The Generative AI Turn in the Productivity J-Curve
Firm-Level Evidence from Earnings Calls*

Giulia Maria Guerrini† Jonathan Rice‡

July 20, 2026

Abstract

Does adopting artificial intelligence pay off? Based on sixteen years of earnings calls from more than 10,000 listed firms in 88 countries, we date each firm's first reported use of AI and follow adopters against closely matched firms that had not adopted. For a decade, adoption looked like investment without return. Adopters grew more slowly in employment and revenue and showed no productivity gain, however long they persisted. The return arrived only when the technology improved. From the first quarter after ChatGPT's release, the productivity of firms that had adopted AI prior to its release rose by about two percentage points a year relative to their matches. No earlier date shows anything similar, and the gain is not explained by the passage of time since adoption. Firms adopting for the first time after the release of ChatGPT no longer pay the early adjustment costs, though they show no clear gain of their own yet. Slower employment growth accompanies adoption throughout. The recovery half of the productivity J-curve was therefore not automatic. It waited for the technology to become capable enough, and it arrived first at the firms that had already reorganised their data, processes, and skills around AI, groundwork laid for earlier systems and not necessarily in anticipation of generative AI.

Keywords. artificial intelligence, technology adoption, firm productivity, intangible capital, difference-in-differences, text as data.
JEL codes. O33, D22, J23, E22, C55.

*The views expressed are the authors' own and do not represent those of the Central Bank of Ireland, the European Central Bank (ECB), or the Eurosystem. We thank participants at the June 2026 meeting of the ECB Working Group on Forecasting and the European Systemic Risk Board (ESRB)–Bank of Finland 2026 Conference on AI and Systemic Risk for their helpful comments. We also thank Frederik Andersen and Lisa Chardon-Denizot for their excellent research assistance and Markus Schwedeler for provision of NL Analytics data. Transcript data are from NL Analytics; financial data are from S&P Capital IQ.

†European Central Bank. giulia_maria.guerrini@ecb.europa.eu.
‡Central Bank of Ireland. jonathan.rice@centralbank.ie.

Non-technical summary

Artificial intelligence is widely expected to raise productivity, yet for most of the past decade the gains have been hard to find in firm data. This paper asks when adoption began to pay off, and for whom. The answer matters for how forecasters and policymakers read the current wave of investment in generative AI, because the history of general-purpose technologies suggests that firms bear the costs of adoption long before they see measurable returns, a pattern known as the productivity J-curve.

We measure adoption from what firms tell their investors. Based on sixteen years of quarterly earnings calls from more than ten thousand listed firms in 88 countries, we identify the quarter in which each firm first reports actually using artificial intelligence in its business, as distinct from merely discussing it. We then compare each adopter with a closely matched firm in the same sector that had not adopted at that date, and follow both over time.

Through the decade before generative AI arrived, adoption brought costs without a visible return. Adopting firms grew more slowly in employment and revenue than their matches and showed no gain in labour productivity, and the picture did not improve the longer they persisted. This is the first half of the J-curve, the years in which firms reorganise work around a new technology at a cost, and for a decade there was little sign that the promised recovery would arrive.

The recovery and productivity gain came when the technology improved with the release of ChatGPT. From the first quarter of 2023, the firms that had already built around artificial intelligence, which the paper calls the installed base, recorded labour-productivity growth roughly two percentage points a year faster than their matched comparators. No earlier date shows a comparable movement, and the timing lines up with the arrival of generative AI rather than with how long any individual firm had been adjusting. We interpret this finding as the gains accruing to firms whose complementary systems and skills were already in place when the technology became capable enough to reward them. We infer that investing in earlier AI technologies required these firms to make deep internal changes that gave them a head start once ChatGPT arrived. This could have included organising and cleaning their data so that machines could use it, redesigning workflows so that a model's output feeds a decision, hiring engineers and analysts who understand what the technology can and cannot do, and giving their managers years of practice in judging where it adds value in their business. These changes were not necessarily made in anticipation of generative AI; much of the groundwork would have served the AI variants firms were already using. When far more capable models arrived, firms with these structures in place would have been better positioned to benefit, while others may still have had the groundwork ahead of them.

The release also changed who adopts. The share of firms adopting AI for the first time each year roughly doubled after the release of ChatGPT, and adoption spread from large, technology-centred, disproportionately American firms towards smaller, more ordinary firms across sectors and countries. By the end of 2025, firms reporting adoption accounted for a large majority of listed-sector employment and market value. These newer adopters do not repeat the early wave's slow and costly adjustment, but they show no

clear productivity gain of their own yet.

The message for the aggregate outlook is that productivity gains from artificial intelligence are neither automatic nor evenly spread. They arrived only when the technology crossed a threshold of capability, and first at the firms that had spent time preparing for it. Throughout, adopters' employment grows more slowly than that of comparable non-adopters, so the gains reflect output rising relative to labour input rather than expansion in hiring. Whether the far larger group of recent adopters eventually converts today's accessible tools into measured productivity is the key margin to watch, and it will determine how much of the current investment wave appears in aggregate output.

1 Introduction

Artificial intelligence has diffused rapidly through firms, yet its effects on firm performance remain uncertain. The release of ChatGPT in late 2022 marked a visible change in the capabilities and accessibility of the technology, and generative AI subsequently spread through work at a pace comparable to earlier general-purpose technologies (Bick et al., 2026). Rapid diffusion need not translate immediately into measured productivity. Firms must identify suitable tasks, redesign workflows, train workers, integrate data and software, and establish new forms of oversight before a new technology can be used effectively. These activities absorb resources before their benefits are fully realised or recorded, producing the temporary productivity slowdown described by the productivity J-curve (David, 1990; Brynjolfsson et al., 2019, 2021). Generative AI models may alter that adjustment path by making useful applications cheaper to identify and deploy, but they may also raise complementary investment as firms attempt more ambitious reorganisations. The same process may affect employment and revenue as firms reorganise, and it may differ between firms that had built AI capabilities before generative AI models became widely available and those that adopted later. Looking only at average outcomes after 2022 would conflate this technological shift with changes in the composition and readiness of adopters. This paper asks whether the arrival of generative AI changed the firm-level returns associated with reported AI adoption, and in particular whether the recovery that the J-curve predicts arrived on each firm's own adjustment clock or on the technology's.

We address this question by constructing a dated measure of first reported AI deployment or committed investment from sixteen years of earnings-call transcripts for more than 10,000 listed firms in 88 countries. A broad keyword screen first identifies sentences that refer to artificial intelligence. A language-model classifier then reads each flagged sentence, with one sentence of context on either side, and judges whether it shows the firm actually deploying or committing investment to AI rather than offering general commentary, aspiration, or exposure. Alongside that judgement it returns a score from 0 to 100 for how strongly the passage indicates adoption, and we date a firm's adoption event at the first quarter in which any sentence receives a positive judgement with a score of at least 80. We also verify the measure manually. On 500 sentences labelled by hand, the classifier agrees with the human consensus close to nine times in ten. This procedure yields a consistent quarterly signal across the periods before and after ChatGPT and distinguishes operationally relevant statements from the much larger set of AI mentions.

To compare reporting adopters with plausible counterfactual firms, we construct a matched panel. Each adopter is paired within sector and calendar quarter with firms that did not disclose AI adoption and that were similar in size, prior employment and revenue growth, profitability, leverage, and related observable characteristics. The design therefore addresses large observable differences between firms that report AI activity and those that do not, while preserving the timing of each reporting event. It also permits event-time comparisons over a common local window rather than attributing all differences between eventual adopters and non-adopters to AI. Our preferred specification uses the interaction-weighted staggered-adoption estimator of Sun and Abraham (2021); conventional static two-way fixed-effects estimates are retained only as sen-

sitivity summaries. The regressions use matching weights and compare changes within physical firms over common calendar quarters. They identify matched-sample associations under conditional parallel trends, not the effect of randomly assigned AI adoption. In particular, firms choose when to disclose, and that choice may reflect managerial strategy or expected performance that we cannot observe. Alternative matching windows and staggered-adoption estimators, together with richer time effects, are reported as robustness exercises.

To make the comparison across technological eras explicit, we divide reporting adopters at 2023Q1, the first full quarter after ChatGPT's release and our empirical boundary for the widespread availability of generative AI. Firms whose first high-scoring disclosure occurs before 2023Q1 form the installed base entering 2023, and we observe their outcomes on both sides of that boundary; those whose first occurs from 2023Q1 form the new reporting-adopter population in the generative AI era, and their post-adoption outcomes fall from 2023Q1 onwards by construction.

Three findings organise the empirical results. First, before 2023Q1, reported adoption is followed by slower employment and revenue growth than at matched firms, with no improvement in labour productivity, and the shortfall does not fade with time since adoption. The descending half of the J-curve is visible in the data; the ascending half is not. Second, the recovery arrives at a calendar date rather than at any point determined by time since adoption. From the first full quarter after ChatGPT's release, the labour productivity of the installed base rises by about two percentage points a year relative to its matches. The change is not explained by time since adoption, holds across the adoption vintages we can separate, and appears at no earlier boundary we test. The slower relative employment growth that accompanies adoption shows no such turn. The installed base's employment differential does not narrow at the release, if anything widening slightly, so the productivity gain reflects output rising relative to labour input rather than a recovery in relative hiring. Third, the release changed who adopts. Entry roughly doubles from 2023, the firms entering are about half the size of the earlier adopters and less concentrated in the United States, and they sit at the same point of their sectors' productivity distributions as the adopters that came before them. Their post-adoption record through 2025 shows neither the revenue and productivity costs earlier cohorts paid nor a significant gain of their own.

These findings contribute in two ways. On measurement, the paper provides a validated, high-frequency event date for firms' first reported deployment or committed investment, complementing surveys, measures of workforce skills and vacancies, and text measures of exposure or attention (Babina et al., 2024; McElheran et al., 2024; Kalyani et al., 2025). The regular quarterly event measure is well suited to observing dynamics, and its long pre-generative-AI history also aids the central comparison. Discussion of AI surged after ChatGPT, and the firms disclosing adoption after the release may differ from those disclosing before it. A measure built only from that period could therefore not tell whether measured returns changed because adoption began to pay off or because the population of disclosing firms changed. Because the installed base is identified from disclosures made years before the release, the change in its outcomes at the boundary reflects a change in the returns to adoption rather than the surge in AI discussion that followed. On economic

interpretation, the results revise the productivity J-curve account rather than confirming it. In the standard account, returns arrive once each firm's complementary investment matures. In our data the costs behave that way, but the returns do not. They arrive together, at the public release of generative AI, and firms adopting after the release show no sign of the costly phase over the horizons we observe. The payoff was waiting on the capability of the technology, not on the completion of each firm's adjustment. Other developments after ChatGPT may contribute to the pattern, so the paper offers evidence on when the returns to adoption changed rather than a claim that AI alone caused the outcomes.

The paper proceeds as follows. Section 2 surveys the literature on general-purpose technologies and firm AI adoption. Section 3 describes the text measure, validation, outcomes, and sample. Section 4 sets out the matched empirical design. Section 5 presents the average patterns, the installed-base turn, and the profile and early record of the firms adopting after the release. Section 6 assesses robustness and interpretation, and Section 7 concludes.

2 Related literature

The paper first relates to research on general-purpose technologies, complementary investment, and delayed productivity gains. David (1990) linked the lag between electrification and factory productivity to the reorganisation of production, while Bresnahan and Trajtenberg (1995) emphasised downstream co-invention. Models of learning, adjustment, and organisation capital generate similar delays (Helpman and Trajtenberg, 1998; Greenwood and Yorukoglu, 1997; Atkeson and Kehoe, 2007), and firm evidence shows that information technology is most valuable when combined with workplace reorganisation and skills (Bresnahan et al., 2002). Brynjolfsson et al. (2021) bring these mechanisms together in the productivity J-curve. Firms accumulate complementary intangible capital before its output is fully realised or measured. This framework has both a real and a measurement implication. Implementation can temporarily divert labour and management effort from current production, while internally generated software, training, and organisational capital may not appear as investment in the conventional accounts (Corrado et al., 2009; Eisfeldt and Papanikolaou, 2013). We apply this perspective to AI by dating first reported deployment or committed investment and asking whether the associated growth pattern changes once generative AI models become widely available. The installed-base comparison is particularly informative for this interpretation because it takes the firms that had already disclosed AI adoption before ChatGPT's release and asks how their outcomes changed at that common technological date. The evidence is consistent with a shorter early productivity drag, but it does not directly observe complementary investment or identify a single organisational mechanism.

A second literature develops measures of firm AI adoption. Official surveys and censuses show that adoption is concentrated among larger firms and varies substantially across industries, although such measures are often self-reported or observed infrequently (Zolas et al., 2020; McElheran et al., 2024; Calvino and Fontanelli, 2023; Bonney et al., 2024). Labour-market measures infer investment from AI skills in employee profiles and vacancies (Alekseeva et al., 2021; Babina et al., 2024), while exposure indices describe

the tasks or occupations that AI could affect (Webb, 2020; Felten et al., 2021; Eloundou et al., 2024). These approaches measure different margins. Surveys ask about use, labour-market data capture specialised inputs, and exposure measures describe technical potential rather than realised deployment. Corporate text provides a complementary source of dated information. Prior work uses filings and earnings calls to measure product-market similarity, political risk, climate exposure, and technological diffusion (Hoberg and Phillips, 2016; Hassan et al., 2019; Sautner et al., 2023; Kalyani et al., 2025). Recent studies use corporate language to study firms' proximity to frontier AI research, exposure to generative AI, and managers' anticipated responses to ChatGPT (Soto, 2025; Ca' Zorzi et al., 2025; Jha et al., 2024). Earnings calls provide high-frequency timing and management's description of current activity, but they also reflect disclosure incentives and investor attention. Our measure has a narrower purpose. It validates sentence-level evidence and dates the first disclosure that reports deployment or committed investment rather than attention alone. This distinction matters because managerial language can be promotional and because a public report need not coincide with first operational use (Barrios et al., 2025).

The firm-level evidence on AI, productivity, and employment is mixed. Measures based on skills and vacancies associate AI investment with faster sales and employment growth, product innovation, and workforce upgrading, but not with a detectable increase in revenue per worker (Babina et al., 2025, 2024). Survey and patent measures often find positive productivity differences among adopters (Czarnitzki et al., 2023; Yang, 2022), and Aldasoro et al. (2026) estimate a labour-productivity gain using matched European investment-survey data. Other survey and administrative evidence finds limited recent effects on employment, earnings, hours, or measured productivity even where tasks change (Yotzov et al., 2026; Humlum and Vestergaard, 2025). Studies based on occupational exposure also find labour-market adjustment concentrated in particular establishments, tasks, and worker groups rather than a uniform aggregate response (Acemoglu et al., 2022; Hampole et al., 2025; Brynjolfsson et al., 2025a). These findings need not conflict if returns depend on adoption timing, complementary investment, or the margin being measured. Our long quarterly panel adds evidence on dynamics around first reported adoption. It shows the clearest positive productivity association for the installed base entering 2023, and it shows that firms first adopting in the generative AI era no longer pay the early costs that earlier cohorts paid. The slower relative employment growth in our baseline is a difference between matched firms rather than evidence of aggregate displacement, and automation effects may well contribute to it. We do not read displacement as the overriding explanation, because the gap does not move significantly when far more capable models arrive, but task-based theories in which automation displaces labour from some activities while creating or complementing others accommodate both readings (Acemoglu and Restrepo, 2018; Ide and Talamàs, 2025).

Finally, the paper helps connect large task-level experimental gains with slower firm-level outcomes. Experiments find sizeable improvements in speed or quality when generative AI is applied to tasks within the model's capabilities, including writing, customer support, and coding (Noy and Zhang, 2023; Brynjolfsson et al., 2025b; Peng et al., 2023; Cui et al., 2026). Benefits vary with worker skill and task type, however, and performance can worsen outside the technology's effective frontier (Dell'Acqua et al., 2026; Otis et al.,

2023; Becker et al., 2025; Dillon et al., 2025). Translating a task gain into firm productivity also requires adoption, workflow redesign, data integration, oversight, and complementary skills; some of this investment may be intangible and absent from conventional balance sheets. Exposure-based aggregate forecasts must therefore make assumptions about task coverage, adoption, and the share of technical gains that firms convert into output (Acemoglu, 2025; Filippucci et al., 2024). Consequently, broad exposure and rapid diffusion can coexist with modest average firm effects and substantial heterogeneity. Our results offer one explanation for this gap. Task-level gains become firm-level productivity only after a firm has built the systems, workflows, and skills to use the technology, and in our data the gains from generative AI arrived first at the firms that had already done that building.

3 Data and measurement

3.1 Measuring reported AI adoption

Most listed firms hold earnings calls on a regular reporting schedule, and the dated transcripts record management's account of current operations and investment. We use these calls to identify the quarter in which a firm first reports deploying AI or making a committed AI investment. The measure has two stages. First, a deliberately broad keyword screen selects sentences containing general AI terms, named model families and products, or application terms. All transcripts are analysed in English, with calls originally delivered in another language translated into English before the screen is applied. The screen is intended to retain genuine mentions even at the cost of admitting false positives. A sentence can mention AI while discussing a competitor, an analyst's general question, or a possible future development. Appendix A reports the complete keyword list and the audit used to remove ambiguous terms.

