Bonfiglioli, Alessandra; Crinò, Rosario; Filomena, Mattia; Gancia, Gino

Working Paper

Data, Power and Emissions: The Environmental Cost of AI

CESifo Working Paper, No. 12158

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Bonfiglioli, Alessandra; Crinò, Rosario; Filomena, Mattia; Gancia, Gino (2025) : Data, Power and Emissions: The Environmental Cost of AI, CESifo Working Paper, No. 12158, Munich Society for the Promotion of Economic Research - CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/331624

CES ifo Working Papers

12158
2025

Original Version: September 2025
This Version: November 2025

Data, Power and Emissions: The Environmental Cost of AI

Alessandra Bonfiglioli, Rosario Crinò, Mattia Filomena, Gino Gancia

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

Data, Power and Emissions: The Environmental Cost of AI

Alessandra Bonfiglioli[a,*], Rosario Crinò[b], Mattia Filomena[c] and Gino Gancia[d]

[a] *University of Bergamo, QMUL and CEPR*

[b] *University of Bergamo, CEPR and CESifo*

[c] *University of Bergamo*

[d] *University of Milano-Bicocca and CEPR*

November 5, 2025

Abstract

We study the environmental impact of artificial intelligence (AI) using a novel dataset that links measures of AI penetration, the location of data centers and power plants, and CO_2 emissions across US commuting zones between 2002 and 2022. Our analysis yields four main findings. First, exploiting a shift–share identification strategy, we show that localities more exposed to AI experience relatively faster emissions growth. Second, decomposition results indicate that scale effects dominate, while changes in industrial composition exert at most a weak mitigating effect; at the same time, electricity generation becomes more carbon intensive. Third, AI penetration raises dependence on non-renewable electricity. Fourth, proximity to data centers is a key driver of this effect, as nearby power plants shift toward greater fossil fuel use. These findings suggest that, absent a rapid decarbonization of power generation, the diffusion of AI is likely to exacerbate environmental externalities through the energy demand of data centers.

Keywords: Artificial Intelligence, Data Centers, Environment, Emissions, Pollution

JEL Classification: O33, Q55, R11.

---

*Corresponding author. Department of Economics, University of Bergamo, via dei Caniana 2, 24127, Bergamo, Italy. E-mail addresses: alessandra.bonfiglioli@unibg.it (A. Bonfiglioli), rosario.crino@unibg.it (R. Crinò), mattia.filomena@unibg.it (M. Filomena), ginoalessandro.gancia@unimib.it (G. Gancia). We thank Antonin Bergeaud, Philipp Hartmann, Isabelle Mejean, Laura Ogliari and participants at the 2nd TECH Workshop on Technology, Employment, Change Management and Human Well-Being (Gaeta, 2025) for useful comments. Rosario Crinò gratefully acknowledges financial support from the Italian Ministry for University and Research under the PRIN 2022 program (project title: AUTOmation, PROductivity and Wage INequality (AUTOPROWIN): Firms and Workers in Times of Economic Turmoil; project number: 2022FLBY7J; CUP: F53D23003060006), M4, C2, Investment 1.1 - financed by the European Union – Next Generation EU, D.D. MUR 104, February 2, 2022.

1 Introduction

Quantifying the carbon footprint of AI is a complex yet increasingly urgent task. According to self-reported data, Google's carbon emissions rose by nearly 50% between 2019 and 2023, primarily due to increased energy consumption in data centers—the core infrastructure supporting AI models. These facilities, which house the servers and hardware necessary for data storage and processing, are projected to account for 8% of US electricity demand by 2030, up from 3% in 2022 (Davenport et al., 2024).[1] Such a surge in power needs is likely to increase the use of electricity from fossil fuels and delay closings of obsolete coal-fired plants. Supporting this view, anecdotal evidence suggests that the rising energy needs from AI-related projects has already postponed planned plant closures in Kansas City, West Virginia and in the Salt Lake City region.

At the same time, AI is frequently promoted as a tool for mitigating climate change. Machine learning and other AI technologies are increasingly employed to optimize energy consumption, enhance resource efficiency and support low-carbon innovation. Moreover, AI-intensive industries generally exhibit lower emission intensities compared to traditional manufacturing sectors. However, comprehensive metrics quantifying the energy savings attributable to AI are currently lacking. As a result, the net environmental impact of AI remains an open and pressing question.

In this paper, we examine the relationship between AI adoption and air pollution, with a particular focus on the role of energy demand and data centers. To do so, we construct a novel dataset that integrates information on AI penetration, the geographic location of data centers and power plants, and local emissions of carbon dioxide (CO_2) and other air pollutants across 722 US commuting zones (CZs) over the period 2002-2022. This period coincides with the rise of the digital economy, cloud computing and early AI applications. To capture the carbon footprint of all these phenomena, we take a broad definition of AI as algorithms applied to big data, and we measure its penetration using changes in employment in data-intensive occupations. Since AI penetration is likely endogenous—potentially influenced by local productivity or demand shocks—OLS estimates cannot be interpreted causally.[2] To identify causal effects, we employ a shift-share instrumental variables strategy that leverages variation in AI exposure across CZs. This variation is driven by historical (pre-AI) differences in industrial composition across localities, combined with national trends in AI penetration at the industry level over time.

