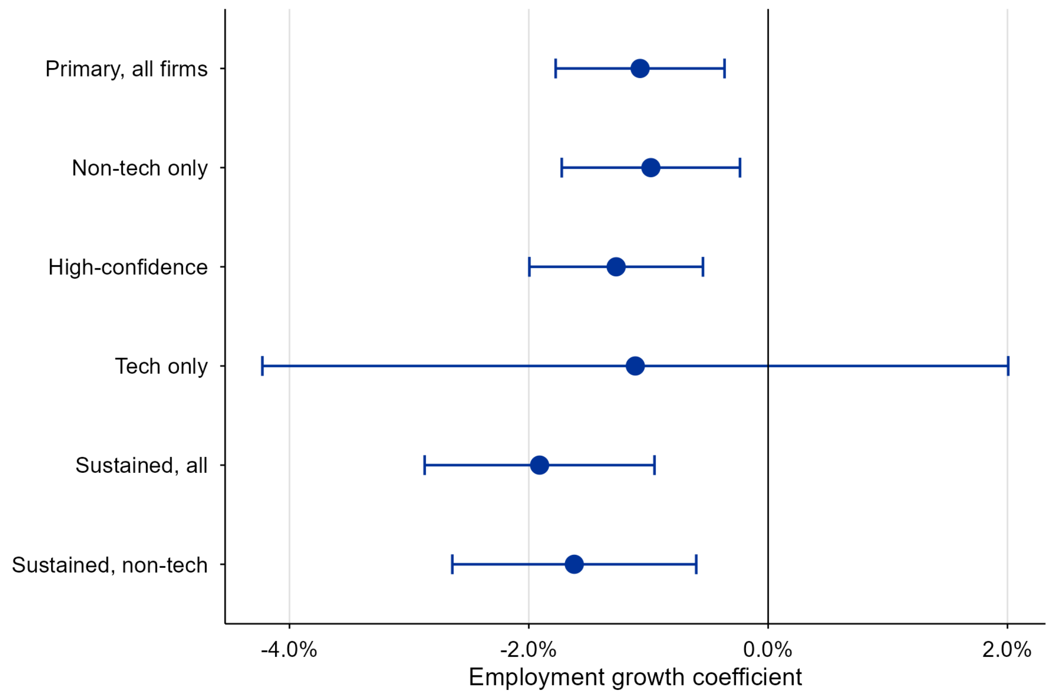



Figure 14: Employment Growth Effect Across Specifications
*Notes:* Point estimates and 95% confidence intervals for the employment growth coefficient across the robustness specifications reported in Table 8.

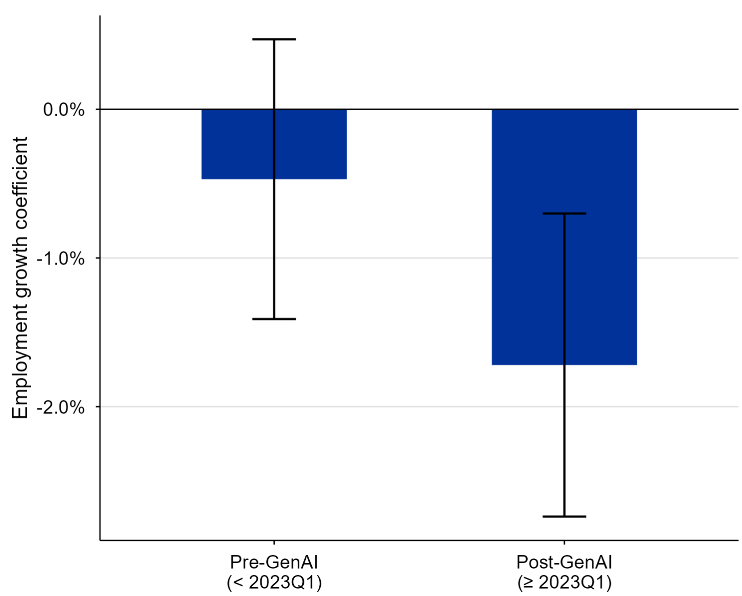



Figure 15: Employment Displacement by Adoption Cohort
*Notes:* Point estimates and 95% confidence intervals for the employment growth DiD coefficient, separately for pre-GenAI (adopted before 2023Q1) and post-GenAI (adopted 2023Q1 or later) cohorts on cohort-matched firms (all firms).