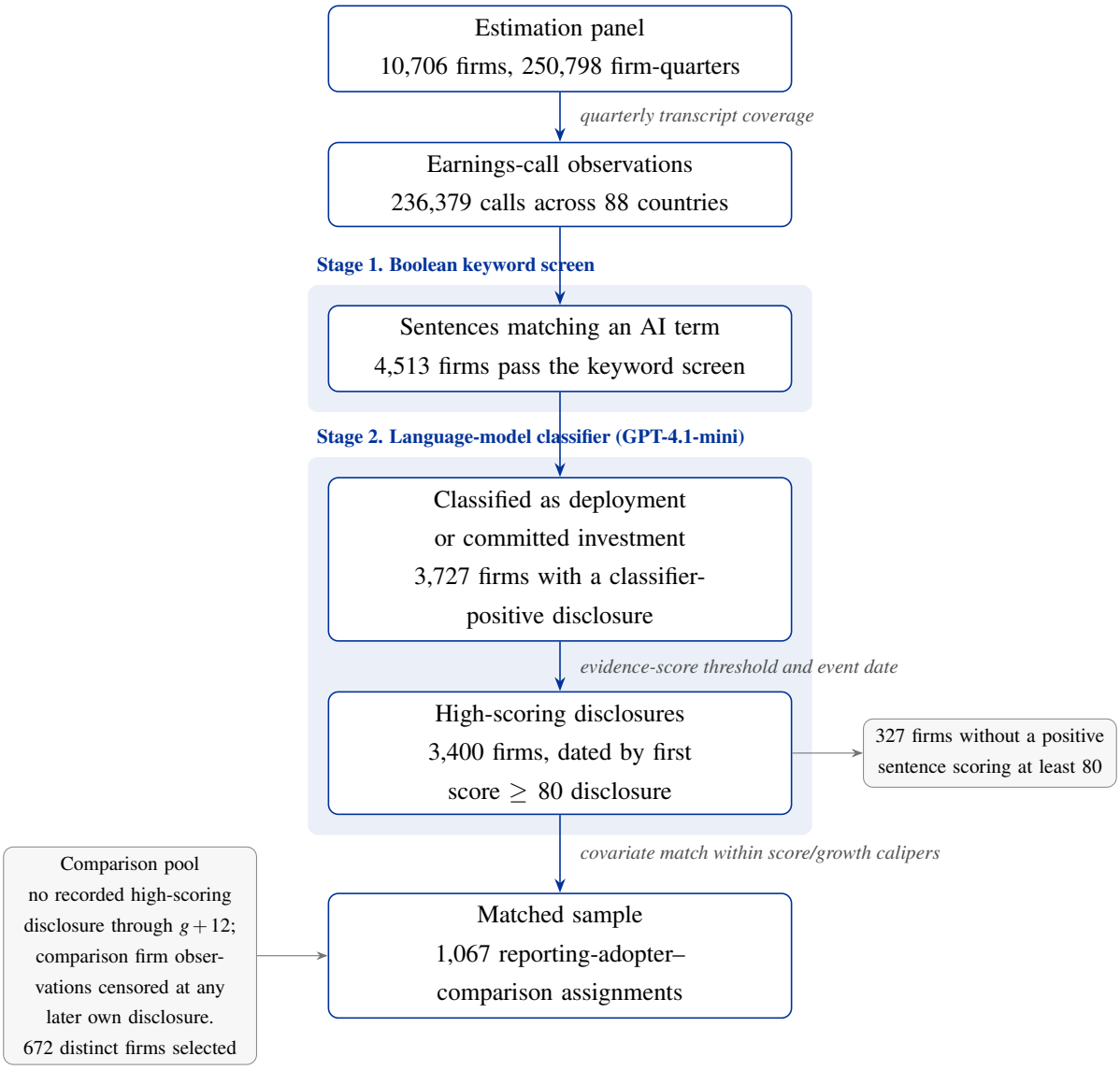
Second, GPT-4.1-mini classifies each screened sentence including its surrounding sentences for context. A positive label requires evidence of the firm's own current deployment, use, integration, pilot, or committed investment in AI within its operations, products, or strategy. Internal tools and proprietary systems count, but general commentary, aspiration without a commitment, and reference to another firm's activity do not. The model sees the flagged sentence together with the sentence before and the sentence after it, and earnings-call-specific instructions concerning speaker references and active deployment, but not the rest of the transcript. It returns a binary label and an evidence score from 0 to 100. The score measures the strength of textual support for the classification, with 90–100 indicating near certainty that the firm is disclosing current adoption of AI and 0–10 near certainty that it is not. Appendix B shows the production prompt used.

We filter AI-adopting firms as those with a positive AI disclosure quarter and a score of 80 or above, which reduces ambiguous events while retaining enough firms for the matched design. The threshold was chosen using the validation evidence below, rather than to imply a probabilistic cut-off. The first qualifying quarter supplies the event date even if the firm continues to discuss AI in later calls. Figure 1 summarises this measurement and sample pipeline. Of the 3,400 firms with a dated high-scoring disclosure, 1,557 have

the required pre-event covariates and 1,067 obtain a comparison firm within the matching calipers.

A single qualifying sentence is sufficient to date a firm-quarter, and additional sentences do not move that event date.

Figure 1: From the panel to the matched sample



Notes. The shaded bands mark the keyword and classification stages. A comparison firm's observations are censored at its own first high-scoring disclosure. The event is dated by the first classifier-positive sentence scoring at least 80.

The resulting measure is at the firm-quarter level. A firm can have no qualifying disclosure for many years and then enter the reporting-adopter sample when it first crosses the rule. The measure concerns deployment or committed investment, rather than sentiment, expected exposure, or the frequency of generic AI discussion. The validation exercise in the next subsection assesses sentence-level classification accuracy, and the empirical design compares outcomes between treated firms, those reporting high-scoring AI adop-

tion, and control firms, which include firms whose earnings calls contain no AI keywords at all and firms whose sentences the classifier judges not to indicate adoption.

3.2 Validation

We validate the classifier against 500 randomly sampled screened sentences. The two authors and two research assistants labelled each sentence independently, based on whether they believed the text implied the firm was currently adopting or investing in AI. They then resolved disagreements to produce a consensus benchmark. The validation set is close to balanced, with 52 per cent positive sentences, so the reported accuracy does not arise from a predominantly negative sample.

The production classifier agrees with the consensus in 87 per cent of cases. Its precision is 85 per cent, its recall is 91 per cent, its F_1 score is 0.88, and Cohen's chance-corrected agreement is $\kappa = 0.74$ (Table 1). Across the five alternative frontier models and two prompt variants we tested, the production combination of model and prompt offered the best balance of performance and cost. Appendix B reproduces the production prompt and both alternatives verbatim and summarises the model comparisons.

The validation sentences are drawn from sentences that had already passed the keyword screen, so the agreement rates measure the classifier, not the keyword list. Two things therefore sit outside the validation. A firm that uses AI but never discusses it on a call cannot be detected by any text measure, and a genuine AI discussion phrased without any screened term would be missed; the deliberately broad vocabulary and the keyword audit in Appendix A limit the second risk.

Table 1: Classifier performance against human labels

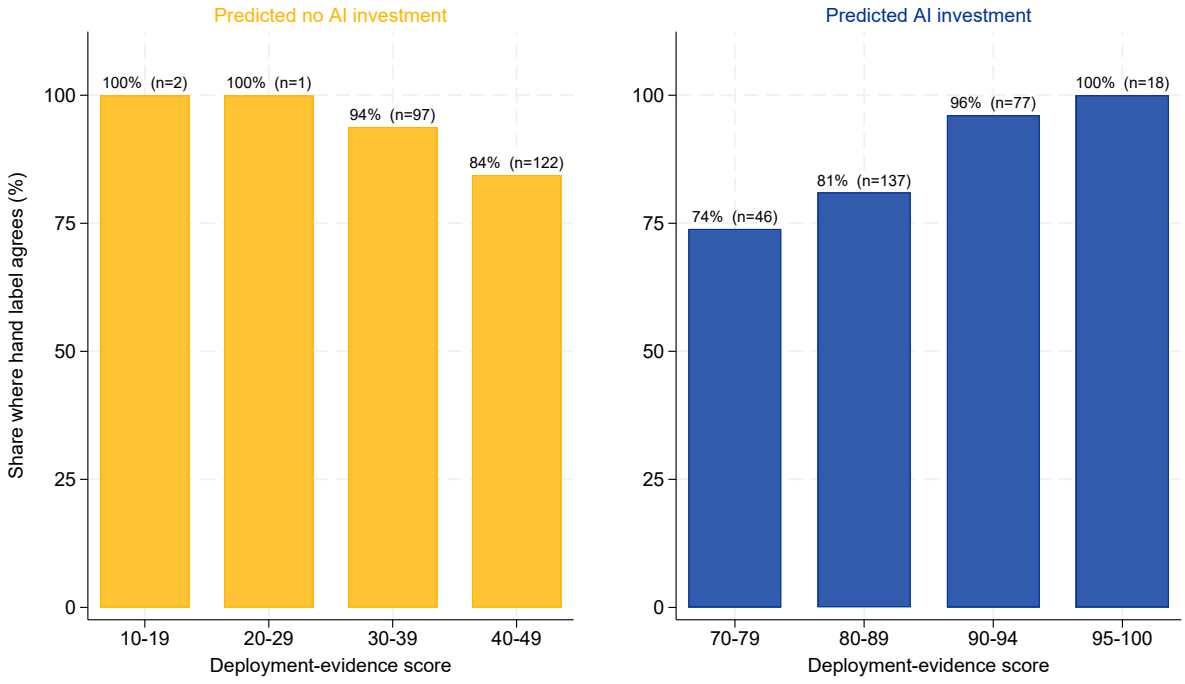
Overall (500 sentences)		*At the deployment-evidence threshold*	
Accuracy	0.868	Precision, score ≥ 80 ($n = 232$)	0.875
Precision	0.853	Precision, score ≥ 90 ($n = 95$)	0.968
Recall	0.905	*Human positive rate*	0.52
F_1	0.878	*Cohen's* κ	0.735

Notes. GPT-4.1-mini predictions compared with consensus human labels for 500 randomly drawn screened sentences. Precision is the share of classifier-positive sentences confirmed by the human label; recall is the share of human-positive sentences recovered; κ is Cohen's chance-corrected agreement. The final column reports precision at or above the stated evidence-score threshold.

Figure 2 shows how agreement with the consensus human label varies with the evidence score, separately for sentences the classifier judges to indicate adoption and sentences it judges not to. Within each panel the relationship is monotonic. Among predicted negatives, agreement rises as the score falls, from 84 per cent in the 40–49 band to 94 per cent at 30–39 and 100 per cent below 30, the last on three sentences; among predicted positives, it rises as the score climbs, from 74 per cent in the 70–79 band to 81 per cent at 80–89,

96 per cent at 90–94, and 100 per cent at 95–100. Accuracy against the hand-labelled sample is therefore weakest in the bands nearest the middle of the scale, where an evidence measure should be least certain, and high in the outer cohorts, the predicted negatives scoring below 40 and the predicted positives scoring 90 or above. The score is an ordinal measure of textual evidence rather than a calibrated probability, but its ordering carries the information the design relies on, and the event-dating rule draws only on scores of at least 80, where agreement starts at 81 per cent and rises with the score.

Figure 2: Classifier agreement by evidence-score band



Notes. Share of 500 hand-labelled sentences on which the human consensus label agrees with the classifier, by evidence-score band and predicted label. Each bar reports the agreement share and number of sentences. The score measures textual support and is not a calibrated probability.

The company accounts provide a complementary check that these disclosures mark real commitments of resources. In the J-curve account, the costs of adopting a new general-purpose technology are largely intangible, spending on process redesign, data, and skills that accounting treats as current expense rather than as investment, so adoption should surface in how firms finance themselves rather than in recorded assets (Brynjolfsson et al., 2019, 2021), and that is the pattern the accounts show. In the matched comparison introduced in the next section, reported adoption is associated with an increase of about 1.2 percentage points in debt to assets before 2023Q1, while capital expenditure, R&D, and recorded intangible-asset ratios show no significant change; the from-2023Q1 debt estimate is smaller and less precise. Total assets themselves show no significant differential in either era, so the leverage association operates through the debt side rather than through a shrinking denominator. Recorded investment lines omit internally generated

organisational capital and many implementation costs, so these estimates are consistent with, but do not identify, unrecorded complementary investment. What they document is that adopting firms leave traces in their financing without corresponding traces in recorded assets, which is how intangible adjustment costs, as distinct from intangible assets, would appear in company accounts. Full estimates are reported in Appendix Table 16, with the specification excluding financial firms in Appendix Table 17.

3.3 Outcomes, sample, and diffusion

We merge the AI adoption measure described above with quarterly financial data from S&P Capital IQ. The principal outcomes are year-on-year log growth in employment, revenue, and revenue per employee. The annual difference reduces seasonal reporting effects and gives each coefficient the interpretation of an adopter-comparison growth differential in percentage points per year. Revenue per employee is a revenue-based proxy for labour productivity, not a physical productivity measure. It can change because of quantities, prices, product mix, exchange rates, or workforce composition, although we aim to control for much of this variation through sector-matched comparisons and firm and time fixed effects. We use this outcome to describe revenue generated per worker while retaining that limitation in the interpretation, and refer to it as labour productivity throughout.

Secondary outcomes describe cost structure and recorded investment. Selling, general, and administrative expense (SG&A) over revenue is the overhead ratio, while cost of goods sold (COGS) over revenue is the production-cost ratio. Lower ratios may be consistent with operating efficiencies, but can also reflect pricing, product mix, revenue recognition, and accounting classification. Capital expenditure, research and development (R&D), intangible assets, and debt are each scaled by assets. These balance-sheet ratios capture recorded accounting changes rather than the full set of complementary expenditures that an AI programme may require. Ratio outcomes are expressed in percentage points of their denominator. Appendix D provides the source items and transformations.

The panel contains 236,379 calls for 10,706 listed firms in 88 countries and 250,798 firm-quarters from 2010Q1 through 2025Q4. The international coverage permits comparisons across a wide range of firms, but the estimation population remains listed companies that hold earnings calls and can be linked to quarterly accounts. Calls are fewer than firm-quarters because some firms, mainly outside the US, report semi-annually. When calls are separated by one or two quarters, intervening rows retain the most recent call's classification but contribute no mentions or evidence score. A carried row therefore cannot create an event or raise the early-activity measure. A gap of three quarters ends the firm's panel because longer silence commonly accompanies delisting or acquisition. Appendix G.1 removes all carried rows and reports materially similar headline estimates.

The matched sample contains 1,067 reporting adopters and the same number of controls, drawn from 672 distinct firms. The distinction between controls and distinct firms matters because a suitable comparison firm may be selected for more than one adopter. Section 4 describes the corresponding weights and physical-

firm clustering. Recent disclosure cohorts necessarily have shorter follow-up because the accounting panel ends in 2025Q4; the event-time aggregation uses the cohorts observed at each horizon.

Table 2 documents selection before matching. Covariates are measured before the reporting adopter's event and at the same anchor quarters for eligible comparison firms. Reporting adopters are larger and differ in margins, leverage, liquidity, and growth histories from the pool. These are unconditional differences and do not describe post-disclosure effects. They reinforce the need to compare firms within sector and calendar quarter and to balance their prior trajectories, as in Section 4.

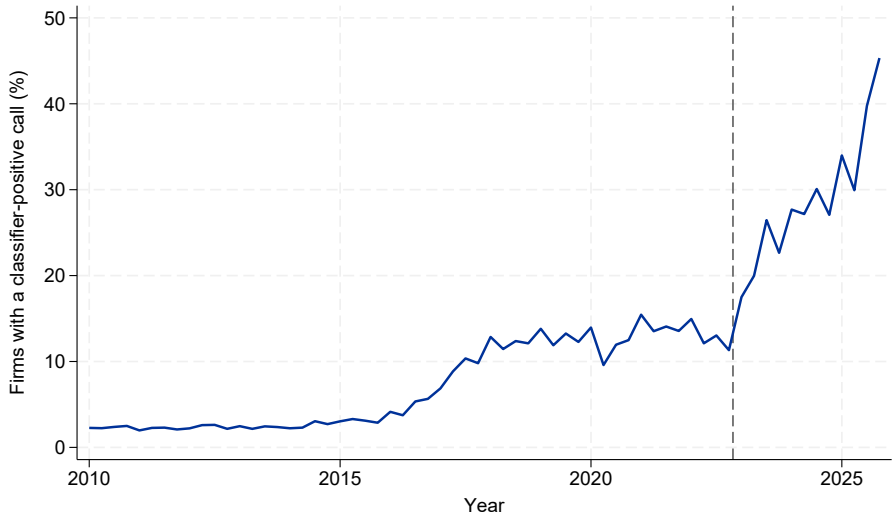
Table 2: Candidate reporting adopters and the eligible comparison pool

Covariate	Candidate adopters	Non-adopter pool
Revenue (USD millions, quarterly)	412	287
Employees	1,780	800
Total assets (USD millions)	2,601	2,005
EBITDA margin	0.04	0.03
Operating margin	-0.03	-0.07
Gross margin	0.43	0.37
Overhead ratio	0.31	0.28
Debt/assets	0.28	0.30
Current ratio	2.24	2.50
Revenue growth, prior 8q (%)	15.66	13.35
Employment growth, prior 8q (%)	8.02	7.38
Employment growth, prior 4q (%)	6.63	6.50
Revenue growth, prior 4q (%)	12.54	12.30
Firms	1,557	3,978

Notes. Means of matching covariates one quarter before first high-scoring disclosure for candidate adopters and at the same anchor quarters for eligible comparison firms. Size rows are geometric means, with quarterly revenue; matching uses logarithms. Growth rates are year-on-year percentages; margins and ratios are fractions. The final row counts firms.

Figure 3 shows the quarterly share of firms with any classifier-positive call, regardless of evidence score. It rises gradually during the late 2010s, roughly doubles from 13 to 26 per cent in the year around ChatGPT's release, and reaches about 45 per cent by 2025Q4. The sharp common increase means that estimates for recent adopters are based on shorter post-event horizons. It also coincides with economy-wide changes around generative AI, making calendar-time adjustment central to the matched comparison.

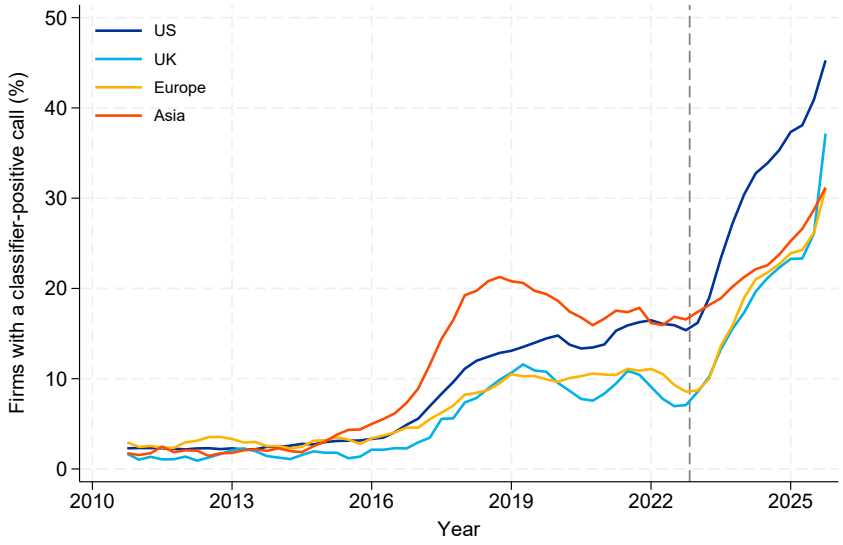
Figure 3: The diffusion of reported AI deployment



Notes. Share of panel firms whose earnings call contains at least one classifier-positive sentence in the quarter, at any evidence score, 2010–2025. The event-date measure elsewhere requires a positive score of at least 80. The dashed line marks ChatGPT's public release in November 2022.