Our first result is that increased AI penetration leads to significantly higher emissions. This re-

[1] As further evidence of their growing significance, employment in US data centers increased by more than 60% from 2016 to 2023, according to the US Census Bureau's Quarterly Workforce Indicators.

[2] For example, local demand or productivity shocks may simultaneously drive both AI adoption and investment in green technologies.

sult is robust to a wide set of fixed effects, additional controls, alternative variable definitions and various strategies to account for local trends and unobserved shocks. Quantitatively, the difference in AI penetration between Kansas City (top quartile) and New Orleans (bottom quartile) explains approximately 41% of the observed difference in CO_2 emissions growth between the two CZs. The estimates also imply that, absent any AI penetration, the average CZ would have reduced its CO_2 emissions by 34.3% rather than the observed 25%, i.e., a 37% greater decline. While these figures should not be interpreted as counterfactual exercises, since nation-wide effects are differenced out in our empirical strategy, they nonetheless suggest that local AI penetration increases emissions relative to less exposed areas.

Next, following a conventional decomposition framework (Levinson, 2009), we investigate whether the higher emissions are due to a scale effect (an expansion in the size of economic activity), a composition effect (changes in the industry mix) or a technique effect (changes in average emission intensity). Our second result is that scale effects play a primary role, while there is weak evidence that composition effects may have contributed to reducing emissions. However, an important exception concerns emissions from electricity generation. Specifically, we find that AI penetration increases the carbon footprint of electricity generation even on a per capita basis, thereby raising its emission intensity.

To understand this result, we turn to highly detailed data on power plants, the facilities where electricity is generated. Our dataset includes total annual net electricity generation for each power plant, disaggregated by renewable and non-renewable energy sources. Our third result is that AI penetration increases the reliance of electricity generation on non-renewable sources. Consequently, the rise in emissions from electricity generation is largely driven by a shift toward more carbon-intensive energy mixes. This is likely to be because data centers, which are needed to support AI, require a stable and high-capacity energy supply, which is easier to obtain from non-renewable power.

To corroborate this hypothesis, we leverage granular data on the geographic distribution of data centers and power plants. Our fourth result is that proximity to data centers is associated with increased electricity generation from non-renewable sources. To address potential endogeneity in the location of data centers, we construct an instrumental variable based on the distance-weighted average AI exposure of neighboring CZs. We find that in areas with higher AI exposure, data centers tend to locate closer to power plants. Moreover, proximity to data centers is associated with power plants generating higher CO_2 emissions and relying more heavily on non-renewable energy sources.[3] These

[3] Although electricity can be transmitted over long distances, proximity remains important due to transmission losses. According to the US Energy Information Administration, average line losses amount to 7% of total electricity generated. As a result, in the average US state, approximately 80–90% of consumed electricity is produced locally.

findings reinforce concerns that rising energy demand from data centers may be slowing down the transition to cleaner energy systems.

Our findings contribute to a growing body of literature on the environmental impact of digital technologies. Earlier studies have primarily focused on ICT. For example, Lange et al. (2020) find that, despite some offsetting effects, overall digitalization increases energy consumption. More recent work attempts to estimate the carbon footprint of large language models (see, e.g., Luccioni et al., 2023). Other papers aim to assess how future expansions of US data centers will affect emissions. EPRI (2024) argues that in the short-run the process is likely to increase the use of fossil fuels, while the long-run impact depends on policy and clean energy procurement. Knittel et al. (2025) argue that data center load flexibility lowers costs but has ambiguous emissions effects, reducing CO_2 in renewable-rich grids while increasing it in fossil-heavy ones. Feher et al. (2025) find no effects of data centers on local electricity prices, while Benetton et al. (2023) show that cryptomining raises them.

To the best of our knowledge, no paper has yet causally linked AI diffusion to emissions. Our results challenge the prevailing view of AI as an environmentally friendly technology, suggesting instead that its current deployment may exacerbate environmental externalities. These effects are likely to persist unless AI adoption is accompanied by substantial improvements in energy-efficient infrastructure. An important caveat of our results is that our sample does not include the most recent innovations, such as large language models. While these technologies promise future efficiency gains that could facilitate decarbonization, training and operating today's large language models are considerably more energy-intensive than earlier AI applications captured in our data. Absent massive investments in clean energy, the next generation of AI may thus exert even greater short-run pressure on emissions.

From a methodological perspective, this paper refines the measure of AI penetration based on occupational data proposed by Bonfiglioli et al. (2025) and employs a shift-share identification strategy inspired by Autor and Dorn (2013) and Acemoglu and Restrepo (2020), among others. We are the first to apply this approach to study the carbon footprint of AI and to combine it with geolocated data on data centers and power plants. In doing so, we bring together the literature on the local effects of new technologies and environmental economics.

The remainder of the paper is structured as follows. Section 2 develops a theoretical framework for understanding how AI affects emissions and how these effects can be decomposed. Section 3 introduces the data and documents key stylized facts. Section 4 outlines the empirical framework and identification strategy. Section 5 estimates the impact of AI penetration on overall emissions and disentangles the contributions of the three channels identified in Section 2. Section 6 provides robustness

checks and addresses potential threats to identification. Section 7 uses granular data on power plants and data centers to examine the effect of AI penetration on CO_2 emissions from electricity generation. Section 8 concludes.