# C AI Classification Details

Table 12: Full Model × Prompt Classification Sweep

| Model | Prompt | Accuracy (%) | Precision | Recall | F1 |
|---|---|---|---|---|---|
| GPT-4o-mini | Prompt 1 | 85.0 | 0.835 | 0.889 | 0.861 |
| GPT-4o-mini | Prompt 2 | 69.6 | 0.637 | 0.973 | 0.770 |
| GPT-4o-mini | Prompt 3 | **87.0** | 0.866 | 0.889 | **0.878** |
| GPT-4o | Prompt 1 | 83.4 | 0.894 | 0.775 | 0.830 |
| GPT-4o | Prompt 2 | 78.8 | 0.724 | 0.962 | 0.826 |
| GPT-4o | Prompt 3 | 86.0 | 0.900 | 0.824 | 0.861 |
| GPT-4.1-mini | Prompt 1 | 86.0 | 0.831 | 0.920 | 0.873 |
| GPT-4.1-mini | Prompt 2 | 73.8 | 0.672 | 0.977 | 0.796 |
| GPT-4.1-mini | Prompt 3 | 86.6 | 0.868 | 0.878 | 0.873 |
| GPT-4.1 | Prompt 1 | 84.0 | 0.825 | 0.882 | 0.852 |
| GPT-4.1 | Prompt 2 | 78.8 | 0.725 | 0.958 | 0.826 |
| GPT-4.1 | Prompt 3 | 84.8 | 0.878 | 0.824 | 0.850 |
| GPT-5-mini | Prompt 1 | 77.8 | 0.707 | 0.985 | 0.823 |
| GPT-5-mini | Prompt 2 | 68.2 | 0.623 | 0.996 | 0.767 |
| GPT-5-mini | Prompt 3 | 81.4 | 0.744 | 0.985 | 0.847 |
| GPT-5.2 | Prompt 1 | 86.4 | 0.837 | 0.920 | 0.876 |
| GPT-5.2 | Prompt 2 | 74.8 | 0.677 | 0.992 | 0.805 |
| GPT-5.2 | Prompt 3 | 86.0 | 0.836 | 0.912 | 0.872 |

*Notes:* All 18 model–prompt combinations evaluated on 500 manually labelled sentences. Prompt 1 includes a confidence score and a three-month horizon instruction. Prompt 2 is a simplified binary classification without confidence. Prompt 3 adds explicit negative-class guidance. Prompt 2 achieves near-perfect recall but low precision across all models, indicating it labels too many sentences as positive. Prompt 3 generally improves precision relative to Prompt 1 at a small cost to recall. GPT-5-mini shows a similar pattern to Prompt 2, over-predicting positives across all prompts. All models use temperature = 0.

**Production prompt (Prompt 1).** The system message sent to GPT-4.1-mini for each earnings call sentence is reproduced below. The user message appends the sentence text (together with the immediately preceding and following sentences for context) after the instruction "Classify the following text per the instructions. Return ONLY the JSON object specified."