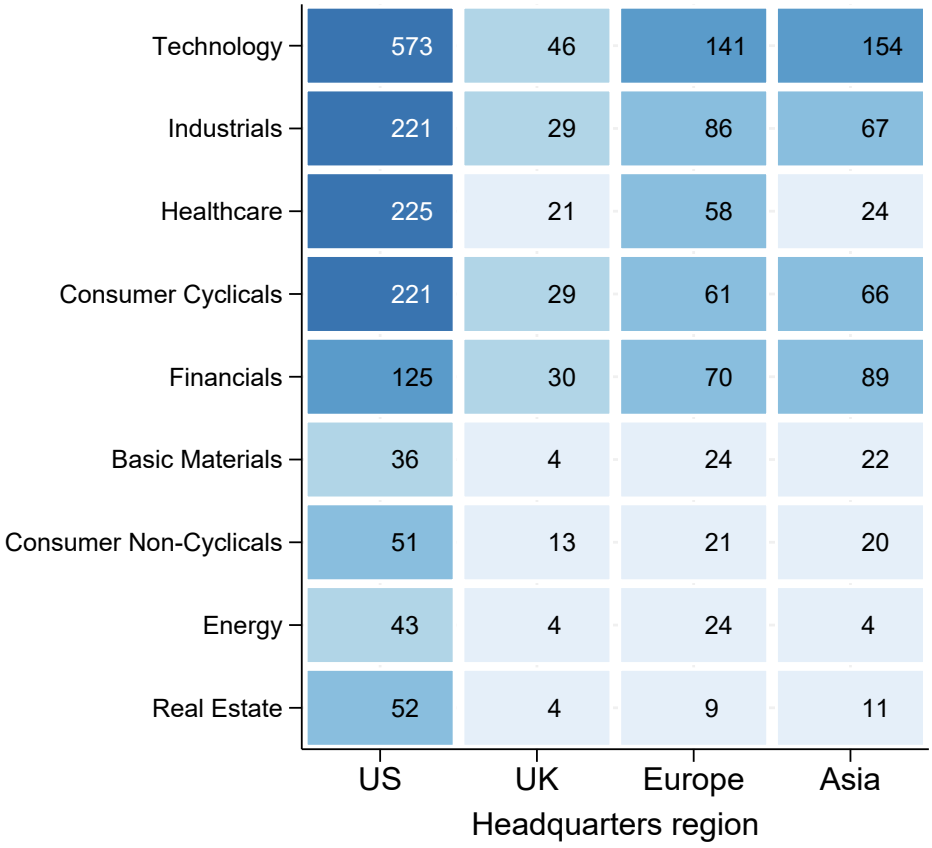
Figures 4 and 5 show that reported AI activity diffuses at different rates across regions and is concentrated in the US and the technology sector. The regional series rises broadly after ChatGPT, while the stock of firms ever making a high-scoring disclosure remains uneven across sector–region cells. Exact matching on sector and anchor quarter addresses part of this compositional variation, but matched firms need not share a country. Physical-firm fixed effects absorb permanent geographic differences and calendar-quarter effects absorb shocks common to the estimation sample. Neither absorbs time-varying shocks specific to an industry or location. As robustness, we therefore further report specifications with sector-by-quarter, region-by-quarter, and country-by-quarter effects.

Figure 4: Reported AI deployment diffuses at different rates across regions



Notes. Share of firms headquartered in each region whose earnings call in the quarter contains at least one classifier-positive sentence, at any evidence score, as in Figure 3. Europe is continental Europe and the European Free Trade Association (EFTA); Asia comprises China, India, Japan, Hong Kong, Taiwan, Singapore, South Korea, and the smaller Asian markets. The dashed line marks the release of ChatGPT.

Figure 5: Reported adoption is concentrated by sector and region



Notes. Number of firms that ever make a high-scoring deployment or committed-investment disclosure, by economic sector and headquarters region. Cells are blank where the sector–region pair contains fewer than ten panel firms in total, adopters or not; the displayed numbers count adopters only, so a shown cell can contain fewer than ten adopters. Sectors with no qualifying cell are omitted. The four regions shown cover 2,678 of the 3,400 reporting adopters; the remainder are headquartered outside these regions or sit in omitted cells. Europe is continental Europe and EFTA; Asia comprises China, India, Japan, Hong Kong, Taiwan, Singapore, South Korea, and the smaller Asian markets.

4 Empirical strategy

4.1 Matched comparison

Firms choose when to deploy AI and when to disclose it, and reporting adopters differ observably from firms without a qualifying disclosure. An unmatched comparison would therefore combine any post-disclosure association with differences in size, performance, and sector composition that pre-date the event. We address observable selection by matching each reporting adopter to a comparison firm observed in the same economic sector and anchor quarter. Matching cannot make disclosure random, but it makes the outcome paths more comparable on measured pre-event characteristics.

Let g_a denote adopter a's first high-scoring disclosure quarter. A candidate comparison must have no recorded high-scoring disclosure at g_a or through $g_a + 12$. For anchors near the end of the sample, eligibility is checked through the available end date, 2025Q4, rather than requiring twelve future quarters to be observed. If a selected comparison later makes its own high-scoring disclosure, its observations are censored from that quarter onwards. The eligibility rule protects the observed post-event comparison window from a measured disclosure.

The match uses 15 pre-event characteristics. Size is represented by eight-quarter averages of log employment, revenue, and assets. Profitability and cost structure comprise gross and operating margins, earnings before interest, tax, depreciation and amortisation (EBITDA) margins, and the overhead ratio, while balance-sheet position includes debt to assets and the current ratio. Employment- and revenue-growth histories include their eight-quarter averages, recent four-quarter averages, and trends. These variables are measured before the adopter's event and at the same anchor quarter for candidate comparisons.

A logit estimated on the combined adopter and candidate rows, with covariates standardised over the scoring sample, gives firm j the propensity score

$$\hat{p}_j = \Lambda\left(z_j'\hat{\beta} + r_j'\hat{\gamma}\right), \tag{1}$$

where z_j contains the standardised covariates and r_j the sector indicators. We transform the fitted probability to log odds, $\hat{\ell}_j = \log[\hat{p}_j/(1-\hat{p}_j)]$, remove sector and anchor-quarter effects, and set the score caliper to one quarter of the residual standard deviation. Within each exact sector–anchor-quarter cell, candidates satisfying $|\hat{\ell}_c - \hat{\ell}_a|$ within this caliper are ranked by standardised covariate distance,

$$c(a) = \arg\min_{c} \sum_{m=1}^{15} \left(z_{m,a} - z_{m,c}\right)^2. \tag{2}$$

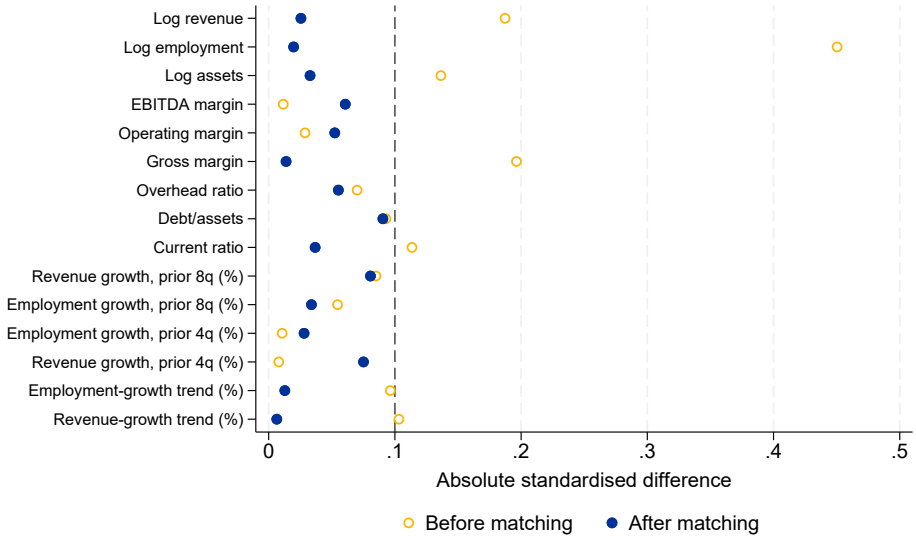
Because employment and revenue growth are the key outcome variables, we additionally require the eight-quarter average, recent four-quarter average, and trend in employment growth to lie within 0.6 standard deviations, and place a caliper on the three equivalent revenue-growth terms. The revenue caliper takes the loosest searched width at which every employment- and revenue-growth history covariate balances within 0.1 standard deviations after matching; Appendix Table 12 reports the search. These hard growth-path calipers discard a distant candidate rather than merely lowering its rank. The nearest qualifying candidate is selected, and adopters without one are dropped rather than forced to a poor match. The procedure is thus nearest-neighbour covariate matching inside propensity-score and growth-path calipers, not matching on the propensity score alone. It produces 1,067 adopter-comparison assignments.

Comparison firms may be reused when they are the nearest eligible match for more than one adopter. The estimation data retain a separate assignment for each use, and matching weights represent these assignments. The same underlying physical-firm identifier is nevertheless used for fixed effects and clustering, so repeated selection does not create independent firms for inference. We also build a five-nearest-neighbour variant

of the match, reported in the robustness section; in that variant each adopter retains unit weight and its comparison weight is divided equally among the qualifying neighbours.

Figure 6 shows the resulting balance. The initial employment-size difference of almost half a standard deviation falls to 0.02, revenue and assets balance within 0.04, and every one of the fifteen matching covariates balances below 0.1, with the largest remaining difference at 0.09 and the largest among the growth histories at 0.08, for the eight-quarter revenue-growth average. The productivity-growth history, which is deliberately not among the matching covariates, balances closely as a by-product. Matching therefore achieves the balance the design asks of it, but it cannot eliminate differences in unobserved adoption readiness, disclosure incentives, or later shocks.

Figure 6: Covariate balance before and after matching



Notes. Absolute standardised differences between reporting adopters and eligible comparison firms before matching, and between weighted adopters and controls after matching. The dashed line marks 0.10. Appendix E reports propensity-score overlap; Appendix Table 11 reports an overlap-free rematch.

4.2 Specifications

The matched data retain separate assignments because one physical firm can serve as a control more than once. Let s index an assignment, $f(s)$ its underlying physical firm, t calendar quarter, A_s an indicator for a reporting-adopter assignment, and g_s the anchor quarter of assignment s. The observed outcome is $Y_{st} = Y_{f(s)t}$, and treatment status is $D_{st} = A_s \mathbf{1}\{t \geq g_s\}$. As a sensitivity benchmark, we also estimate the static specification

$$Y_{st} = \alpha_{f(s)} + \gamma_t + \theta D_{st} + \varepsilon_{st}, \tag{3}$$

where $\alpha_{f(s)}$ and γ_t are physical-firm and calendar-quarter fixed effects. Physical-firm effects remove time-invariant differences between firms, while quarter effects remove shocks common to the matched sample

in a period. Regressions use matching weights and cluster standard errors by physical firm. A control firm selected for several adopters can consequently receive the appropriate matching weight without being treated as several independent firms. This sensitivity regression pools the matched sample rather than imposing pair-by-quarter effects. The coefficient θ is therefore a weighted post-disclosure difference in the matched sample, conditional on firm and common-quarter effects, rather than an average of 1,067 own-pair differences.

Conventional two-way fixed-effects estimates can be difficult to interpret when adoption is staggered and post-disclosure associations differ across cohorts (Goodman-Bacon, 2021; Sun and Abraham, 2021). Our preferred estimator is the interaction-weighted approach of Sun and Abraham (2021). It estimates cohort-by-event-time coefficients relative to the quarter before disclosure, $k = -1$,

$$Y_{st} = \alpha_{f(s)} + \gamma_t + \sum_g \sum_{k \neq -1} \beta_{gk} A_s \mathbf{1}\{g_s = g\} \mathbf{1}\{t - g_s = k\} + \varepsilon_{st}, \tag{4}$$

where g indexes disclosure cohorts. The coefficient β_{gk} compares cohort g at event time k with pooled comparison observations, relative to that cohort's reference quarter. It is not estimated using only cohort g and its own comparison firms. We aggregate these coefficients as

$$\beta_k = \sum_g \omega_{gk} \beta_{gk}, \qquad \omega_{gk} = \frac{N_{gk}}{\sum_h N_{hk}},$$

where N_{gk} is the number of reporting adopters from cohort g observed at event time k. The weights therefore adapt to the cohort support available at each horizon rather than extrapolating recent cohorts beyond the end of the panel. Single post-disclosure averages then weight the supported event-time estimates β_k equally across the stated horizon.

Reporting-adopter observations enter for $k = -6, \ldots, 12$, so distant history does not determine their in-window fixed effect. Control firms enter at their available dates before their own high-scoring disclosure, so their physical-firm effects can use observations outside the event windows of the adopters they are matched to. To assess this asymmetry, the symmetric-window check restricts both sides of each assignment to the same calendar dates spanning $k = -6$ through $k = 12$. The post-event horizon also matches the maximum disclosure-free period imposed when comparisons are selected.

To compare calendar eras, we classify each post-disclosure cohort–event-time cell by whether its calendar quarter falls before or from $c = 2023Q1$, the first full quarter after ChatGPT's public release. For era $e \in \{\text{before}, \text{from}\}$, let $\mathcal{K}_e$ be the supported post-event times and aggregate the coefficients from (4) as

$$\theta_e = \frac{1}{|\mathcal{K}_e|} \sum_{k \in \mathcal{K}_e} \sum_{g:\, g+k \in e} \frac{N_{gk}}{\sum_{h:\, h+k \in e} N_{hk}} \beta_{gk}, \qquad \Delta = \theta_{\text{from}} - \theta_{\text{before}}. \tag{5}$$

Cells are weighted by the number of reporting adopters within each event time and then averaged equally over the supported post-event times in that era. Both era averages and Δ are linear combinations of one Sun–

Abraham coefficient vector, so inference uses their full physical-firm-clustered covariance matrix. Using the first full post-release quarter avoids assigning October and most of November 2022 to the new era. The appendix reports an alternative convention that places the boundary at 2022Q4, the release quarter, for the pooled era contrast, the installed-base cohort, and the entry-fixed activity ranking.

The Sun–Abraham estimator is our preferred staggered-adoption estimator because it accommodates the fixed event window, the matching weights, the calendar-era aggregations, and the joint intensity interactions used below. The appendix reports a pooled-distinct-control implementation of the group-time estimator of Callaway and Sant'Anna (2021). That exercise collapses the stacked data to one observation per physical firm and quarter, assigns treatment cohorts to reporting adopters, and uses comparison firms before their own disclosure. Because it discards repeated assignments and pair weights, applies different cohort-time weights, and covers a broader effective horizon, it is a robustness check rather than an estimate of the identical matched-window effect.

The installed base is the stock of firms whose first high-scoring disclosure precedes 2023Q1, the adopters already in place when generative AI arrived. The installed-base exercise fixes membership to these 683 firms and asks whether their differential against the matched comparisons shifted at the common 2023Q1 boundary. In the notation of equation (3), restricting the sample to assignments whose anchor quarter precedes 2023Q1, we estimate

$$Y_{st} = \alpha_{f(s)} + \gamma_t + \lambda D_{st} + \tau D_{st}\,\mathbf{1}\{t \geq 2023\text{Q1}\} + \varepsilon_{st}, \tag{6}$$

with matching weights and physical-firm clustering as before. The coefficient λ is the ordinary post-adoption differential between these adopters and their comparisons over the quarters before 2023Q1, and τ is the amount by which that differential changes from 2023Q1 onwards, net of the calendar shocks common to both groups. Because every member of the cohort adopted before the boundary, τ switches on at the same date for every firm, so the design is a common-date fixed-cohort difference-in-differences rather than an event-window aggregation. To confirm that this change belongs to the calendar date rather than to time since adoption, the main text re-estimates it with controls for time since disclosure, restricts the cohort to early vintages, and estimates the same construction at placebo boundaries inside the pre-generative-AI period using pre-2023 data only. For the firms first disclosing from 2023Q1, the paper reports the composition of the entering population and two summary estimates, the average post-adoption differential through 2025Q4 and the same average restricted to each firm's first two years after disclosure.

4.3 Identification

A causal interpretation requires conditional parallel trends. Absent reported adoption, adopter outcomes would have followed the pooled matched comparison sample after conditioning on the match and fixed effects. It also requires that disclosure timing is not chosen in response to an outcome shock, that firms do not differentially anticipate the event during the lead window, and that no unobserved shock jointly changes disclosure and the outcome. Exact sector and anchor-quarter matching, pre-event balance, growth-path

calipers, and censoring comparison observations at their own disclosure improve comparability, though they do not make adoption truly random.

Event-study leads are a diagnostic for differential pre-trends rather than proof of the identifying assumption. They are small and jointly insignificant for the principal outcomes, supporting comparable pre-event paths. The robustness section reports a placebo event date and an alternative reference period; the appendix reports symmetric-window estimates, alternative matches, and richer time effects. These checks probe particular threats without eliminating endogenous timing or disclosure concerns. We therefore interpret the estimates as post-disclosure differences in the matched sample under conditional parallel trends, not as the effect of randomly assigned AI use or the date of first operational use. Comparisons across adoption cohorts carry a further caveat, because firms that first report adoption after ChatGPT differ from earlier adopters in observable and unobservable ways, and their pre-disclosure growth already runs above that of their matched comparisons. The summary estimate for these adopters is therefore measured against their own pre-disclosure differential, so that this selection is not absorbed into it.