2 Theoretical Framework: Emissions, AI and Data Centers

We now sketch a theoretical framework that illustrates how AI can affect emissions. We start by decomposing emissions due to scale, composition and technique of production, using the method of Levinson (2009). Let $E(Y_c)$ be the total emissions from production in CZ $c \in C$, θ_{ci} the share of sector $i \in I$ in total output, and e_{ci} emissions per unit of output in sector i. Then, total emissions from production can be written as:

$$E(Y_c) = Y_c \cdot \sum_{i \in I} \theta_{ci} \cdot e_{ci}.$$

In differences:

$$dE(Y_c) = \underbrace{\bar{e}_c \sum_{i \in I} dY_{ci}}_{\text{scale}} + \underbrace{Y_c \sum_{i \in I} (e_{ci} - \bar{e}_c) d\theta_{ci}}_{\text{composition}} + \underbrace{Y_c \sum_{i \in I} \theta_{ci} \cdot de_{ci}}_{\text{technique}},$$

where $\bar{e}_c$ is the average emissions per unit of output. This decomposition shows that emissions can change for three reasons: (i) more output increases emissions (scale effect); (ii) changing the structure of the economy affects emissions (composition effect); and (iii) improved technologies or regulations reduce emissions per unit of output (technique effect). The technique effect can be further decomposed into the industry-level emission intensity plus an industry-CZ residual:

$$de_{ci} = de_i + d\epsilon_{ci}.$$

To study the effect of AI on emissions, we assume that different sectors adopt and benefit from AI to varying degrees. We posit a production function of the form:

$$Y_{ci} = F_i(L_{ci}, K_{ci}, A_{ci}),$$

where L_{ci} is employment, K_{ci} is capital and A_{ci} is the AI input. AI adoption, defined as an increase in A_{ci}, can influence total emissions through the three channels above. First, AI can increase overall productivity and output ($\partial Y_{ci}/\partial A_{ci} > 0$), which may raise emissions unless offset by cleaner

techniques or composition shifts. Second, AI may shift output shares, θ_{ci}. Third, AI can reduce e_{ci} by improving energy efficiency, optimizing processes and enabling better environmental monitoring ($\partial e_{ci}/\partial A_{ci} < 0$); or increase it, if it raises energy demand ($\partial e_{ci}/\partial A_{ci} > 0$).

To empirically test these effects, we need a measure of AI adoption. We first focus on sectors using AI and posit the following AI production function:

$$A_{ci} = D_{ci}^{\delta} H_{ci}^{1-\delta},$$

where D_{ci} are data and H_{ci} are specialized AI workers. Data are freely traded, while AI workers are mobile and paid a wage w_H. Demand for AI workers is given by:

$$H_{ci} = (1-\delta)\frac{\omega}{w_H} A_{ci},$$

where ω is the price of one unit of A. This implies that we can read AI adoption from the change in employment of AI workers:

$$\frac{dH_{ci}}{H_{ci}} = \frac{dA_{ci}}{A_{ci}}.$$

Next, we assume that data must be stored in data centers. The data supplied by data center d is:

$$D_d = \gamma Y^{\alpha} H_d^{1-\alpha},$$

where $Y = \sum_{c \in C} Y_c$ and H_d are AI workers. The factor γY^{α} is the total data generated as a by-product of economic activity, and $H_d^{1-\alpha}$ captures the labor requirement for data storage. Assuming a fixed cost F to build a data center, free entry pins down its size:

$$H_d = \frac{1-\alpha}{\alpha}\frac{F}{w_H}.$$

We do not model the location choice of data centers. However, the model suggests that the presence of AI workers captures both AI adoption and the activity of data centers. Data centers generate emissions:

$$E(D_d) = D_d \cdot e_d,$$

where e_d is emissions per unit of data. Total emissions in CZ c are:

$$E_c = E(Y_c) + \sum_{d \in \Omega_c} E(D_d),$$

where Ω_c is the set of data centers in CZ c.

In sum, AI affects emissions directly via data centers and indirectly through changes in scale, intensity and composition of output. Following the model, in the rest of the paper we will measure the overall AI penetration at the CZ level from changes in employment of AI-related workers, i.e., $\sum_{i \in I} H_{ci} + \sum_{d \in \Omega_c} H_d$. First, we will study how AI penetration affects total emissions, E_c. Next, we will explore the distinct role of the three components of emissions from production and of energy demand from data centers.

3 Data and Stylized Facts

Our sample includes 722 commuting zones that collectively cover the entire mainland US. Each CZ is observed at five-year intervals over 2002–2022.[4] In this section, we begin by presenting our main data sources; a more detailed discussion and descriptive statistics are provided in Appendix A. Next, we outline a number of stylized facts.