> *You are an expert in analysing earnings call transcripts for references to artificial intelligence (AI).*
>
> *Context: The text comes from an earnings call and can be either (a) a statement by company representatives (management), or (b) a question from analysts. When company representatives speak, 'we' and 'our' refer to the company. When analysts speak, 'you' refers to the company.*
>
> *Task: Classify the text as '1' (AI employed) if the company is currently deploying, using, implementing, integrating, piloting, or investing in AI or AI-powered tools in its operations, products, or strategy. Proprietary tools or internal applications count. Concrete hiring for AI projects counts. Classify as '0' (AI not employed) if the text contains general commentary about AI without evidence the company itself is deploying or investing in AI, or if it announces future plans that will clearly not materialise within the next three months.*
>
> *Provide an integer confidence score from 0 to 100 representing how strongly the snippet suggests evidence of AI employment. Confidence must be consistent with the label: higher than 50 if label = 1, lower than 50 if label = 0. Output as JSON: {"label": 0 or 1, "confidence": 0–100}.*

**Prompt 2 (simplified).** This prompt omits the confidence score and the three-month horizon instruction, testing whether the classifier performs comparably with a simpler task definition.

> *You are an expert in analyzing earnings call transcripts for references to artificial intelligence (AI).*
>
> *Context: The text comes from an earnings call and can be either: (a) a statement by company representatives (management), or (b) a question from analysts or other audience members directed at the company. When company representatives speak, 'we', 'our', and the company name refer to the company itself. When analysts or audience members speak, 'you' and the company name refer to the company.*
>
> *Task: Classify the text as: '1' (AI employed) if the company is deploying/using/implementing/integrating/pilo in AI or AI-powered tools in its operations, products, or strategy. Proprietary tools or internal applications count. '0' (AI not employed). Output STRICTLY as valid JSON: {"label": 0 or 1}.*

**Prompt 3 (expanded negative examples).** This prompt adds explicit negative-class guidance, listing specific types of statements that should receive label = 0. It also adds hiring for AI projects and concrete pilots to the positive class.

> *You are an expert in analyzing earnings call transcripts for references to artificial intelligence (AI).*
>
> *Context: [Same as Prompt 2.]*
>
> *Task: Classify the text as: '1' (AI employed) if the company is deploying/using/implementing/integrating/pilo in AI or AI-powered tools in its operations, products, or strategy. Proprietary tools or internal applications count. This includes mentions of hiring staff to work on AI-related projects, concrete pilots, or systems already being developed/rolled out. '0' (AI not employed): General or abstract commentary about AI (industry trends, competitor activity, customer demand) without evidence the company itself uses, invests in, or implements AI. Vague future plans, intentions, or aspirations to use AI without indication that AI projects are planned. Mentions of building, expanding, or investing in data centres where it is not specifically stated these will be used for AI systems or AI workloads internal to the company. Output STRICTLY as valid JSON: {"label": 0 or 1}.*

**Illustrative excerpts.** Table 13 presents examples of earnings call sentences from the validation set to illustrate the types of language the classifier encounters. Panel A shows sentences manually labelled as genuine AI adoption. These range from concrete product deployments (Concentrix, XPO) to internal process automation (Max Financial Services) and hiring for AI projects (Hinge Health). Panel B shows sentences that mention AI but were manually labelled as *not* reflecting current adoption. These include analyst questions about AI strategy (Exlservice), firms with "AI" in their company name (SES AI Corp, Perimeter Medical Imaging AI), and forward-looking aspirational statements without concrete current deployment (Infineon Technologies).