5 Main results

Unless otherwise stated, regressions use matching weights, physical-firm and calendar-quarter fixed effects, and standard errors clustered by physical firm. Growth outcomes are expressed in percentage points per year. The results are organised around one question. The productivity J-curve says that adoption costs come first and returns follow once complementary investment matures, so the timing of the recovery should depend on how long each firm has been adjusting. In our data the costs do depend on time since adoption, but the recovery does not. It arrived at a single calendar date, the first full quarter after ChatGPT's release, and it arrived for the firms that had already adopted. Firms adopting after that date no longer show the early revenue and productivity costs, over the horizons we can observe so far. Two designs provide the principal evidence, and a further subsection describes the firms that adopt after the release. Staggered event studies trace outcomes by time since adoption, and a common-date difference-in-differences tests whether the installed base's differential shifts upwards at 2023Q1. Each subsection states what its design compares and why before reporting what it finds.

5.1 Average patterns and the shift from 2023Q1

Table 3 opens the results with the paper's pooled estimates around first reported adoption. Panel A asks the simplest question, how adopters fare on average over the whole post-adoption period relative to their matched comparisons. It reports two summaries. The first is a conventional static two-way fixed-effects estimate, retained as a sensitivity check because a single post dummy can weight heterogeneous cohorts perversely. The second is the preferred Sun–Abraham interaction-weighted average over event quarters $k = 0$ to 12, which estimates every adoption-cohort by event-quarter cell separately before averaging, so no

cohort's effect contaminates another's. Panel B asks whether that association changed when generative AI arrived. It uses the same Sun–Abraham cells but classifies each post-adoption cell by whether its calendar quarter falls before or from 2023Q1, the first full quarter after ChatGPT's public release, and tests the difference between the two era averages with the full covariance matrix. Calendar-quarter effects absorb shocks common to reporting adopters and comparison firms, so the era contrast measures the change in adopters' relative outcomes between the two periods.

Table 3: Growth and cost-ratio estimates around first reported AI adoption

Panel A. Average post-adoption estimates

Outcome	Static TWFE (sensitivity)	Sun–Abraham, $k = 0{:}12$
Employment growth	-0.69^{***}	-0.54^{**}
	(0.26)	(0.26)
Revenue growth	-0.66	-0.65
	(0.43)	(0.60)
Labour productivity growth	$+0.04$	-0.10
	(0.40)	(0.57)
Overhead ratio	$+0.60$	$+0.59$
	(0.49)	(0.38)
Production-cost ratio	$+0.60^{*}$	-0.04
	(0.36)	(0.34)

Panel B. Sun–Abraham estimates by calendar era

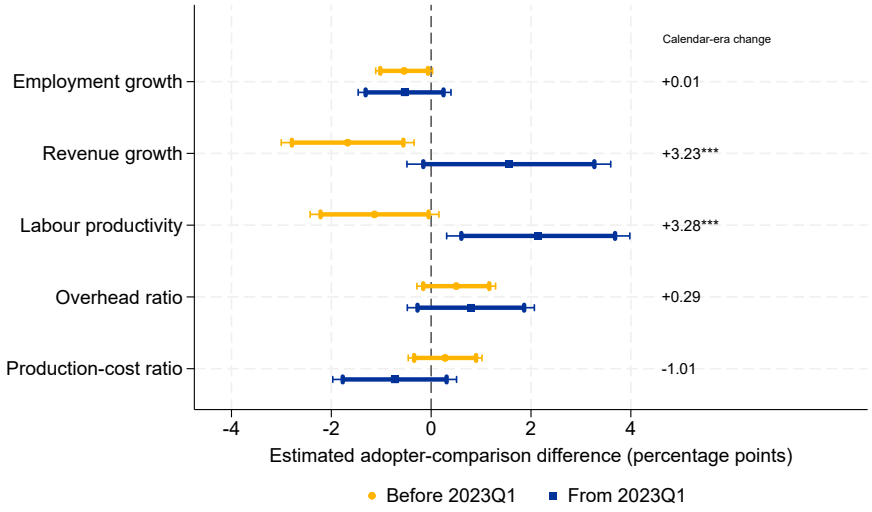
Outcome	Before 2023Q1	From 2023Q1	Change
Employment growth	-0.54^{*}	-0.53	$+0.01$
	(0.29)	(0.47)	(0.53)
Revenue growth	-1.67^{**}	$+1.56$	$+3.23^{***}$
	(0.68)	(1.04)	(1.18)
Labour productivity growth	-1.13^{*}	$+2.14^{**}$	$+3.28^{***}$
	(0.66)	(0.94)	(1.08)
Overhead ratio	$+0.50$	$+0.79$	$+0.29$
	(0.40)	(0.65)	(0.67)
Production-cost ratio	$+0.28$	-0.73	-1.01
	(0.38)	(0.63)	(0.70)
Pairs with post-adoption quarters	683	609	

Notes. Estimates compare firms after their first high-scoring disclosure with the matched comparison sample. Panel A reports a conventional static TWFE sensitivity estimate and the preferred Sun–Abraham average over $k = 0{:}12$. Panel B reports Sun–Abraham estimates before and from 2023Q1 and their from-2023Q1-minus-before-2023Q1 calendar-era change, with each cohort–event-time cell measured against that cohort's final pre-event quarter. Every estimate uses the same 1,067 matched pairs; observations vary with outcome availability. The final row of Panel B counts the pairs whose adopter contributes at least one post-adoption quarter, $k = 0$ to 12, to that calendar era, with counts for individual outcomes slightly lower where an outcome is unavailable. *, **, and *** denote significance at the 10%, 5%, and 1% levels.

Employment grows more slowly after reported adoption, by 0.54 percentage points per year under Sun–Abraham and 0.69 under the static specification, and the gap is almost identical before and from 2023Q1, so it shows no response to the technology's generative AI turn. Revenue and productivity behave differently,

and their all-period averages are close to zero precisely because they conceal the era change. Revenue growth runs 1.67 points behind the comparisons before 2023Q1 and 1.56 ahead, imprecisely, from 2023Q1, a change of +3.23 points ($p = 0.006$). Labour productivity runs 1.13 points behind the comparisons before 2023Q1 and 2.14 points ahead from 2023Q1 ($p = 0.023$ on the post-era estimate), a change of +3.28 points ($p = 0.002$). The pooled group-time estimator of Callaway and Sant'Anna, which uses different comparison pooling and cohort-time weights, gives era changes of +1.84 for revenue and +2.33 for productivity, both significant, and an employment change of −0.50 that is not (Appendix Table 13). The two accounting cost ratios offer little evidence of an operating-efficiency change in either era; the production-cost ratio moves by −1.01 points from 2023Q1 but the change is not significant, and pricing, product mix, and accounting classification can move a cost share in any case. Figure 7 puts the era estimates on a common scale.

Figure 7: Matched-sample estimates before and from 2023Q1



Notes. Sun–Abraham interaction-weighted estimates before 2023Q1 and from 2023Q1. Thick and thin bars are 90% and 95% confidence intervals. Growth outcomes are in percentage points per year and cost ratios are in percentage points of revenue. The right-hand column reports the from-2023Q1-minus-before-2023Q1 calendar-era change.

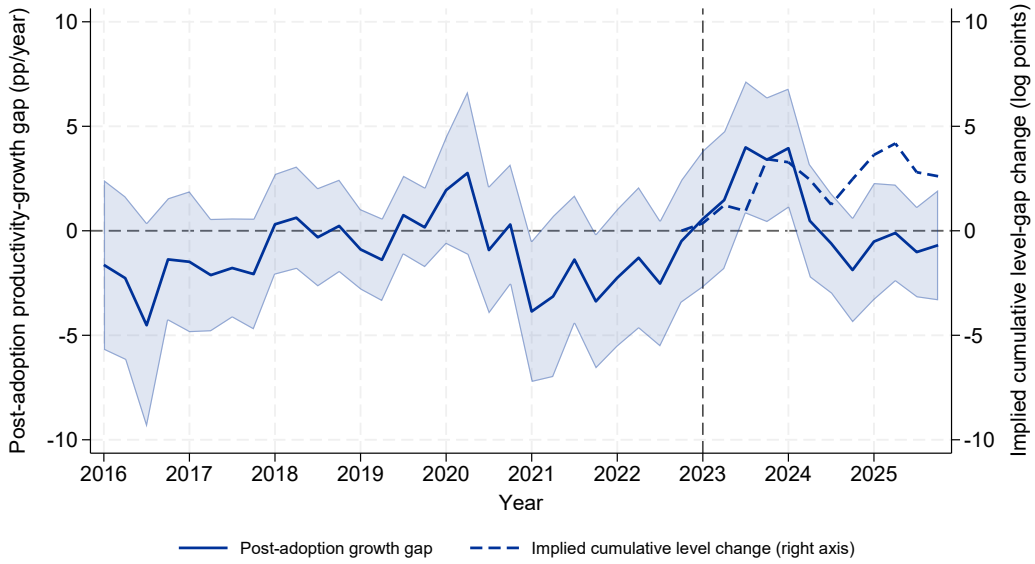
The pooled comparison bundles two sources of variation. The technology environment changes after ChatGPT, and the set of adoption cohorts contributing to each calendar era changes with it, because firms keep adopting. The rest of this section separates the two. The next subsection fixes the cohort, the 683 firms that had reported adoption before 2023Q1, and asks whether their path changed at that date. A later subsection turns to the firms first adopting after the release, describing who they are and what their early record shows.

5.2 The installed base entering 2023

Figure 8 shows the paper's central fact. The line is the productivity-growth differential between the installed-base adopters and their matched comparisons, in calendar time, computed only over each adopter's post-

disclosure quarters. Because every post-disclosure quarter carries its own coefficient and no quarter is omitted, each point is measured against the firm's own pre-disclosure average, net of calendar effects. The event-time estimates earlier normalise on the quarter before disclosure because event time aligns that quarter across firms; in calendar time the cohort's disclosure quarters differ, so the firm's own pre-disclosure average is the natural reference, and it is the same reference the common-date regression below uses. From 2016 to the end of 2022 the differential is negative in most quarters, averaging -1.1 points per year, and shows no sign of closing as firms age past their adoption dates. From 2023Q1 it turns positive, averaging $+2.3$ points through mid-2024. The surge recedes from mid-2024 and the differential is mildly negative over the final six quarters. A one-off gain in the level of productivity appears in year-on-year growth as a temporary hump followed by a return towards zero, so this shape is consistent with a completed level adjustment rather than a permanently faster growth rate, as shown by the cumulative change in the dashed line. It cumulates the same information into the implied change in the productivity-level gap, and by 2025Q4 it stands at 2.6 log points, so the level gained through the surge is retained rather than reversed.

Figure 8: The installed base's post-adoption productivity-growth differential in calendar time



Notes. The solid line is the quarterly adopter–comparison gap in year-on-year labour-productivity growth for the 683 matched pairs whose adopter first produced a high-scoring disclosure before 2023Q1, estimated over post-disclosure quarters only. Each treated post-disclosure quarter has its own coefficient and none is omitted, so every point is measured against the adopter's own pre-disclosure mean, net of calendar-quarter effects; the shaded region is a pointwise 95% confidence band. The dashed line, read against the right axis, is the implied cumulative change in the productivity-level gap, the sum from 2023Q1 of directly estimated quarter-on-quarter growth-gap coefficients net of their equal-quarter 2018Q1–2022Q4 mean, anchored at zero in 2022Q4 and shown as a point path. The two axes share one numerical range so their zero lines coincide. The vertical line marks 2023Q1.

Table 4 estimates τ of equation (6), the change in the installed base's differential from 2023Q1. The design fits the situation of the installed base, whose members are all already treated, many of them years

past their adoption dates, so there is no staggered treatment to aggregate; the only question is whether their differential shifted at one shared calendar date. The regression contains two indicators alongside the firm and calendar-quarter fixed effects. The first switches on for an adopter from its own disclosure quarter onwards, and its coefficient λ is the ordinary post-adoption gap between adopters and their matched comparisons over the quarters before 2023Q1. The second switches on for an adopter only from 2023Q1 onwards, and its coefficient τ measures how much the gap between the installed base and its matched comparisons changed at 2023Q1, over and above the gap those same firms had shown since adopting, with calendar-quarter effects removing shocks common to both groups. A positive estimate means the installed base improved relative to its comparisons at the boundary, whatever level its differential had held before. The equation looks like the static specification of Table 3, but the concern with that specification does not apply here. A single post-adoption indicator is problematic when adoption dates are staggered, because it averages cohorts at different event times with uneven weights; the 2023Q1 indicator instead switches on in the same quarter for every firm in a fixed cohort, so there is nothing to weight unevenly. The first two rows of Panel A report, for each outcome and sample, the change from 2023Q1 in that outcome's adopter–comparison differential relative to its value over the same firms' earlier post-disclosure quarters; the final two rows, taken up later in the section, report the generative AI era adopters' post-adoption averages rather than a change at the boundary. Productivity rises by 1.99 points per year (standard error 0.93, $p = 0.033$). The split of that change between its two margins is less precise. The revenue change is $+1.18$ and not significant, and the employment change is -0.84, marginally short of significance at 5 per cent ($p = 0.056$), so the productivity change appears to draw on some of each margin, though neither is significant on its own. The second sample row excludes US and Chinese technology firms, where reported adoption is most concentrated, and leaves the productivity change at 1.84 points, significant at the 10 per cent level.

Table 4: The installed base's change from 2023Q1, the generative AI era adopters' averages, and treatments of the specification

Panel A. The change in the adopter–comparison differential from 2023Q1

Sample	Employment	Revenue	Labour productivity
All installed-base adopters	-0.84^{*}	$+1.18$	$+1.99^{**}$
	(0.44)	(1.02)	(0.93)
Installed base, excluding US and CN tech firms	-0.86^{*}	$+0.97$	$+1.84^{*}$
	(0.46)	(1.10)	(1.00)
GenAI era adopters	-0.70	-0.16	$+0.58$
	(0.45)	(1.25)	(1.25)
GenAI era adopters, first 2 yrs after adoption	-0.73^{*}	-0.25	$+0.52$
	(0.43)	(1.18)	(1.19)

Panel B. The installed base's productivity-growth change from 2023Q1, under each treatment of the specification

Specification	Change	(SE)	Adopters
Baseline	$+1.99^{**}$	(0.93)	683
With event-time profile controls	$+3.03^{***}$	(1.13)	683
Excluding the 2020 and 2021 quarters	$+1.85^{*}$	(0.99)	683
Estimated on 2018 onwards only	$+2.34^{**}$	(0.98)	670
Adopted 2020Q4 or earlier only	$+2.13^{**}$	(1.04)	535
Adopters still observed from 2023Q1	$+2.00^{**}$	(0.97)	559
Two-way clustered (firm and quarter)	$+1.99^{*}$	(1.05)	683

Notes. Panel A reports τ from equation (6) for each outcome, the change from 2023Q1 in the adopter–comparison differential relative to its value over the same firms' earlier post-disclosure quarters, for the 683 matched pairs whose adopter first disclosed before 2023Q1. The regression includes a post-disclosure indicator and firm and calendar-quarter fixed effects, so each firm is measured against its own history and shocks common to adopters and comparisons are removed. The second row of Panel A re-estimates the same specification excluding pairs whose adopter is a US or Chinese technology firm. Panel B re-estimates the change in productivity growth from 2023Q1 under treatments of the specification. Event-time profile controls add indicators for bands of quarters since disclosure, so the change is identified from the calendar date holding the adoption life-cycle fixed. The excluding-2020-and-2021 row drops those calendar quarters from the estimation sample. The 2018-onwards row measures the change against the recent pre-boundary differential only. The early-vintage row restricts the cohort to adopters at least nine quarters past disclosure at the boundary. The still-observed row keeps the 559 adopters with productivity observations from 2023Q1, so exits before the boundary cannot move the reference. The final two rows of Panel A are not changes at the boundary but average post-adoption differentials of the 384 matched adopters first disclosing from 2023Q1, with event quarters weighted by the number of firms observed and measured against that group's average pre-disclosure differential; the first averages all observed post-adoption quarters through 2025Q4 and the second restricts to each firm's first two years after disclosure. Standard errors cluster by physical firm except in the two-way row. *, **, and *** denote significance at the 10%, 5%, and 1% levels.