3.1 Data

3.1.1 Employment in Data-Intensive Occupations

Direct measures of the local diffusion and economic significance of AI remain unavailable. To proxy for its penetration across CZs, we start from a broad definition of AI as the application of algorithms to big data. This definition encompasses the major AI-related and algorithmic technologies that expanded rapidly over our sample period, such as search engines, targeted advertising, recommendation systems and chatbots, as well as more recent breakthroughs in large language models and generative AI, which began to materialize only toward the end of the period. At the core of these technologies lie two essential building blocks: large-scale data and advanced computational methods. Leveraging these building blocks in practice requires specialized software tools, which play two critical roles. First, they enable the collection, organization and analysis of big data, as well as the design, customization and deployment of machine learning algorithms that integrate AI applications into firms' operations. Second, they support the infrastructure on which AI depends. AI systems rely on vast

[4] CZs are clusters of counties characterized by strong internal commuting ties and relatively weak ties with other zones (Tolbert and Sizer, 1996). The CZs used in our analysis are the same as those in Autor and Dorn (2013), Autor et al. (2013) and Acemoglu and Restrepo (2020). When data for a variable are missing at an endpoint of an interval, we use the closest available year and express the resulting change as a five-year equivalent.

Table 1: Data-Intensive Occupations

SOC Code	SOC Definition	SOC Code	SOC Definition
151211	Computer Systems Analysts	151251	Computer Programmers
151221	Computer and Information Research Scientists	151252	Software Developers
151232	Computer User Support Specialists	151253	Software Quality Assurance Analysts and Testers
151241	Computer Network Architects	151254	Web Developers
151242	Database Administrators	151299	Computer Occupations, All Others
151243	Database Architects	152051	Data Scientists
151244	Network and Computer Systems Administrators		

Notes: Occupations are classified according to the 6-digit level of the 2018 Standard Occupational Classification (SOC).

databases, requiring data centers to store and process information at scale, a task that in turn depends on specialized software for storage management, data retrieval and systems administration.

In turn, operating the software tools that enable AI adoption and data center management requires specialized expertise, which is typically concentrated in a narrow set of data-intensive occupations within computer science. To capture AI penetration across CZs, we therefore measure employment in occupations where such expertise is most likely to be applied, following a similar approach to Bonfiglioli et al. (2025). As detailed in Appendix A.1.1, we identify these data-intensive occupations based on the specialized knowledge they demand in advanced data processing and machine learning software. The classification draws on a novel section of the O*NET database, called "Hot Technologies", which records the software tools most frequently cited in all recent US job postings, mapped to occupations under the 2018 Standard Occupational Classification (SOC) system. This procedure yields 13 distinct job titles, listed in Table 1 alongside their corresponding 6-digit SOC codes. We view these as the key data-intensive and AI-related occupations, as they are most directly associated with intensive data use in the context of AI. As shown in Appendix B.3, our evidence holds under alternative definitions.[5]

We merge the list of data-intensive occupations with micro-level data from the 2000 US Census and the American Community Survey (ACS), which provide information on each worker's SOC occupation. Using these data, we compute employment in data-intensive occupations in each CZ and year. This measure captures employment both at firms that adopt AI technologies in their operations and within the infrastructure that supports AI, such as data centers. As a result, it provides the basis

[5] Bonfiglioli et al. (2025) consider a broader range of AI-related software and consequently identify a larger set of AI-related occupations. We refine their methodology to identify the subset of occupations with high data intensity. In untabulated results, we also constructed an alternative definition of data-intensive occupations based on a list of specialized software compiled by computer scientists, rather than by GPT-5 as in our main measure. The results remain unchanged, suggesting that our classification is consistent with the broader expert consensus.

for a comprehensive measurement of AI penetration within CZs. To isolate the specific contribution of the AI support infrastructure, we will further integrate this measure with novel and highly granular data on the location and characteristics of data centers, which are introduced in the next section.

3.1.2 Data Centers

Data centers form the backbone of AI infrastructure, supplying the storage capacity and computational power needed to operate AI systems at scale. These activities are normally argued to involve significant energy consumption. We obtain the names and addresses of all data centers currently active in the US by scraping the website of *Datacenters.com*, a global technology marketplace and directory that enables businesses to find, compare and procure data center infrastructure. The platform aggregates listings from hundreds of providers, covering more than 6,300 facilities across 108 countries. The total number of US data centers is 2,194. Using each data center's address, we apply forward geocoding to retrieve its geographic coordinates (latitude and longitude). These coordinates are then combined with the shapefile of US CZs to assign each data center to a specific CZ. For a subset of 1,445 data centers, *Datacenters.com* also provides information on their physical size (floor space, measured in million square feet).

3.1.3 Carbon Dioxide and Other Air Pollutants

The main air pollutant we consider is carbon dioxide (CO_2). CO_2 emissions are the result of complete fuel combustion, so their concentration is closely linked to energy consumption, especially in electricity systems that rely heavily on fossil fuels. AI can affect energy demand both directly—through the computational power required to run these technologies and operate data centers—and indirectly, by inducing changes in scale, composition and production techniques.