Table 13: Illustrative Earnings Call Excerpts

| Firm | Sector | Excerpt |
| --- | --- | --- |
| *Panel A: Genuine AI adoption (manual label = 1)* | | |
| Concentrix Corp | Industrials | "This complements our autonomous AI assistant product, iX Hello, giving clients an array of AI solutions to meet their needs for both full automation and human augmentation." |
| XPO Inc | Industrials | "Recently, we deployed new AI-driven linehaul models, designed to improve freight flows across our network." |
| Max Financial Services | Financials | "Additionally, a gen AI based SEWA bot pilot was launched for customer service team query resolution specific to queries on SEWA products." |
| Hinge Health Inc | Healthcare | "And he is driving personally from the founder level, AI adoption across our R&D organization, and we are measuring how many of our engineers are using AI tools to build their products." |
| London Stock Exchange Group | Financials | "We integrated some of the TORA functionality, and we launched the AI-driven advanced dealing for FX." |
| *Panel B: Not genuine AI adoption (manual label = 0)* | | |
| GEE Group Inc | Industrials | "The only thing I would add is, obviously, AI is top of mind for everyone today." |
| Exlservice Holdings | Industrials | "But as you think about the impact of AI as opposed to the growth opportunity around new opportunities." |
| SES AI Corp | Technology | "Hello everyone and welcome to the SES AI first-quarter 2025 earnings release and call." |
| Infineon Technologies | Technology | "From a midterm perspective, growth dynamics remain favorable, underpinned by structural trends, in particular, the rapid proliferation of artificial intelligence is an extremely attractive opportunity for us." |

*Notes:* Examples drawn from the 500-sentence manual validation set. Panel A sentences describe concrete, current AI deployment, integration, or investment. Panel B sentences mention AI but reflect generic commentary, analyst questions, company names containing "AI", or aspirational statements without evidence of current adoption.

**Use-type classification.** For sentences classified as AI-positive, a second LLM prompt classifies the AI use as product-facing, internal, both, or unclear. Against the 262 manually labelled positive sentences, the use-type classifier achieves 67.9% accuracy, with per-category accuracy of 72.9% for product (n = 166), 75.0% for internal (n = 24), 62.5% for both (n = 32), and 47.5% for unclear (n = 40). The lower accuracy on "both" and "unclear" categories reflects the inherent ambiguity in distinguishing overlapping use types from short text segments.

Table 14: AI Keyword List for Boolean Pre-Filtering

| Category | Terms |
|---|---|
| General AI terminology | ai, artificial intelligence, generative-ai, genai, gen-ai, machine learning, deep learning, natural language processing, nlp, computer vision, augmented intelligence, aigc |
| Foundation & generative models | generative model*, foundation model*, large language model*, llm*, diffusion model*, transformer model* |
| Specific LLM families | ChatGPT (incl. 3.5, 4, 4o), GPT (incl. 2, 3, 3.5, 4, 4-turbo, 4o), Claude (incl. 2, 3 Haiku/Sonnet/Opus), Gemini (incl. 1.5 Flash/Pro, Ultra 1.0, Nano), PaLM (incl. 2, PaLM-E), Bard, Llama (incl. 2, 3 variants), Falcon (7B/40B/180B), Mistral (incl. 7B, Large), Mixtral (8x7B/8x22B), Jurassic (1/2 variants), Jamba, Gemma (2B/7B), Grok (incl. 1, 1.5, 4), ERNIE Bot (4.0/4.5), Qwen (1.5/2 variants), Tongyi Qianwen, Yi (34B/9B), Phi (3 Mini/Small/Medium), Command R/R+, PPLX (7B/70B variants) |
| Image generation models | DALL-E (2/3), Stable Diffusion (v1.5/v2.1/XL/Turbo), Midjourney (v1–v6.1), Firefly (Image 1–4, Video), Imagen (2/3), Ideogram (v1–v3) |
| Video generation models | Runway (Gen-1/Gen-2/Gen-3 Alpha), Pika (1.0/1.1/1.5), Sora, Veo |
| Application terms | text-to-image, text to image, text-to-video, text to video, code completion, retrieval-augmented generation, chatbot*, copilot, co-pilot, prompt engineering |
| Technical vocabulary | neural network*, reinforcement learning, supervised learning, unsupervised learning, backpropagation, hyperparameter*, hallucination*, supercomput*, quantum comput* |

*Notes:* The 176 terms used for boolean pre-filtering of earnings call sentences. An asterisk (*) denotes wildcard matching (e.g., "chatbot*" matches "chatbots" and "chatbot-based"). Model family entries are grouped for readability; each variant listed in parentheses is a separate keyword. Case-insensitive matching is applied for all terms.