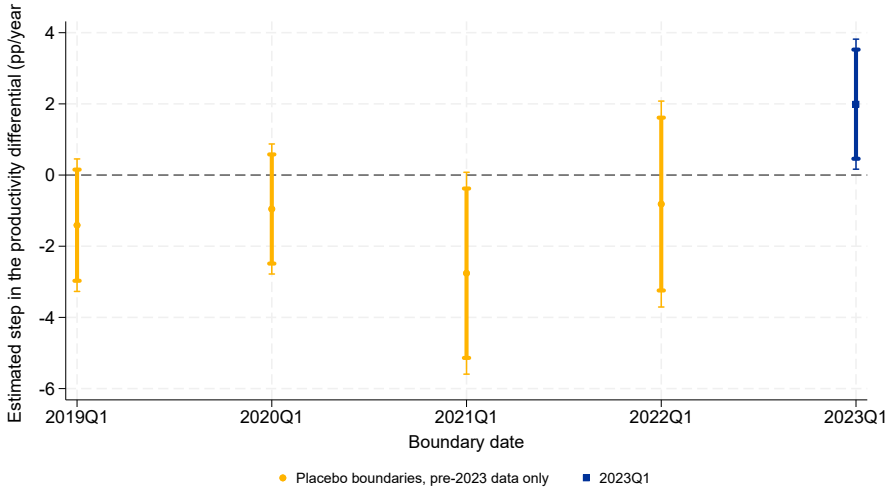
Panel B re-estimates the change in productivity growth from 2023Q1 under treatments of the specification, and each row answers one objection. The first objection is that the installed base is mechanically deeper

into its adoption life-cycle by 2023, so an ordinary J-curve recovery arriving with time since adoption would produce a rise at 2023Q1 even if the date itself meant nothing. The second row adds controls for bands of time since disclosure, so the change is identified from the calendar date holding the adoption life-cycle profile fixed, and it rises to 3.03 points ($p = 0.007$) rather than falling. The estimated band controls also explain why the estimate rises rather than falls. Within the pre-generative-AI period the differential worsens with time since adoption, from about -0.5 points in the first year to about -2.5 points beyond five years, so time since adoption alone produces no recovery that a calendar break could pick up. Part of that deepening is the pandemic rather than the ageing of adoptions. Splitting the eight-to-eleven-quarter band by vintage puts it at -4.1 for the 2018 and 2019 adopters, whose late window falls in 2020 and 2021, against -3.0 for the earlier vintages and -1.0 for the 2020 to 2022 vintages, and dropping the 2020 and 2021 calendar quarters reduces the band to -1.8, so the pandemic explains part, though not all, of the late deepening. The third row therefore drops the 2020 and 2021 calendar quarters from the estimation entirely, and the change is 1.85 points. The next objection is that a long panel lets old history into the reference, since a firm adopting in 2016 contributes pre-boundary quarters from up to seven years before the release. The fourth row measures the change against the differential from 2018 onwards only and gives 2.34 points ($p = 0.017$). The fifth row addresses vintage composition by restricting the cohort to firms that had disclosed by 2020Q4, at least nine quarters before the boundary, and gives 2.13 points ($p = 0.042$). The sixth row asks whether firms leaving the panel drive the result. A firm belongs to the installed base if it disclosed before 2023Q1, whether or not it is still observed afterwards; 559 of the 683 adopters report productivity from 2023Q1, while the rest exit the panel before the boundary and so enter only the pre-boundary average. If those leavers were unusual firms, the pre-boundary reference could be tilted by their exit. Restricting the estimate to the adopters observed on both sides of the boundary gives 2.00 points, essentially the baseline, so exits are not driving the change. Comparison firms can also leave the panel, since a control's observations are censored once it makes its own first qualifying disclosure, and these exits accelerate after the release; 262 of the 683 installed-base controls eventually disclose, 188 of them from 2023Q1 onwards, so the comparison group thins over the generative AI era. Two facts limit what this turnover can explain. The eligibility rule means no control is censored within three years of its anchor, and the median censoring arrives 19 quarters after it, so through the end of 2023, by which point the rise in the differential is already visible in Figure 8, 79 per cent of the original comparison group is still in place. And the controls that later disclose were not a faster-improving group beforehand; their productivity growth over 2021 and 2022 is within half a percentage point of that of the controls that never disclose, so removing them does not weaken the comparison group. The final row keeps the baseline estimate and recomputes the standard errors clustering by firm and calendar quarter jointly, which gives a p-value of 0.062. Taken together, the change lies between 1.85 and 3.03 points across the rows, is centred near 2, and is significant at the 5 per cent level in five of the seven rows.

Figure 9 asks the remaining question, whether 2023Q1 is special or whether this construction finds a rise at any boundary. It re-estimates the same specification at boundaries placed at the start of 2019, 2020, 2021, and 2022, using only data before 2023 and only pairs whose adopter disclosed before the placebo boundary,

so the true turn cannot enter. Every placebo estimate is negative. The placebos sit below zero rather than at it because the deteriorating time-since-adoption profile shows through at any interior boundary, which is also why controlling for that profile raises the 2023Q1 estimate. The 2023Q1 boundary is the only date in the sample at which the installed base's differential shifts upwards.

Figure 9: The change in the installed base's differential appears only at the generative AI boundary



Notes. Each point is the estimated change in the installed base's productivity-growth differential at the stated boundary, from the same specification as Panel A of Table 4. The placebo estimates use only data before 2023Q1 and only pairs whose adopter disclosed before the placebo boundary, so the true turn cannot enter them. Thick and thin bars are 90% and 95% confidence intervals.

The economic reading of these estimates is straightforward. These firms adopted, paid the costs, and did not see returns while the technology remained in its pre-generative-AI state, however long they waited. When substantially more capable models arrived, the same firms improved together across the adoption vintages we can separate. Our interpretation of these findings is that the gain arrived first at firms that had already reorganised their data, processes, and skills around AI, and that none of this reorganisation need have been undertaken in anticipation of generative AI; the complements that earlier AI systems required, organised data, redesigned workflows, and technical and managerial capability, overlap heavily with those generative AI draws on, so groundwork laid for one form of the technology would have positioned these firms to benefit from the next. That pattern is what one expects if the binding constraint on the payoff was the capability of the technology rather than the completion of each firm's adjustment. The design cannot rule out another force arriving at the same calendar date and favouring the same firms, but the placebo boundaries, the life-cycle controls, and the vintage restriction remove the mechanical explanations, no boundary tested in the preceding four years produces the pattern, and the calendar-time differential in Figure 8 shows no earlier upturn.

5.3 Adopting after the release of generative AI

Table 4 showed that the installed base's gains arrived when the technology changed. The release changed something else as well, namely who adopts. Before the release, deploying AI typically meant building bespoke systems around a firm's own data and engineering, an investment that took scale (Babina et al., 2024). Generative AI tools arrive packaged, work through natural language, and can be tried without rebuilding core systems, so the barrier to entry should fall and the firms entering from 2023Q1 should look different from those that came before. That is the diffusion path of a general-purpose technology, moving from technical early users to the mainstream as computing and the internet did before it (David, 1990; Bresnahan and Trajtenberg, 1995).

Entry roughly doubles at the release. Among active firms that had not yet reported adoption, the share making a first qualifying disclosure averaged 7.4 per cent a year over 2016 to 2022; it rises to 13.4 per cent in 2023, holds at 13.7 per cent in 2024, and reaches 17.8 per cent in 2025. More than two fifths of all the first disclosures we ever observe occur in these twelve quarters. Table 5 profiles the entrants. The entrants' median employment is 741 against 1,558 for the pre-generative-AI adopters, and their median revenue little more than half as large. Adoption also broadens across sectors and regions. The lower rows of the table report, for each group, the share of its firms that had disclosed adoption by the end of each column's period, so the second column is cumulative and the difference between the columns is what the generative AI era added. Before the release, thirteen years of diffusion had reached 35 per cent of Technology firms and 43 per cent of US technology firms, against 11 per cent of firms outside the four largest sectors. In the three years from 2023Q1 every group then adds at least half of what it had accumulated over those thirteen, taking Technology to 59 per cent, Financials to 39, and the sectors outside the four largest to 22. The proportional gains are generally largest where adoption had been rare and smallest among the pre-release leaders. The United Kingdom's share more than doubles, from 14 to 31 per cent, while US technology, still ahead of every other group in levels, adds just over half of its pre-release share. The mode of reported use also shifts, with mainly-product adopters falling from 61 to 42 per cent of entrants across the two periods and mainly-internal adopters rising from 25 to 38 per cent; the remainder are evenly split or unclear. What has not changed is the entrants' standing within their industries, a mean productivity percentile of 47 against 44. Reported adoption has spread from large, disproportionately American, technology-centred early movers to ordinary firms across sectors and countries, consistent with falling barriers to entry.

Table 5: Who adopts, before and after the release of generative AI

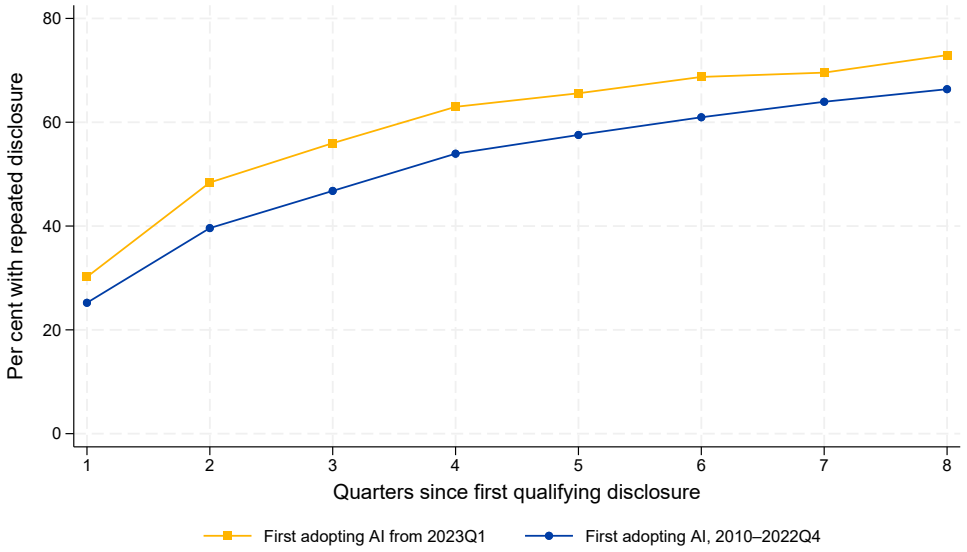
	First adopting AI, 2010–2022Q4	First adopting AI from 2023Q1
Adopting firms	1,939	1,410
Mean annual first-disclosure rate among active firms (%)	7.4	15.0
Median employment	1,558	741
Median revenue (USD millions, annualised)	1,143	657
Mean productivity percentile within sector–quarter	44	47
Mainly internal use (%)	25.0	37.7
Mainly product use (%)	61.3	42.3
Share of the group's firms having disclosed adoption (%)		
	By end-2022Q4	By end-2025Q4
Technology	35.2	59.3
Industrials	16.8	31.7
Consumer Cyclicals	18.7	32.7
Financials	24.3	39.2
Other sectors	11.2	21.8
US	21.7	35.2
United Kingdom	14.1	31.1
Europe	17.3	32.0
Asia	15.8	30.5
Other regions	13.5	22.7
US technology	42.6	66.5

Notes. All firms in the panel with a qualifying disclosure, split by the date of the first. The second row is the share of firms with calls and no prior qualifying disclosure that produce their first one in a year, averaged over 2016 to 2022 and over 2023 to 2025; it counts all entrants, while the other rows exclude firms first disclosing in 2022Q4, the quarter of ChatGPT's release. Characteristics are measured over the four quarters up to and including that disclosure. Medians are taken across firms; revenue is an annualised US-dollar value measured over each firm's disclosure window, so the two columns reflect different calendar periods. The productivity percentile ranks revenue per employee within the firm's economic sector and calendar quarter across all panel firms, adopters and non-adopters alike. The lower rows report, for each group, the share of its panel firms whose first qualifying disclosure has occurred by the end of the column's period, so the second column is cumulative, contains the first, and continues to exclude the 2022Q4 cohort. Sectors follow the TRBC economic sector classification, with other sectors grouping the seven remaining sectors; regions group headquarters countries as in Figure 4, with remaining countries as other regions. Firms without a recorded sector appear only in the region rows. The internal- and product-use rows classify each adopter's classifier-positive quarters within the column's period with the use-type classifier of Appendix C; mainly internal or mainly product means strictly more of the firm's quarters carry that label than the other. Shares are in per cent.

The entrants of the generative AI era also follow through on their first disclosures. Because the classifier

requires a confident statement of current adoption, a qualifying disclosure is already a substantive claim; what Figure 10 adds is that the claim persists. For each horizon, the figure shows the share of entering firms whose first qualifying disclosure is followed by at least one further high-scoring disclosure within that many quarters, computed among firms observable for that long. The generative AI era entrants sit above the earlier entrants at every horizon, with 63 per cent reporting again within four quarters against 54 per cent, and 73 against 66 within eight. Firms that begin reporting adoption after the release are, if anything, more likely to continue reporting it than the early adopters were, so the entry wave records sustained engagement rather than a transient response to the release of ChatGPT.

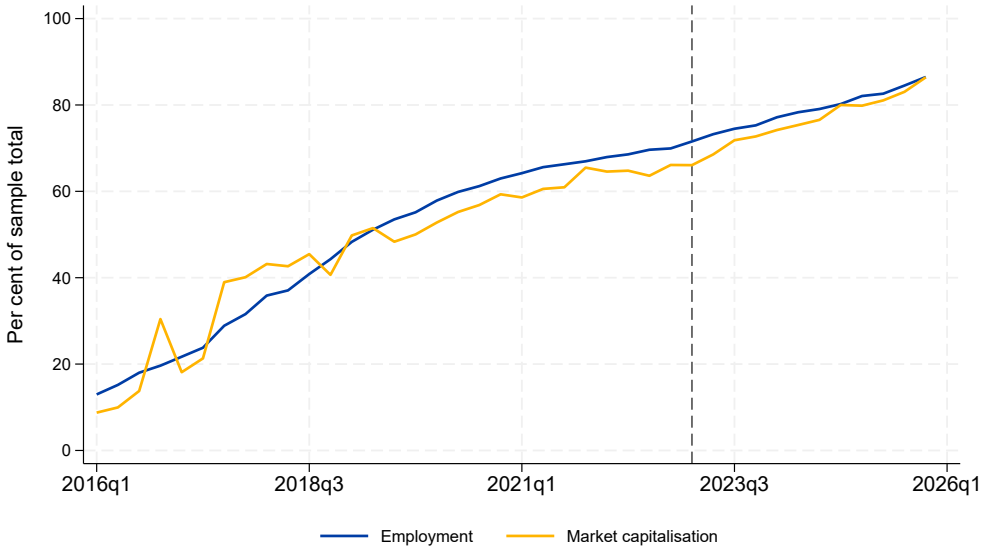
Figure 10: The share of first disclosures followed by continued reporting



Notes. For each horizon on the horizontal axis, the share of entering firms whose first qualifying disclosure is followed by at least one further high-scoring disclosure within that many quarters, computed among firms observable for that many quarters. Firms first disclosing in 2022Q4, the quarter of ChatGPT's release, are excluded, as in Table 5.

The cumulative effect of this entry is that reported adoption now spans most of the listed sector. Figure 11 tracks the share of sample employment and market capitalisation held by firms whose first qualifying disclosure has occurred. Firms that had disclosed adoption accounted for just over half of sample employment, and half of its market capitalisation, at the end of 2019, and for 70 per cent of employment and two thirds of market capitalisation when ChatGPT was released; by the end of 2025 they account for 86 per cent of each. The productivity patterns in this paper therefore concern the bulk of the listed sector by economic weight rather than a niche of early movers.

Figure 11: The share of the listed sector accounted for by reporting adopters



Notes. Share of total sample employment and of total market capitalisation held by firms whose first qualifying disclosure has occurred by each quarter, across all panel firms. The dashed vertical line marks 2023Q1.

Whether the entrants gain is a question their short history can only begin to answer, but the early record is already informative. For the 384 of them in the matched estimation sample, the last two rows of Panel A in Table 4 summarise it. Averaged over all observed post-adoption quarters through 2025Q4, the entrants' differentials are statistically indistinguishable from zero on every outcome. Restricting the average to each firm's first two years after adoption, the build-phase window in which the earlier cohorts' costs first appeared, changes little; revenue and productivity remain flat, and employment growth is slightly slower, significant only at the 10 per cent level. The productivity estimate stays insignificant under every aggregation convention we examine (Appendix Table 18), and how much a firm discusses AI at adoption shows no robust relationship with its subsequent productivity either (Appendix Table 19).

The two flat estimates say different things, and together they form the paper's central contrast. The absence of early costs is a clear change from the pre-generative-AI decade. Firms first disclosing adoption in the pre-generative-AI decade entered a build phase in which revenue growth ran about 1.7 points behind their comparisons (Table 3); the entrants of the generative AI era show no such shortfall even in their first two years. But the absence of a significant early gain is equally striking, because gains are precisely what the installed base received, almost immediately, when the technology improved. Panel B shows the installed base's productivity growth rising by about two points at 2023Q1 and holding under every treatment of the specification; the entrants adopting that same technology from 2023Q1 onwards show no gain that is significant at conventional levels over horizons of similar length.

Both findings are consistent with the entrants adopting an essentially finished technology. Where the earlier cohorts built bespoke systems and reorganised around them, paying visibly for years, a firm adopting

from 2023Q1 can take a packaged, capable model off the shelf, which is why the costly phase no longer appears. Yet the installed base's experience suggests that the gains came from combining the newly capable technology with years of accumulated complementary investment, and an off-the-shelf adoption involves no such accumulation. The flat productivity path is then what deployment without deep internal reorganisation would produce. The environment the entrants face points the same way. By 2023 generative AI tools are available to every firm, and employees at comparison firms plausibly use them without any announcement, so a qualifying disclosure marks formal adoption, the deliberate integration of the technology into a firm's operations rather than ad hoc use by employees, and the estimate in Table 4 is the return to that formal integration relative to firms that may already use the tools informally. A modest differential is what a technology available to everyone should produce, but the contrast with the installed base shows that formal adoption alone, without the accumulated complementary base, has so far conferred no measured advantage.

If the entrants' returns require the complementary investment the installed base undertook, they lie beyond the current panel, and whether they arrive is the clearest test ahead. If complementary investment is what turns capability into measured gains, the entrants should climb as their own investments mature; a flat path beside a persistent installed-base advantage would instead mean the advantage lay in adopting early, before the technology reached everyone.

6 Robustness and interpretation

6.1 Robustness

The event-study leads provide no evidence of differential employment growth immediately before reported adoption. The five leads from $k = -6$ to -2 average $+0.15$ percentage points and are jointly insignificant ($p = 0.29$). A false event date eight quarters before the first high-scoring disclosure also produces no significant employment or labour-productivity difference when the sample is restricted to observations before the true event. Using the mean of the pre-adoption leads rather than the final pre-event quarter leaves the employment estimate negative. These results reduce concern that a short pre-event divergence mechanically generates the baseline estimates.

The principal productivity results are stable when the match itself is constructed differently. Three alternative matches address three distinct concerns. The overlap-free re-match addresses a mechanical overlap between the matching variables and the earliest outcomes. Outcomes are year-on-year growth rates, so the outcome in the first post-adoption quarters is computed partly from levels in the year before adoption, and those same pre-adoption quarters feed the growth histories the baseline match balances. Matching firms on growth over quarters that also enter the early outcome calculations could therefore create some similarity in the early estimates by construction. The re-match instead measures every matching variable over a window that ends before the outcome's base year begins, so no quarter used for matching enters any post-event outcome. The five-nearest-neighbour design gives each adopter five comparison firms rather than one, so

that no single comparison firm drives an estimate. The pre-path-adjusted design adds direct controls for each firm's pre-adoption growth path, absorbing any remaining differences in trajectories that the match left behind. Across these designs the installed-base estimate ranges from 1.65 to 2.42 percentage points, and the pooled productivity reversal, the from-2023Q1-minus-before-2023Q1 change in the productivity differential from Panel B of Table 3, remains positive and significant in every design (Appendix Table 11).