Our primary source of information on CO_2 emissions is the Vulcan dataset (Gurney et al., 2009, 2025). It provides estimates of CO_2 emissions from fossil fuel combustion and cement production across the US in 2002 (version 2.2) and over 2010-2021 (version 4.0). The dataset is constructed using a wide range of federal, state and local data sources, including pollution inventories, energy statistics and infrastructure maps. These inputs are combined to estimate the magnitude of CO_2 emissions and to allocate them across space and time, taking into account the characteristics of combustion technologies, infrastructure and human activity patterns. The annual emissions data are reported at a high spatial resolution, using a 1 km $\times$ 1 km grid. In addition to the gridded data, the dataset also includes aggregated CO_2 emission totals at the county level. We use the county-level figures, together with a crosswalk between counties and CZs (Autor and Dorn, 2013), to compute total CO_2 emissions

for each CZ and year.

A distinctive feature of the Vulcan dataset is its detailed estimation of CO_2 emissions across a broad range of economic sectors and disaggregated by both fuel type and source category. The dataset includes emissions from the following sectors: onroad transportation, nonroad mobile sources, electricity production, residential buildings, commercial buildings, industrial facilities, airports, railroads, commercial marine vessels and cement production. For each sector, fuel use is classified into categories such as coal, natural gas and various petroleum products. Emissions are estimated using detailed activity data, combustion technologies and fuel characteristics. Sources are categorized into point sources, such as power plants and industrial facilities, and non-point sources, which refer to spatially diffuse activities including residential heating and commercial energy use. In our analysis, we aggregate emissions into the following sectoral groups: electric, business, residential and transportation sector. For each aggregated sector, we include total emissions from both point and non-point sources, and across all relevant fuel types. Emissions are expressed in million metric tonnes of carbon.

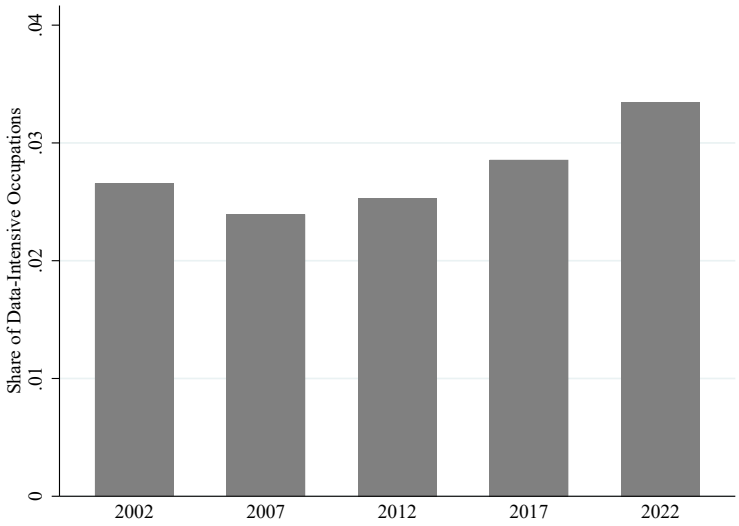
We complement the analysis of CO_2 emissions by examining additional air pollutants: carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2) and particulate matter ($PM_{1.0}$ and $PM_{2.5}$). These pollutants are primarily generated as byproducts of incomplete combustion, high-temperature combustion processes or from impurities in fossil fuels. The analysis of these pollutants offers complementary insights into air quality outcomes typically more localized than CO_2. We use annual satellite-derived data on $PM_{2.5}$ from the Atmospheric Composition Analysis Group over 2002-2022 (V5.NA.04.02 dataset). Data on CO, SO_2, NO_2 and $PM_{1.0}$ are obtained from the gridded Environmental Impacts Frame over 2002-2020. Both datasets provide pollutant concentration estimates at high spatial resolution, based on 0.01-degree grid cells. We combine these data with the shapefile of US CZs to compute annual average concentration levels of each pollutant for every CZ.

3.2 Stylized facts

Figure 1 illustrates the evolution of the employment share of data-intensive occupations in the US over the sample period. This share remained relatively stable at approximately 2.7% throughout the early 2000s. Starting in 2007, however, it began to rise steadily, with a notable acceleration during the 2010s. In 2022, data-intensive occupations constituted about 3.3% of total US employment. This trend aligns with existing evidence indicating that the use of AI was limited in the early 2000s but has significantly intensified over the past decade (e.g., Taddy, 2018; Alekseeva et al., 2021; Bonfiglioli et al., 2025).

The aggregate figures conceal substantial heterogeneity across industries and regions. Leveraging

Figure 1: Employment Share of Data-Intensive Occupations in the US



Notes: The figure shows the employment share of data-intensive occupations in the US between 2002 and 2022. The list of data-intensive occupations is reported in Table 1.

information on workers' industries from the US Census and the ACS, we calculate the employment share of data-intensive occupations across 210 industries, using the 1990 Census Bureau industrial classification. Table 2 presents the ten industries with the highest and lowest shares of employment in data-intensive occupations in 2022. Consistent with Bonfiglioli et al. (2025), the top industries are concentrated in advanced services, including computer and data processing, telecommunications and audiovisual services, R&D and financial services. They also encompass high-tech manufacturing sectors, such as those producing computer equipment, communication devices, audiovisual systems, aerospace products and scientific instruments. In contrast, the bottom industries are predominantly traditional service activities, such as retail, logging, agriculture, food and personal services.

Because industries are not uniformly distributed over space, the employment share of data-intensive occupations also varies significantly across CZs. As shown in the upper panel of Figure 2, this share is particularly high in CZs along the West Coast, in CZs encompassing major cities on the East Coast and in the Great Lakes, and in parts of the South-Central US. The share is also high in a number of CZs in northern states, including Utah, Colorado and Montana, reflecting the emergence of new high-tech hubs around cities like Salt Lake City, Boulder and Bozeman. High shares are also observed in CZs that host federal government facilities linked to the defense and aerospace sectors. On the contrary, the employment share of data-intensive occupations is relatively low in most of the Midwest.

Table 2: Top and Bottom Industries by Employment Share of Data-Intensive Occupations

Top 10 Industries	DI Share	Bottom 10 Industries	DI Share
Computer and Data Processing Services	0.4601	Eating and Drinking Places	0.0015
Computer Related Equipment	0.1771	Services to Dwellings and Other Buildings	0.0015
Telephone Communications	0.1640	Logging	0.0014
Radio, TV, Broadcasting and Cable	0.1607	Child Day Care Services	0.0013
Guided Missiles, Space Vehicles and Parts	0.1245	Retail Bakeries	0.0009
Radio, TV and Communication Equipment	0.1239	Agricultural Production, Livestock	0.0008
Research & Development, and Testing	0.0932	Landscape and Horticultural Services	0.0006
Credit Agencies	0.0922	Beauty Shops	0.0004
Security, Commodity Brokerage and Investment	0.0917	Private Households	0.0002
Scientific and Controlling Instruments	0.0899	Barber Shops	0.0000

Notes: DI share denotes the percentage of industry employment accounted for by data-intensive occupations in 2022. Industries are classified according to the US Census industrial classification.

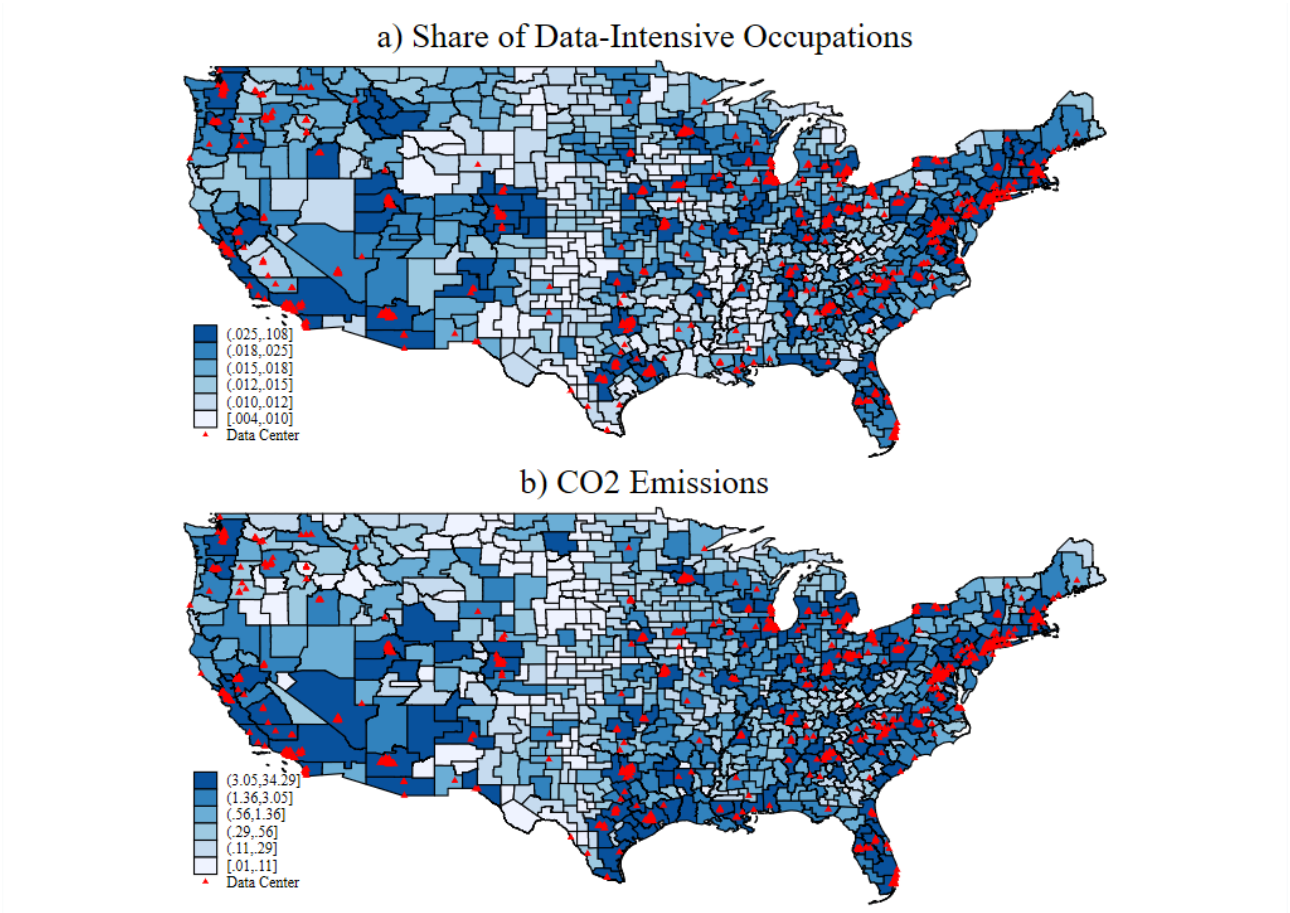
The map also indicates the location of data centers, represented by red triangles.[6] Data centers are not evenly distributed across the country but are concentrated in a relatively small number of CZs (169); the average number of data centers in these CZs is 13 (see Appendix Table A.4 for details). There is a notable degree of spatial correlation between the presence of data centers and the employment share of data-intensive occupations. A regression of this share on a dummy for the presence of at least one data center in a CZ yields a coefficient of 0.014 (s.e. 0.001). As expected, however, the overlap between the two distributions is not perfect: several CZs with high concentrations of employment in data-intensive occupations do not host any data centers. This underscores the multifaceted array of factors that shape the geographic distribution of data centers.