Alternative dating and windowing conventions tell the same story. Requiring each comparison firm to be observed over the same event window as its treated firm, so that the two are compared over identical quarters, gives a productivity change of $+2.62$ points ($p = 0.026$). Moving the era boundary from 2023Q1 to 2022Q4, the quarter of ChatGPT's release, gives a pooled change of $+3.26$ ($p = 0.003$) against the $+3.28$ baseline, and an installed-base change of $+2.10$. Rebuilding the installed-base estimate from the Sun–Abraham cells rather than the common-date regression, feasible for the 225 of 683 adopters with post-boundary support inside the event window, gives $+3.76$ ($p = 0.032$). Finally, replacing the common calendar-quarter effects with sector-by-quarter, region-by-quarter, or country-by-quarter effects leaves the installed-base estimate between 1.92 and 2.33, so a common sectoral cycle or geographic shock is an unlikely sole explanation (Appendix Table 14). The remaining installed-base checks, covering event-time profile controls, the pandemic quarters, placebo boundaries, and survivor restrictions, appear in the main text in Table 4 and Figure 9.

The treatment definition also groups different kinds of adoption, since a qualifying disclosure can describe AI used inside the firm's own operations or AI built into what the firm sells. To check that the installed-base change does not depend on this mix, a second classifier labels every classifier-positive sentence as internal use, product use, both, or unclear; its prompt, its validation against the hand-labelled sample, and the composition of disclosures over time are reported in Appendix C. On the decisive internal-versus-product calls the check relies on, the classifier agrees with the hand labels in 95.7 per cent of sentences. Grouping the installed-base adopters by these labels shows that the change survives with a stable or larger coefficient when adoption is internal use. Adopters whose pre-2023 disclosures are predominantly internal show a change of $+2.82$ points ($p = 0.109$, 187 pairs), those reporting only internal use in the disclosure quarter $+1.78$ ($p = 0.323$, 183 pairs), and the product-dominant remainder $+1.72$ ($p = 0.121$, 410 pairs). The split cuts the sample sharply, roughly doubling the standard errors, so significance is hard to reach at these sizes; both estimates sit just outside the 10 per cent level, and the split is underpowered for a formal internal-versus-product contrast (Appendix Table 9).

Inference accounts for the reuse of comparison firms. A comparison firm can serve in several matched pairs, so standard errors are clustered on the physical firm, which stops the same firm's quarters from being treated as independent pieces of information. Because five era contrasts are tested at once, one could appear significant by chance; a Holm (1979) correction across the five leaves the productivity and revenue changes significant at the 5 per cent level. Finally, we test investment intensity, measured by the frequency of qualifying sentences per earnings call, but see no robust link between investment intensity measured in this way and firm outcomes (Appendix Table 19).

6.2 Limitations and interpretation

The event we study is a firm's first reported AI deployment or committed investment, which is not necessarily its first operational use, and firms choose when to discuss AI. Larger firms and firms with stronger recent growth are more likely to produce a first qualifying disclosure, and the design controls for this observable selection in three ways. The match conditions on size, profitability, leverage, and two years of pre-event growth; firm fixed effects absorb persistent differences in levels; and the overlap-free re-match confirms the results when the quarters used for matching are fully separated from the quarters used for outcomes. The flat pre-event leads and the placebo event date indicate that adopters and their comparisons were on similar paths before disclosure. Selection on characteristics that do not appear in the pre-event data remains possible, and transcript language cannot verify that a reported deployment was implemented, so we interpret the estimates as matched-sample associations under conditional parallel trends rather than as effects of randomly assigned adoption.

The classifier reduces this measurement problem without eliminating it. It assigns scores below the treatment threshold to general commentary, intentions without a commitment, and weakly aspirational statements, and the sentence-level validation shows that high evidence scores correspond closely to current deployment or committed investment, which supports consistent measurement across the full panel. The classifier cannot detect deployment that firms do not report, cannot determine whether a disclosed project succeeded, and cannot prevent firms from emphasising their AI activity strategically, so some misdated or selectively timed reporting can remain.

The outcome we call labour productivity is US-dollar revenue per employee. It can change through quantities, prices, product mix, exchange rates, or workforce composition. Firm fixed effects absorb each firm's average growth, and the sectoral and geographic time effects absorb shocks common to those groups, but neither absorbs a change over time in an individual firm's own prices, product mix, or currency exposure, so the estimates are not direct measures of physical productivity. All estimates are also differences between adopters and matched comparison firms. Consequences of ChatGPT that are common to both groups are removed, and slower relative employment growth at adopters does not imply that aggregate employment fell; the design cannot show whether workers who leave one group of firms are absorbed elsewhere.

The accounting evidence concerns four recorded ratios, debt to assets, capital expenditure, R&D, and recorded intangible assets. As noted in Section 3, the rise in debt to assets reflects an increase in debt rather than a change in total assets, but the estimates do not establish that the borrowing financed AI specifically. The absence of significant change in the three investment ratios is consistent with adjustment costs that are expensed rather than capitalised, but it does not rule out other explanations.

Finally, the sample contains listed firms with earnings calls, which are larger, more visible, and more concentrated in the US than the wider firm population, and firms that use AI without discussing it are not observed. The from-2023Q1 window is short, especially for the new reporting adopters and at longer event horizons, where support declines and confidence intervals widen. A longer panel will show whether the

installed-base improvement persists, whether the newer adopters improve as their complementary investments mature, and whether the employment-growth gap closes.

7 Conclusion

This paper constructs a validated, dated measure of firms' first reported AI deployment or committed investment from sixteen years of earnings-call transcripts. Matching reporting adopters to observably similar firms provides a common framework for studying employment, revenue, and labour productivity across the arrival of generative AI. We document three core findings in the paper. First, the costs of adoption behave as the productivity J-curve predicts. Before 2023Q1, reported adoption is followed by slower employment and revenue growth with no improvement in revenue per employee, and the productivity differential does not close as firms age past their adoption dates; it widens, with part of the late deepening reflecting the pandemic years rather than adoption itself. Second, the recovery does not behave as the J-curve predicts. It arrives at one calendar date, the first full quarter after ChatGPT's release, when the productivity differential of the 683-firm installed base rises by 1.99 points per year. This result is robust to a range of tests. Third, the release broadened adoption. Entry roughly doubled, the firms entering are about half the size of the earlier adopters and less concentrated in the United States, and their post-adoption record so far shows neither the revenue and productivity costs earlier cohorts paid nor a statistically significant gain of their own. Throughout the sample, adopters' employment grows more slowly than that of their matched comparison firms. This employment-growth gap narrows in none of the baseline designs; at the installed-base boundary it widens slightly, short of significance at 5 per cent.

The reading we propose is that the recovery half of the J-curve was not automatic. In the standard account, returns arrive as each firm's complementary investment matures, on the firm's own clock; here they arrived only when the technology improved. Whether the early adopters' investment would eventually have paid off without generative AI cannot be known, but the timing evidence ties the return to its arrival. On this reading, the jump in capability rewarded internal reorganisation, of data, processes, and skills, that these firms had probably already carried out; such complements would transfer across forms of AI, so the date of the payoff was set by the technology's capability rather than by the maturing of each firm's own investment. The persistent employment-growth gap could reflect continuing reorganisation, labour-saving use, or differences in underlying growth trajectories, and the data do not distinguish among those channels.

The from-2023Q1 window remains short, particularly for new reporting adopters at longer event horizons. The clearest test of this interpretation will come as the panel lengthens; the dataset underlying the paper will be hosted by NL Analytics and kept updated at regular intervals. If accumulated complements are what convert capability into measured productivity, the generative AI era cohorts' gains should emerge as their own complementary investments mature, without another capability jump, and should be led by the firms investing most heavily; if their path instead stays flat while the installed base's gain persists, the advantage from AI was a first-mover phenomenon that a universal technology no longer confers. The same

panel will show whether the installed base's gain was a one-off level adjustment or the start of a faster path, and whether the employment-growth gap fades. The paper contributes a reproducible firm-level measure and evidence that the returns to reported AI adoption arrived with the technology's generative AI turn rather than with the passage of time since adoption.

References

D. Acemoglu. The Simple Macroeconomics of AI. *Economic Policy*, 40(121):13–58, 2025.

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

D. Acemoglu, D. Autor, J. Hazell, and P. Restrepo. Artificial Intelligence and Jobs: Evidence from Online Vacancies. *Journal of Labor Economics*, 40(S1):S293–S340, 2022.

I. Aldasoro, L. Gambacorta, R. Pal, D. Revoltella, C. Weiss, and M. Wolski. AI Adoption, Productivity and Employment: Evidence from European Firms. Technical report, BIS Working Papers No 1325, 2026.

L. Alekseeva, J. Azar, M. Giné, S. Samila, and B. Taska. The Demand for AI Skills in the Labor Market. *Labour Economics*, 71:102002, 2021.

A. Atkeson and P. J. Kehoe. Modeling the Transition to a New Economy: Lessons from Two Technological Revolutions. *American Economic Review*, 97(1):64–88, 2007.

T. Babina, A. Fedyk, A. He, and J. Hodson. Artificial Intelligence, Firm Growth, and Product Innovation. *Journal of Financial Economics*, 151:103745, 2024.

T. Babina, A. Fedyk, A. He, and J. Hodson. Firm Investments in Artificial Intelligence Technologies and Changes in Workforce Composition. In S. Basu, L. Eldridge, J. Haltiwanger, and E. Strassner, editors, *Technology, Productivity, and Economic Growth*, pages 75–117. University of Chicago Press, 2025.

J. M. Barrios, J. L. Campbell, R. G. Johnson, and C. Liu. Signals or Smoke? The Determinants and Informativeness of Corporate Artificial Intelligence (AI) Disclosures. Technical report, SSRN Working Paper 5133107, 2025.

J. Becker, N. Rush, E. Barnes, and D. Rein. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. Technical report, METR, arXiv:2507.09089, 2025.

A. Bick, A. Blandin, and D. J. Deming. The Rapid Adoption of Generative AI. *Management Science*, 2026. doi: 10.1287/mnsc.2025.02523. Published online 20 January.

K. Bonney, C. Breaux, C. Buffington, E. Dinlersoz, L. Foster, N. Goldschlag, J. Haltiwanger, Z. Kroff, and K. Savage. Tracking Firm Use of AI in Real Time: A Snapshot from the Business Trends and Outlook Survey. Technical report, NBER Working Paper 32319, 2024.

T. F. Bresnahan and M. Trajtenberg. General Purpose Technologies: 'Engines of Growth'? *Journal of Econometrics*, 65(1):83–108, 1995.

T. F. Bresnahan, E. Brynjolfsson, and L. M. Hitt. Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence. *Quarterly Journal of Economics*, 117(1):339–376, 2002.

E. Brynjolfsson, D. Rock, and C. Syverson. Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. In A. Agrawal, J. Gans, and A. Goldfarb, editors, *The Economics of Artificial Intelligence: An Agenda*, pages 23–57. University of Chicago Press, 2019.

E. Brynjolfsson, D. Rock, and C. Syverson. The Productivity J-Curve: How Intangibles Complement General Purpose Technologies. *American Economic Journal: Macroeconomics*, 13(1):333–372, 2021.

E. Brynjolfsson, B. Chandar, and R. Chen. Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence. Technical report, Stanford Digital Economy Lab Working Paper, 2025a.

E. Brynjolfsson, D. Li, and L. Raymond. Generative AI at Work. *Quarterly Journal of Economics*, 140(2): 889–942, 2025b.

M. Ca' Zorzi, G. Lopardo, and A.-S. Manu. Verba Volant, Transcripta Manent: What Corporate Earnings Calls Reveal about the AI Stock Rally. Technical report, ECB Working Paper No 3093, 2025.

B. Callaway and P. H. C. Sant'Anna. Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*, 225(2):200–230, 2021. doi: 10.1016/j.jeconom.2020.12.001.

F. Calvino and L. Fontanelli. A Portrait of AI Adopters across Countries: Firm Characteristics, Assets' Complementarities and Productivity. Technical report, OECD Science, Technology and Industry Working Papers 2023/02, 2023.

C. A. Corrado, C. R. Hulten, and D. E. Sichel. Intangible Capital and U.S. Economic Growth. *Review of Income and Wealth*, 55(3):661–685, 2009.

K. Z. Cui, M. Demirer, S. Jaffe, L. Musolff, S. Peng, and T. Salz. The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers. *Management Science*, 2026. doi: 10.1287/mnsc.2025.00535. Published online 27 February.

D. Czarnitzki, G. Fernandez, and C. Rammer. Artificial intelligence and firm-level productivity. *Journal of Economic Behavior & Organization*, 211:188–205, 2023.

P. A. David. The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox. *American Economic Review*, 80(2):355–361, 1990. Papers and Proceedings.

F. Dell'Acqua, E. McFowland III, E. Mollick, H. Lifshitz, K. C. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani. Navigating the Jagged Technological Frontier: Field Experimental Evidence

of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organization Science*, 37(2):403–423, 2026.

E. W. Dillon, S. Jaffe, N. Immorlica, and C. T. Stanton. Shifting Work Patterns with Generative AI. Technical report, NBER Working Paper 33795, November 2025. Revised version.

A. L. Eisfeldt and D. Papanikolaou. Organization Capital and the Cross-Section of Expected Returns. *Journal of Finance*, 68(4):1365–1406, 2013.

T. Eloundou, S. Manning, P. Mishkin, and D. Rock. GPTs are GPTs: Labor Market Impact Potential of LLMs. *Science*, 384(6702):1306–1308, 2024.

E. W. Felten, M. Raj, and R. Seamans. Occupational, Industry, and Geographic Exposure to Artificial Intelligence: A Novel Dataset and Its Potential Uses. *Strategic Management Journal*, 42(12):2195–2217, 2021.

F. Filippucci, P. Gal, and M. Schief. Miracle or Myth? Assessing the Macroeconomic Productivity Gains from Artificial Intelligence. OECD Artificial Intelligence Papers 29, OECD Publishing, Paris, 2024.

A. Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021.

J. Greenwood and M. Yorukoglu. 1974. *Carnegie-Rochester Conference Series on Public Policy*, 46:49–95, 1997.

M. Hampole, D. Papanikolaou, L. D. W. Schmidt, and B. Seegmiller. Artificial Intelligence and the Labor Market. Technical report, NBER Working Paper 33509, September 2025. Revised version.

T. A. Hassan, S. Hollander, L. van Lent, and A. Tahoun. Firm-Level Political Risk: Measurement and Effects. *Quarterly Journal of Economics*, 134(4):2135–2202, 2019.

E. Helpman and M. Trajtenberg. A Time to Sow and a Time to Reap: Growth Based on General Purpose Technologies. In E. Helpman, editor, *General Purpose Technologies and Economic Growth*, pages 55–83. MIT Press, Cambridge, MA, 1998.

G. Hoberg and G. Phillips. Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*, 124(5):1423–1465, 2016.

S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6 (2):65–70, 1979.

A. Humlum and E. Vestergaard. Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI. Technical report, NBER Working Paper 33777, revised March 2026; previously circulated as "Large Language Models, Small Labor Market Effects", 2025.

E. Ide and E. Talamàs. Artificial Intelligence in the Knowledge Economy. *Journal of Political Economy*, 133(12):3762–3800, 2025.

M. Jha, J. Qian, M. Weber, and B. Yang. ChatGPT and Corporate Policies. Technical report, NBER Working Paper 32161, 2024.

A. Kalyani, N. Bloom, M. Carvalho, T. A. Hassan, J. Lerner, and A. Tahoun. The Diffusion of New Technologies. *Quarterly Journal of Economics*, 140(2):1299–1365, 2025.

K. McElheran, J. Li, E. Brynjolfsson, Z. Kroff, E. Dinlersoz, L. Foster, and N. Zolas. AI adoption in America: Who, what, and where. *Journal of Economics & Management Strategy*, 33(2):375–415, 2024.

National Institute of Standards and Technology. The Language of Trustworthy AI: An In-Depth Glossary of Terms. NIST AI Resource Center, `https://airc.nist.gov/glossary`, 2023.

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

OECD. OECD.AI Policy Observatory: AI Terms and Concepts. `https://oecd.ai`, 2024.

N. Otis, R. Clarke, S. Delecourt, D. Holtz, and R. Koning. The Uneven Impact of Generative AI on Entrepreneurial Performance. Technical report, Harvard Business School Working Paper 24-042, 2023.

S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. Technical report, Working paper, arXiv:2302.06590, 2023.

Z. Sautner, L. van Lent, G. Vilkov, and R. Zhang. Firm-Level Climate Change Exposure. *Journal of Finance*, 78(3):1449–1498, 2023.

P. E. Soto. Research in Commotion: Measuring AI Research and Development through Conference Call Transcripts. Finance and Economics Discussion Series 2025-011, Board of Governors of the Federal Reserve System, Washington, DC, 2025.

L. Sun and S. Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021.

M. Webb. The Impact of Artificial Intelligence on the Labor Market. Technical report, SSRN Working Paper / Stanford (No. 3482150), 2020.

C.-H. Yang. How Artificial Intelligence Technology Affects Productivity and Employment: Firm-Level Evidence from Taiwan. *Research Policy*, 51(6):104536, 2022.

I. Yotzov, J. M. Barrero, N. Bloom, P. Bunn, S. J. Davis, K. M. Foster, A. Jalca, B. H. Meyer, P. Mizen, M. A. Navarrete, P. Smietanka, G. Thwaites, and B. Z. Wang. Firm Data on AI. Technical report, NBER Working Paper 34836, 2026.

N. Zolas, Z. Kroff, E. Brynjolfsson, K. McElheran, D. Beede, C. Buffington, N. Goldschlag, L. Foster, and E. Dinlersoz. Advanced Technologies Adoption and Use by U.S. Firms: Evidence from the Annual Business Survey. Technical report, NBER Working Paper 28290 (also U.S. Census CES Working Paper 20-40), 2020.

A Keyword screen

The first stage sends every transcript sentence that matches at least one term in Table 6 to the classifier. Matching is case-insensitive and uses whole-word or phrase boundaries, except where a wildcard permits a suffix. The list combines general terms from the National Institute of Standards and Technology (NIST) and the Organisation for Economic Co-operation and Development (OECD) glossaries (National Institute of Standards and Technology, 2023; OECD, 2024) with model families, products, and applications found in corporate language. Candidate terms were audited by reading samples of the additional calls they flagged. Ambiguous abbreviations, notably "ml", were removed when false matches dominated. The screen is deliberately broad, and the classification stage that follows removes the sentences that do not indicate adoption.