Turning to CO_2 emissions, Appendix Figure A.1 shows the well-known declining trend that occurred at the US level over the period 2002–2022, reflecting the sustained efforts towards the green transition. The largest share of emissions originates from the electricity sector (approximately 35%), indicating that electricity generation remains the single most significant contributor. The transportation sector follows closely, accounting for around 34% of total emissions. In contrast, emissions from the business and residential sectors are considerably smaller, representing 20% and 10% of the national total, respectively.

Similar to employment in data-intensive occupations, CO_2 emissions are not evenly distributed over space. As shown in the bottom panel of Figure 2, CO_2 concentration levels are particularly high along both coasts, in the South-Central US and in the Great Lakes region, while they are relatively lower across much of the interior of the country. Notably, the spatial distributions of CO_2 emissions

[6]Each triangle corresponds to a data center site. In some cases, a single site may host multiple data centers.

Figure 2: Data-Intensive Occupations, Data Centers and CO_2 Emissions in US Commuting Zones



Notes: Panel a) displays the employment share of data-intensive occupations in each CZ in 2022, while panel b) shows the total CO_2 emissions in each CZ for the same year. Red triangles indicate the presence of a data center site.

and employment in data-intensive occupations are positively correlated, with a correlation coefficient of 0.51 across CZs. CO_2 emissions also tend to be systematically higher in CZs that host at least one data center. A regression of CO_2 emissions on a data center dummy yields a coefficient of 4.105 (s.e. 0.419). These patterns point to a positive association between AI-related activity and CO_2 emissions across US localities. In the following sections, we systematically study this relationship and explore the underlying mechanisms.

4 Econometric Framework

In this section, we illustrate the empirical framework and our identification strategy.

4.1 Regression Equation

Our main specification takes the following form:

$$\Delta E_{ct} = \alpha_c + \alpha_{st} + \beta\, AIpen_{ct} + \mathbf{X}'_{ct}\gamma + \varepsilon_{ct}, \tag{1}$$

where c and t denote CZs and time periods, respectively, and ε_{ct} is an error term. We estimate eq. (1) using stacked first differences corresponding to four five-year periods over 2002-2022. ΔE_{ct} represents the change in emissions of a given air pollutant in CZ c over period t, while $AIpen_{ct}$ measures the penetration of AI in the same CZ and time period.

$AIpen_{ct}$ is defined as follows:

$$AIpen_{ct} = \frac{L^{DI}_{c\tau_1} - L^{DI}_{c\tau_0}}{L_{c0}}, \tag{2}$$

where $L^{DI}_{c\tau_0}$ and $L^{DI}_{c\tau_1}$ denote employment in data-intensive occupations in the first year (τ_0) and last year (τ_1) of period t, while L_{c0} is the total employment (across all occupations) in CZ c in 2000. Hence, for each CZ, $AIpen_{ct}$ measures the quinquennial change in the relative importance of data-intensive occupations relative to initial total employment. Given the definition of $AIpen_{ct}$, eq. (1) is a changes-on-changes regression. The model estimates the overall relationship between AI penetration and changes in emissions. In the subsequent analysis, we will augment eq. (1) to study the underlying mechanisms and disentangle the contribution of data centers from AI adoption.

Eq. (1) includes CZ fixed effects (α_c), state×year fixed effects (α_{st}) and a host of covariates, collected in the vector $\mathbf{X}_{ct}$. These controls fall into two broad categories. First, $\mathbf{X}_{ct}$ includes interactions

between year dummies and CZ-level characteristics. Specifically: (i) the initial size and economic structure of each CZ, proxied by population and the share of manufacturing employment in 2000, respectively; (ii) initial land composition, proxied by six dummy variables corresponding to distinct land cover categories in 2000; and (iii) topographical characteristics, proxied by the difference between the maximum and minimum elevation within the CZ, as well as by the standard deviation of elevation. Second, the vector $\mathbf{X}_{ct}$ includes proxies for two types of shocks that may have occurred in a given CZ over a period: (i) weather-related shocks, proxied by five-year changes in specific humidity, precipitation, average temperature, temperature excursion and wind speed; and (ii) prominent labor market shocks, i.e., changes in import competition from China (Autor et al., 2013) and exposure to industrial robots (Acemoglu and Restrepo, 2020).[7]

The coefficient of interest is β. The inclusion of CZ fixed effects ensures that this coefficient is identified from within-CZ variation over time. The state$\times$year fixed effects absorb period-specific shocks that may simultaneously affect all CZs within a state. The covariates $\mathbf{X}_{ct}$ account instead for heterogeneous trends across CZs with different initial characteristics, as well as for contemporaneous weather- and labor market-related shocks that may influence both emissions and AI penetration.