The vocabulary is intentionally broader than the event definition. General terms capture longstanding machine-learning systems, while named model families and application terms capture language introduced during the generative AI era. Each proposed addition was evaluated at the margin. The audit counted the extra calls admitted by that term and reviewed a sample of those calls directly. Terms were retained when they recovered genuine AI discussion without being dominated by unrelated corporate language. The screening stage therefore errs towards including sentences. A sentence dates adoption only if the classifier subsequently labels it as current deployment or committed investment with an evidence score of at least 80.

Table 6: Keyword-screen terms

Category	Terms
General AI and techniques	ai, artificial intelligence, generative-ai, genai, gen-ai, machine learning, deep learning, natural language processing, nlp, computer vision, augmented intelligence, aigc, neural network∗, reinforcement learning, supervised learning, unsupervised learning, backpropagation, hyperparameter∗, hallucination∗, supercomput∗, quantum comput∗.
Model types	generative model∗, foundation model∗, large language model∗, llm∗, diffusion model∗, transformer model∗.
Model families	ChatGPT (and -3.5, -3.5-turbo, -4, -4-turbo, -4o variants), GPT, GPT-2, GPT-3, GPT-3.5, GPT-3.5-turbo, GPT-4, GPT-4-turbo, GPT-4o, Claude, Claude 2, Claude 3 Haiku, Claude 3 Sonnet, Claude 3 Opus, Gemini, Gemini 1.5 Flash, Gemini 1.5 Pro, Gemini Ultra 1.0, Gemini Nano 1, Gemini Nano 2, PaLM, PaLM 2, PaLM-E, Bard, Llama, Llama 2 (7B, 13B, 70B), Llama 3 (8B, 70B), Falcon (7B, 40B, 180B), Mistral, Mistral 7B, Mistral 7B Instruct, Mistral Large, Mixtral, Mixtral 8x7B, Mixtral 8x22B, Jurassic, Jurassic-1 (Jumbo, Grande), Jurassic-2 (Ultra, Mega, Mid, Lite), Jamba, Jamba Instruct, Gemma, Gemma 2B, Gemma 7B, Grok, Grok-1, Grok-1.5, Grok-1.5 Vision, Grok 4, Grok 4 Heavy, ERNIE Bot, ERNIE Bot 4.0, ERNIE Bot 4.5, Qwen, Qwen-1.5-7B, Qwen-1.5-14B, Qwen2-7B, Qwen2-72B, Tongyi, Tongyi Qianwen, Yi, Yi-34B, Yi-34B-Chat, Yi-9B, DeepSeek∗, Phi, Phi-3 (Mini, Small, Medium), Command, Command R, Command R+, PPLX, PPLX-7B-Online, PPLX-70B-Online, PPLX-70B-Chat.
Image generation	DALL-E (and 2, 3), Stable Diffusion (v1.5, v2.1, XL), SDXL, SDXL 1.0, SDXL Turbo, Midjourney (v1–v6.1), Firefly, Firefly Image Model 1–4, Firefly Video Model, Imagen, Imagen 2, Imagen 3, Ideogram (v1, v2, v3).
Video generation	Runway, Runway Gen-1, Gen-2, Gen-3 Alpha, Pika, Pika 1.0, 1.1, 1.5, Sora, Veo.
Applications	text-to-image, text to image, text-to-video, text to video, code completion, retrieval-augmented generation, chatbot∗, copilot, co-pilot, prompt engineering.

Notes. Terms are matched case-insensitively on whole-word and phrase boundaries. "∗" permits a suffix.

B Classifier prompt and validation

Production prompt. GPT-4.1-mini classifies each screened sentence. The model receives the flagged sentence together with the sentence before and the sentence after it, and knows the text comes from an earnings call, but it does not receive the rest of the transcript.

```
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 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): indicates 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, investing in, or implementing AI. Or
  announcements about future plans to invest in AI that will clearly not
  happen within the next 3 months.

Provide a confidence score:
- Provide an integer confidence score for each earnings call snippet,
  ranging from 0 to 100, representing how strongly the earnings call
  snippet suggests evidence of AI employment.
- Confidence score must be consistent with the label: higher than 50 if
  label=1, lower than 50 if label=0.

Output STRICTLY as valid JSON (no extra text, no markdown) matching this
schema:
{
  "label": 0,
  "confidence": 0
}
Where "label" is 0 or 1, and "confidence" is an integer from 0 to 100.
```

The prompt names the 0–100 field "confidence", but the output is an ordinal deployment-evidence score rather than a calibrated probability. A firm-quarter qualifies when it contains at least one sentence labelled 1 with a score of at least 80, and the first such quarter dates the event.

Alternative prompt 1.

```
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): indicates the company is deploying/using/implementing/integrating/piloting/investing 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 (no extra text, no markdown), following
exactly this schema:
{
  "label": 0
}
Where "label" must be 0 or 1.
```

Alternative prompt 2.

```
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): indicates the company is deploying/using/implementing/integrating/piloting/investing 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 (no extra text, no markdown), following
exactly this schema:
{
  "label": 0
}
Where "label" must be 0 or 1.
```

The validation sample contains 500 randomly selected screened sentences. The two authors and two research assistants labelled every sentence independently under the same deployment definition, then resolved disagreements to a consensus benchmark. The sample is close to balanced, with 52 per cent labelled positive by the human benchmark. As reported in main-text Table 1, the production classifier attains 87 per cent accuracy, 85 per cent precision, and 91 per cent recall. Consensus labels provide a common reference for model comparisons, but they do not eliminate the judgement involved in interpreting short transcript excerpts.

Main-text Figure 2 reports agreement with the consensus label by evidence-score band and predicted label.

Table 7 compares the production configuration with five alternative models available in December 2025 under the three prompts reproduced above. The empirical panel was classified using GPT-4.1-mini with the production prompt, and those classifications are cached and frozen. The production configuration attains 87 per cent accuracy. Alternative prompt 1 removes the score field and uses a shorter definition of employment; alternative prompt 2 adds explicit treatment of hiring, pilots, aspirations, and non-specific data-centre investment. Each model-prompt configuration in the validation comparison was evaluated once, at temperature zero where supported. Temperature zero reduces sampling variation but does not guarantee exact determinism. Later application programming interface (API) or model updates need not reproduce every classification exactly.

Table 7: Classifier accuracy by model and prompt

Model	Prompt	Accuracy	Precision	Recall
GPT-4.1-mini	Main†	0.87	0.85	0.90
	Alt. 1	0.74	0.67	0.98
	Alt. 2	0.87	0.88	0.87
GPT-4.1	Main	0.86	0.84	0.90
	Alt. 1	0.79	0.73	0.97
	Alt. 2	0.84	0.87	0.82
GPT-4o	Main	0.84	0.89	0.80
	Alt. 1	0.81	0.74	0.96
	Alt. 2	0.85	0.90	0.81
GPT-4o-mini	Main	0.85	0.83	0.90
	Alt. 1	0.70	0.64	0.97
	Alt. 2	0.85	0.83	0.90
GPT-5-mini	Main	0.78	0.71	0.99
	Alt. 1	0.69	0.63	1.00
	Alt. 2	0.81	0.74	0.99
GPT-5.2	Main	0.87	0.84	0.93
	Alt. 1	0.77	0.70	0.98
	Alt. 2	0.86	0.84	0.90

Notes. Performance against 500 hand-labelled sentences. Models are those available in December 2025. "Main" is the production prompt; "Alt. 1" and "Alt. 2" are the alternatives reproduced above. † marks the production configuration used to classify the empirical panel. Precision is the share of classifier-positive sentences confirmed by the human label; recall is the share of human-positive sentences recovered.

C What the reported AI is used for

The classifier that dates adoption records whether a firm deploys AI, not what for. A second pass assigns each classifier-positive sentence, at any evidence score, to one of four use types, product, internal, both, or none, using the same model (GPT-4.1-mini) and the system prompt below, reproduced from the production code. The production label none covers sentences whose use type cannot be determined, and Table 8 reports that category as unclear. The internal-use check in Section 6 rests on this classification, so we validate it the same way as the primary classifier.

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

```
Task:
Determine whether the text indicates that the company is:
- Using AI INTERNALLY for its own operations (label = "internal")
- Developing/producing AI PRODUCTS OR SERVICES for sale to customers
  (label = "product")
- Both (label = "both")
- Neither / unclear (label = "none")

Definitions:
- Internal use includes AI deployed for automation, workflow
  optimisation, efficiency, internal analytics, internal software tools,
  or improvements to the company's own processes.
- Product development includes AI models, AI-enabled software,
  platforms, analytics services, or AI-powered features intended for
  external customers or the market.
- If text mentions both internal use and external products, classify as
  "both".
- If the text is generic or speculative without clear indication,
  classify as "none".

Output STRICTLY as valid JSON (no extra text, no markdown), following
exactly this schema:
{
  "label": "internal" | "product" | "both" | "none",
  "rationale": "short explanation"
}
```

Of the 500 hand-labelled test sentences, 238 carry a final consensus hand label for use type, and Table 8 compares the production classifier against them. The classifier agrees with the hand label in 74.8% of sentences, against 64% from always guessing the most common category. Most disagreements involve the both and unclear categories; the classifier calls product where the human says internal, or the reverse, in only six sentences. On the margin the check uses, sentences where both the human and the classifier make a decisive product-or-internal call, agreement is 95.7%. The firm-level dominant-mode rule averages over multiple positive disclosures. If classification error is nondifferential, the resulting mixing tends to attenuate an internal–product contrast; differential error could instead bias it in either direction. The limited number of manually labelled internal sentences further motivates treating this analysis as supporting evidence rather than a formal contrast.

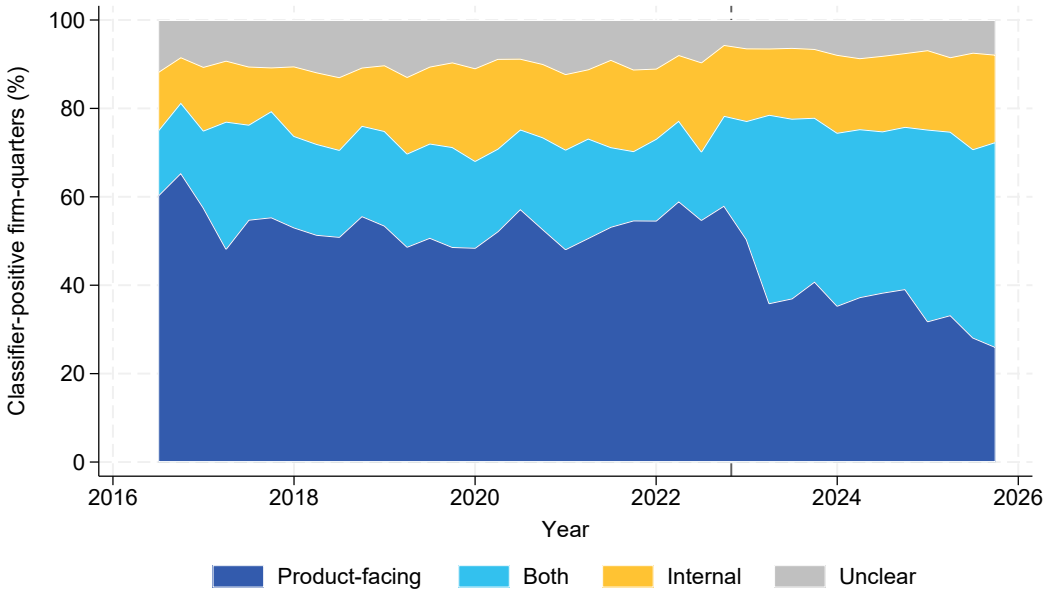
Table 8: Use-type classifier against manual labels

Manual label	Sentences	Classifier agreement (%)
Product-facing	153	76.5
Internal operations	23	78.3
Both	29	65.5
Unclear	33	72.7
All labelled sentences	238	74.8
Product against internal, decisive pairs	141	95.7

Notes. Agreement between the production use-type classifier and the final hand labels on the 238 labelled sentences of the validation sample. The final row restricts to sentences where both the human label and the classifier prediction are either product or internal, the margin used in the internal-use check of Section 6.

The classification also shows how the character of adoption shifted over time. Figure 12 plots the composition of classifier-positive firm-quarters at any evidence score. The unclear share falls steadily as disclosures grow more specific, and after ChatGPT the mix tilts towards firms reporting product-facing and internal use at once. The "both" category rises from about a fifth of classifier-positive firm-quarters before ChatGPT to two-fifths after, while the purely product-facing share falls from just over half to a third. Adoption broadens from a product feature into a tool used across the firm.

Figure 12: What firms use AI for, over time



Notes. Composition of classifier-positive firm-quarters by use type, from a secondary classification of classifier-positive sentences at any evidence score. After ChatGPT the share reporting both product-facing and internal deployment rises sharply. Quarters before mid-2016, with fewer than fifty classified firm-quarters, are omitted.

Table 9 re-estimates the installed base's change from 2023Q1 with the adopters grouped by these labels. Internal use only at disclosure requires an internal label and no product label in the disclosure quarter itself; internal-dominant and product-dominant adopters have more internal than product labels, or the reverse, across their pre-2023 classifier-positive quarters. The change is positive in every group and stable or larger when adoption is internal use. The grouping cuts the 683 pairs to between 183 and 410, roughly doubling the standard errors, which is why estimates that match or exceed the baseline magnitude sit at or outside conventional significance levels; the split is accordingly underpowered for a formal internal-versus-product contrast.

Table 9: The installed base's change from 2023Q1, by adoption mode

Adopter group	Change	(SE)	p	Pairs
All installed-base adopters	$+1.99^{**}$	(0.93)	0.033	683
Internal use only at disclosure	$+1.78$	(1.80)	0.323	183
Internal-dominant before 2023Q1	$+2.82$	(1.76)	0.109	187
Product-dominant before 2023Q1	$+1.72$	(1.11)	0.121	410

Notes. The common-date fixed-cohort productivity specification of Table 4, re-estimated keeping only the matched pairs whose adopter belongs to the stated group. Groups are defined from non-imputed classifier-positive quarters; adopters with equal internal and product counts appear only in the first row. Significance as in Table 3.

D Outcome definitions and sample composition

Table 10 lists the financial outcomes, source items, and transformations. The estimation panel contains 236,379 earnings calls for 10,706 listed firms in 88 countries and covers 250,798 firm-quarters from 2010Q1 to 2025Q4. Matching retains 1,067 reporting-adopter–comparison assignments drawn from 672 distinct comparison firms. Table 2 reports the pre-match composition.

The growth measures use four-quarter log differences to remove regular seasonality and firm scale. Revenue and revenue per employee are converted to US dollars in the source data and remain nominal. Cost and balance-sheet ratios are level outcomes rather than growth rates. The matched regressions use every non-missing outcome observation, so observation counts differ across financial variables even when the number of matched assignments is fixed. Employment, research and development, and cost-of-goods-sold coverage vary especially across sectors and countries.

Table 10: Outcome definitions and transformations

Outcome	From the accounts	Transformation	Reported as
Employment growth	Staff count	Year-on-year log change, $\Delta_4 \log x_t = \log x_t - \log x_{t-4}$	pp per year
Revenue growth	Net revenue	Year-on-year log change $\Delta_4 \log x$	pp per year
Labour productivity growth	Net revenue / staff count	Year-on-year log change, equal to revenue growth less employment growth	pp per year
Overhead ratio	SG&A expense / net revenue	Level of the ratio	pp of revenue
Production-cost ratio	Cost of goods sold / net revenue	Level of the ratio	pp of revenue
Capital expenditure	Capital expenditure / total assets	Level of the ratio, using the absolute value of capital expenditure, which the source records as a negative cash flow	pp of assets
Research and development	R&D expense / total assets	Level of the ratio	pp of assets
Intangible assets	Intangible assets / total assets	Level of the ratio	pp of assets
Debt	Total debt / total assets	Level of the ratio	pp of assets

Notes. Quarterly accounts are from S&P Capital IQ. Growth outcomes are four-quarter log differences multiplied by 100 and are read in percentage points per year. Ratios are percentage points of the denominator. Each outcome is winsorised at the first and 99th percentiles of the full sample.

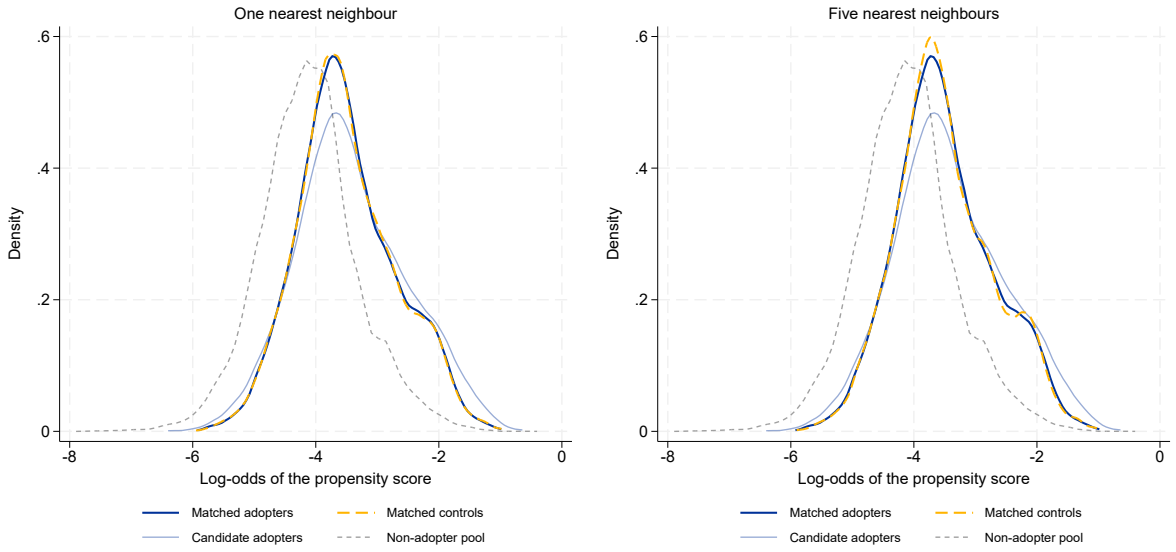
E Matching and overlap diagnostics

Figure 13 plots the propensity-score overlap behind the baseline one-neighbour match and the five-neighbour variant. Candidate reporting adopters have higher scores than the eligible comparison pool, while the selected comparison firms closely track matched adopters. Matching is exact on sector and anchor quarter and applies the propensity-score and growth-path calipers defined in Section 4. Control firms may contribute more than one assignment when reused and are censored before their own qualifying disclosure.

The overlap plot should be read together with the standardised-difference plot in main-text Figure 6. The propensity score summarises the multivariate adoption prediction, while the growth-path restrictions impose direct balance on the variables most closely related to the employment and revenue outcomes. Neither diagnostic tests balance on unobserved readiness or disclosure incentives. The alternative designs therefore

change both the breadth of the comparison pool and the treatment of the pre-adoption outcome path rather than relying on a single matching diagnostic.

Figure 13: Propensity-score distributions before and after matching



Notes. Kernel densities of the log-odds of the propensity score in equation (1). Candidate adopters are measured at their disclosure quarter. Comparison-pool observations use the anchor quarters at which a firm is eligible; selected comparison firms are weighted by reuse. Scores at numerical bounds from perfectly separated cells, about eight per cent of observations, are omitted. Left, the baseline one-nearest-neighbour match; right, the five-nearest-neighbour variant.

Table 11 reports the productivity estimates under the overlap-free re-match, direct adjustment for pre-adoption growth paths, and up to five nearest neighbours. The productivity reversal column is the calendar-era change of Panel B of Table 3, the from-2023Q1 average of the adopter–comparison productivity differential minus its before-2023Q1 average, re-estimated under each design. It stays positive and significant in every design, and the installed-base change stays between 1.65 and 2.42; the entry-fixed Sun–Abraham contrast remains positive but imprecise.

Table 11: Productivity estimates under alternative matched designs

Design	Pairs	Productivity, before 2023Q1	Productivity reversal	Installed-base change	Entry-fixed high vs low
Baseline match	1,067	-1.13^{*}	$+3.28^{***}$	$+1.99^{**}$	$+3.43$
Five nearest neighbours	1,067	-0.99	$+2.83^{***}$	$+2.42^{***}$	$+3.35$
Overlap-free re-match	1,137	-1.62^{**}	$+3.82^{***}$	$+2.12^{**}$	$+1.60$
Pre-path adjustment	1,067	-0.69	$+3.72^{***}$	$+1.65^{*}$	$+2.84$

Notes. The productivity columns report the before-2023Q1 Sun–Abraham estimate and the productivity reversal, its from-2023Q1-minus-before-2023Q1 calendar-era change; the remaining columns report the common-date fixed-cohort installed-base change and the entry-fixed above-median-minus-at-or-below-median intensity contrast. The intensity column uses a joint group-by-cohort-by-event-time Sun–Abraham regression over $k = 0{:}11$. The five-neighbour row widens the baseline match. The pre-path row adds firm-specific pre-adoption productivity- and employment-growth levels and trends interacted with the post indicator. Significance as in Table 3.

Table 12 records the search that sets the revenue-growth caliper in the baseline match. The rule, fixed before estimation, is that every employment- and revenue-growth history covariate must balance to a standardised mean difference of at most 0.1 after matching, and the caliper takes the loosest width that satisfies it. The loosest width searched already does, so the binding consideration is retention rather than balance; tightening further would discard up to 207 additional adopters for balance the rule does not require.

Table 12: The revenue-growth caliper search

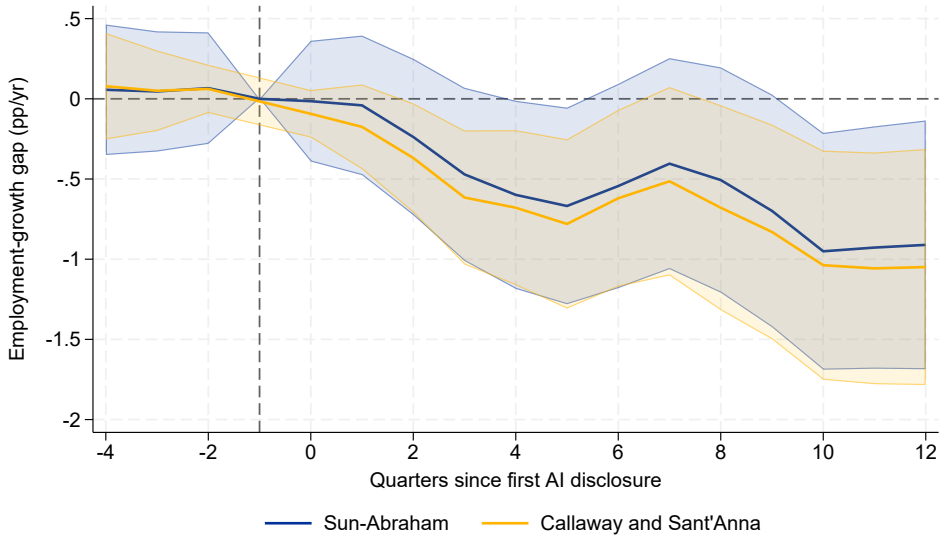
Revenue caliper (SD)	Adopters retained	Largest growth-history SMD
1.5	1,067	0.081
1.2	1,034	0.055
1.0	996	0.040
0.8	942	0.034
0.6	860	0.057

Notes. Each row re-runs the baseline matcher with the stated revenue-growth caliper, in standard deviations, applied to the eight-quarter average, recent four-quarter average, and trend of revenue growth. Retention counts matched adopters among the 1,557 candidates with complete covariates. The final column is the largest post-match standardised mean difference across the six employment- and revenue-growth history covariates. The paper uses the 1.5 row. The employment calipers are tighter than the revenue caliper because both implement the same balance requirement under different distributions. Employment growth is comparatively stable across firms and its three covariates balance well inside 0.1 at the 0.6 width, while revenue growth is more dispersed, so balance is achieved at a looser width and the search shows that tightening it further discards adopters without being required by the rule.

F Alternative estimators and event paths

Figure 14 compares employment paths from the preferred Sun–Abraham model and the pooled-control estimator of Callaway and Sant'Anna. Both are negative after the disclosure event. The second series uses a different comparison construction and effective horizon.

Figure 14: Employment growth around first reported AI adoption



Notes. Event-study paths for year-on-year employment growth with 95 per cent confidence bands. Sun–Abraham is the preferred matched-window estimate. The Callaway and Sant'Anna estimator pools distinct comparison firms observed while eligible and before any own qualifying disclosure, and uses a broader effective horizon. It is not the same matched-window estimand. Event quarter -1 is the reference period.

F.1 Pooled-distinct-control group-time estimates

The pooled-distinct-control group-time check of Callaway and Sant'Anna (2021) first collapses repeated assignments to one observation per physical firm and quarter. For cohort g and quarter t, it computes

$$\widehat{ATT}(g,t) = (\bar{Y}_{g,t} - \bar{Y}_{g,g-1}) - (\bar{Y}_{C,t} - \bar{Y}_{C,g-1}),$$

where C is the pooled set of distinct comparison firms observed in the matched sample while eligible and before any own qualifying disclosure. Quarter $g-1$ is the universal cohort-specific base. Post-adoption cells are aggregated with cohort-size weights, and calendar-era columns pool cells by the calendar quarter in which they occur. The construction does not preserve pair identifiers, reuse weights, or the fixed event window. It includes no regression covariates or firm or calendar-quarter fixed effects. Influence-function contributions, including the contribution from estimated cohort weights, are summed by physical firm before inference. These differences make Table 13 a supporting sign-and-timing check, not an estimate of the

identical matched-window effect.

Table 13: Pooled-distinct-control group-time estimates

Outcome	Overall	Before 2023Q1	From 2023Q1	Change
Employment growth	-0.88^{***}	-0.63^{*}	-1.13^{***}	-0.50
	(0.28)	(0.34)	(0.35)	(0.39)
Revenue growth	-0.97	-1.88^{**}	-0.05	$+1.84^{**}$
	(0.68)	(0.75)	(0.85)	(0.86)
Labour productivity growth	-0.09	-1.25^{*}	$+1.08$	$+2.33^{***}$
	(0.66)	(0.73)	(0.82)	(0.82)
Overhead ratio	-0.24	-0.77	$+0.29$	$+1.06$
	(0.60)	(0.74)	(1.01)	(1.31)
Production-cost ratio	$+0.26$	$+0.47$	$+0.05$	-0.42
	(0.56)	(0.58)	(0.71)	(0.67)

Notes. Group-time differences relative to cohort quarter $g-1$, with cohort-size aggregation. "Overall" pools available post-adoption cells; the era columns pool cells by calendar quarter, and "Change" is from-2023Q1 minus before-2023Q1. Growth outcomes are percentage points per year and cost ratios are percentage points of revenue. Standard errors aggregate influence-function contributions by physical firm. Significance as in Table 3.

G Time effects and missing calls

Table 14 absorbs increasingly fine sectoral and geographic shocks for the installed-base common-date productivity change.

Table 14: Sectoral and geographic time controls

Time controls	Installed-base productivity change
Firm + quarter FE	$+1.99^{**}$
	(0.93)
Sector-by-quarter FE	$+2.33^{***}$
	(0.90)
Region-by-quarter FE	$+2.05^{**}$
	(0.93)
Country-by-quarter FE	$+1.92^{**}$
	(0.94)

Notes. Percentage-point estimates of the common-date fixed-cohort productivity-growth change from 2023Q1 among firms first reporting before 2023Q1. Each row absorbs physical-firm effects and the stated time interactions. Significance as in Table 3.

G.1 Missing call quarters

Firms reporting semi-annually can have gaps in alternate call quarters. Gaps of one or two quarters carry forward the preceding call's classification flags but receive zero mentions and no evidence score; longer gaps truncate the transcript panel. A carried row cannot create an event or add intensity. Carried rows constitute 5.9 per cent of the matched panel. Table 15 removes them. The pooled productivity change and installed-base estimate retain their signs, the entry-fixed Sun–Abraham contrast remains positive, and the employment estimate loses precision.

Table 15: Headline estimates with and without carried firm-quarters

Estimate	Sun–Abraham		Static specifications		Callaway and Sant'Anna	
	With carried quarters	Without carried quarters	With carried quarters	Without carried quarters	With carried quarters	Without carried quarters
Employment growth	-0.54^{**}	-0.43	-0.69^{***}	-0.65^{**}	-0.88^{***}	-0.79^{***}
Revenue growth	-0.65	-0.65	-0.66	-0.67	-0.97	-1.12
Productivity, before 2023Q1	-1.13^{*}	-1.18^{*}	–	–	-1.25^{*}	-1.30^{*}
Productivity, from 2023Q1	$+2.14^{**}$	$+1.93^{**}$	–	–	$+1.08$	$+0.69$
Productivity reversal	$+3.28^{***}$	$+3.11^{***}$	–	–	$+2.33^{***}$	$+1.99^{**}$
Installed-base gain	–	–	$+1.99^{**}$	$+1.80^{*}$	–	–
High-vs-low entry intensity	$+3.43$	$+3.89$	–	–	–	–

Notes. Estimates with carried firm-quarters included and removed. The static-specification columns report the pooled static TWFE estimates for employment and revenue and the common-date fixed-cohort installed-base estimate; the Sun–Abraham and static rows use the pooled matched sample. The Callaway and Sant'Anna estimator uses the distinct-control construction in Appendix F.1. The installed-base row is the common-date fixed-cohort estimate; the entry-fixed intensity row uses the joint group-by-cohort-by-event-time Sun–Abraham estimator over $k = 0{:}11$. A dash marks an unused combination. Significance as in Table 3.

H Accounting correlates

Table 16 reports the leverage and recorded-investment estimates discussed in Section 3, in percentage points of total assets alongside the matched-sample mean and sample sizes. Total-asset growth itself shows no significant adopter–comparison differential in either era, so the leverage association reflects the debt side rather than the denominator. Table 17 re-estimates the leverage ratio and the production-cost ratio of Table 3 without financial firms, whose leverage and cost structures differ from those of other sectors; the leverage estimates remain similar across eras, and neither their change nor the production-cost change is statistically significant.

Table 16: Balance-sheet and recorded-investment correlates of reported adoption

Ratio	Mean	Before 2023Q1	From 2023Q1	From 2023Q1, ex-tech	Obs.	Pairs
Debt / assets	28.20	+1.21**	+0.83	+1.46	106,499	1,067
		(0.57)	(0.91)	(0.92)		
R&D / assets	1.90	+0.01	−0.01	+0.00	45,701	665
		(0.08)	(0.08)	(0.08)		
Capex / assets	1.05	+0.01	−0.03	−0.04	99,972	1,043
		(0.03)	(0.05)	(0.05)		
Intangibles / assets	16.76	+0.09	+0.66	+0.54	55,742	942
		(0.43)	(0.78)	(0.77)		

Notes. Sun–Abraham era-specific estimates in percentage points of total assets. The mean is the matched-sample weighted mean. Capital expenditure uses the absolute value of the reported cash outflow; R&D/assets is winsorised at the first and 99th percentiles. "Ex-tech" excludes matched pairs whose adopter is headquartered in the US or China and has technology as its primary sector. Each outcome uses its available sample. Significance as in Table 3.

Table 17: Leverage and production costs outside finance

Outcome	Before 2023Q1	From 2023Q1	Change
Debt / assets	+1.21**	+0.95	−0.26
	(0.59)	(0.94)	(0.90)
COGS / revenue	+0.46	−0.53	−0.99
	(0.37)	(0.65)	(0.71)

Notes. Sun–Abraham era-specific estimates excluding financial firms. Debt/assets is in percentage points of assets and COGS/revenue in percentage points of revenue. "Change" is from-2023Q1 minus before-2023Q1. Significance as in Table 3.

I The generative AI cohort average and reported activity intensity

A short staggered panel can be averaged in several reasonable ways, so Table 18 re-estimates the post-adoption average for the 384 firms first reporting from 2023Q1 under five conventions, one per row. The first row weights every event quarter from $k = 0$ to 11 equally; because the panel ends in 2025Q4, only the earliest entrants are observed at the late event quarters, so those quarters are estimated from few firms. The second row weights each cohort–event-quarter cell by the number of firms observed in it, so the better-populated early quarters count for more. The two balanced rows keep only the entrants observable over a full eight- or five-quarter window, so exactly the same firms are compared at every event quarter, at the cost of a smaller sample. The final row is the group-time estimator of Callaway and Sant'Anna, which averages each

cohort over whatever horizon it is observed for. Under every convention the productivity point estimate is positive, between about 1.4 and 2.2 points, and significant under none, and the employment estimate is small and negative throughout, so the conclusion does not depend on the aggregation. The main text summarises these adopters with the support-weighted average measured against the group's average pre-disclosure differential rather than the final pre-disclosure quarter; netting off that pre-disclosure differential is why the figure quoted there is smaller than the entries here.

Table 18: The generative AI cohort average under five aggregation conventions

	Labour productivity		Employment		
Aggregation	Estimate	(SE)	Estimate	(SE)	Adopters
Equal event-time weights	+1.36	(1.83)	−0.38	(0.52)	384
Support-weighted	+1.77	(1.23)	−0.43	(0.32)	384
Balanced $k = 0{:}7$	+2.19	(1.74)	−0.33	(0.52)	143
Balanced $k = 0{:}4$	+1.95	(1.21)	−0.51	(0.34)	227
Callaway and Sant'Anna	+1.75	(1.20)	−0.46	(0.31)	384

Notes. Averages of post-adoption estimates for firms first reporting from 2023Q1, relative to matched comparisons. The first four rows are Sun–Abraham estimates with the stated aggregation of cohort–event-time cells; balanced windows restrict to cohorts observable over the full horizon ($k = 0{:}7$ requires first disclosure by 2024Q1, $k = 0{:}4$ by 2024Q4). The final row is the pooled-distinct-control group-time estimator of Appendix F.1. Outcomes are percentage points per year. Significance as in Table 3.

We also asked whether the entrants that discuss AI most at adoption go on to perform differently. Each firm's reported activity is measured as the number of classifier-positive deployment or committed-investment sentences in its entry quarter and the quarter after, and the measure is deliberately fixed at that point. If intensity were instead accumulated over later calls, a firm whose adoption was going well might discuss it more, and the ranking would partly reflect performance rather than predict it. Firms are then split at the median of this count, and Table 19 reports the productivity difference between the above-median and at-or-below-median groups. A finer split is not informative, because the count is discrete and many firms record exactly one sentence, so dividing the distribution into fifths would separate identical firms arbitrarily. The contrast is positive under the preferred joint Sun–Abraham specification, under a ranking that scales the count by the number of calls held in the window, and in a static within-panel specification, but it reaches conventional significance under none of them; the closest is the static contrast at $p = 0.075$, and randomisation inference, which reassigns the high and low labels at random across entrants, puts it at $p = 0.086$. A cumulative ranking over all generative AI era calls produces larger contrasts, but for the reason above we do not use it. The paper accordingly claims no relationship between reported activity intensity and subsequent productivity.

Table 19: Entry-fixed reported activity intensity, above versus at or below the median

	All new adopters	Excl. US & China tech
Sun–Abraham, above vs at/below median	+3.43	+2.95
	(3.12)	(3.47)
Sun–Abraham, ranked by activity per call	+3.70	+3.26
	(3.12)	(3.50)
Static within-panel contrast	+2.66*	+2.53
	(1.49)	(1.66)
Randomisation inference p	0.086	–

Notes. Above-median minus at-or-below-median productivity contrasts among firms first reporting from 2023Q1, with intensity fixed from the entry quarter and the following quarter. The Sun–Abraham rows estimate group–cohort–event-time cells jointly and average $k = 0{:}11$ relative to $k = -1$, with the contrast from the full covariance matrix. The static row interacts the post-disclosure indicator with the above-median flag under firm and calendar-quarter fixed effects. The randomisation-inference row permutes the entry-fixed labels across the 384 adopter pairs 500 times and reports the adjusted permutation p-value, the number of permutations with a static contrast at least as large in absolute value plus one, over 501. The exclusion column drops US and Chinese technology adopters while retaining the full-sample ranking. Outcomes are percentage points per year. Significance as in Table 3.

These specifications address particular forms of sorting but do not assign intensity exogenously. Unobserved management quality, complementary assets, or selective disclosure can affect both intensity and productivity.