Two considerations are worth making at this point. First, our empirical approach identifies the local effect of AI and is not designed to capture general equilibrium effects at the national level, as it leverages variation across CZs. Second, although the inclusion of a rich set of controls and fixed effects addresses many potential confounders, the OLS estimate of β should not be interpreted as causal, because unobservable factors may still drive both AI penetration and emissions. We now turn to a discussion of the key identification challenges and present our empirical strategy for estimating causal effects.

4.2 Identification Strategy

Variation in AI penetration may be influenced by CZ-specific unobservables that also affect emissions. In particular, stricter environmental regulations and growing consumer preference for cleaner products have led firms in some localities to adopt newer production technologies, which are both more AI-intensive and less polluting. These demand shocks tend to induce a spurious negative correlation between $AIpen_{ct}$ and ΔE_{ct}, biasing downward the effect of AI estimated with OLS.[8]

To address this issue, we employ an instrumental variable designed to isolate variation in $AIpen_{ct}$ that is not driven by local demand shocks. Following Bonfiglioli et al. (2025), we construct a shift-

[7] See Appendix A for a detailed description of these variables and data sources.

[8] Such technologies include next-generation industrial machinery that employs AI for process control and monitoring, while achieving lower CO_2 emissions through improved energy efficiency.

share (Bartik) instrument that captures the differential exposure of CZs to aggregate AI developments over time. This is obtained by combining national industry-level shifts in AI penetration over time with the initial (pre-AI) industrial composition of each CZ.[9] The instrument is constructed as follows:

$$AIexp_{ct} = \sum_{i} \omega_{ci0} \times \left(\frac{L^{DI}_{i\tau_1} - L^{DI}_{i\tau_0}}{L_{i0}} \right), \tag{3}$$

where $L^{DI}_{i\tau_1} - L^{DI}_{i\tau_0}$ is the change in national employment of data-intensive occupations within industry i between the first year (τ_0) and the last year (τ_1) of period t, and L_{i0} is the total national employment in industry i in 2000. The term $\omega_{ci0} \equiv \frac{L_{ci,1990}}{L_{c,1990}}$ is the share of industry i in the total employment of CZ c in 1990.

The intuition behind this instrument is as follows. As technological advances reduce the cost of AI and enhance its capabilities, the technology spreads more broadly, especially in industries whose tasks are more compatible with its application. CZs are differentially exposed to these industry-level AI shifts, due to historical differences in their industrial composition, as captured by the employment shares ω_{ci0}. We measure these shares in 1990, i.e., more than a decade prior to the start of the sample period. Because AI technologies were largely nonexistent at that time, using pre-sample employment shares mitigates concerns that the industrial structure of a CZ may have been endogenously influenced by expectations of future AI developments. Moreover, holding these shares fixed over time helps avoid introducing endogenous or serially correlated variation in $AIexp_{ct}$ in the context of our stacked first-differences specification.

Using micro-level data from the US Census and the ACS, we compute industry-level shifts and local employment shares spanning 210 industries that encompass all sectors of the economy. Our instrument, therefore, exploits significantly larger cross-industry variation than the typical Bartik measures used in the recent automation literature. For instance, proxies for exposure to industrial robots based on data from the International Federation of Robotics (IFR) aggregate industry-level shifts for less than twenty broad sectors.

The identifying assumption underlying our approach is that the industry-level shifts are orthogonal to shocks occurring within individual CZs. We believe this to be a reasonable assumption given that most CZs are small relative to the overall US economy, and that our specification includes an extensive set of fixed effects and covariates. Nonetheless, our identification strategy could still be threatened in two cases. First, if some contemporaneous shocks remained that correlate with the outcome ΔE_{ct}

[9] This approach builds on a long tradition of shift-share instruments, initiated by Bartik (1991) and Blanchard and Katz (1992), and subsequently applied to contexts such as the impact of Chinese import competition (e.g., Autor et al., 2013) and industrial robot adoption (e.g., Acemoglu and Restrepo, 2020).

and the instrument $AIexp_{ct}$. Second, if some remaining trends at the CZ level influenced emissions independently of AI. In Section 6, we employ several approaches to account for these identification threats and find that they are unlikely to be driving the results.

5 Baseline Results

In this section, we present the baseline estimates. We begin by examining the impact of AI penetration on overall emissions. Next, we explore the role played by the three components identified in the decomposition framework introduced in Section 2.

5.1 AI Penetration and Emissions

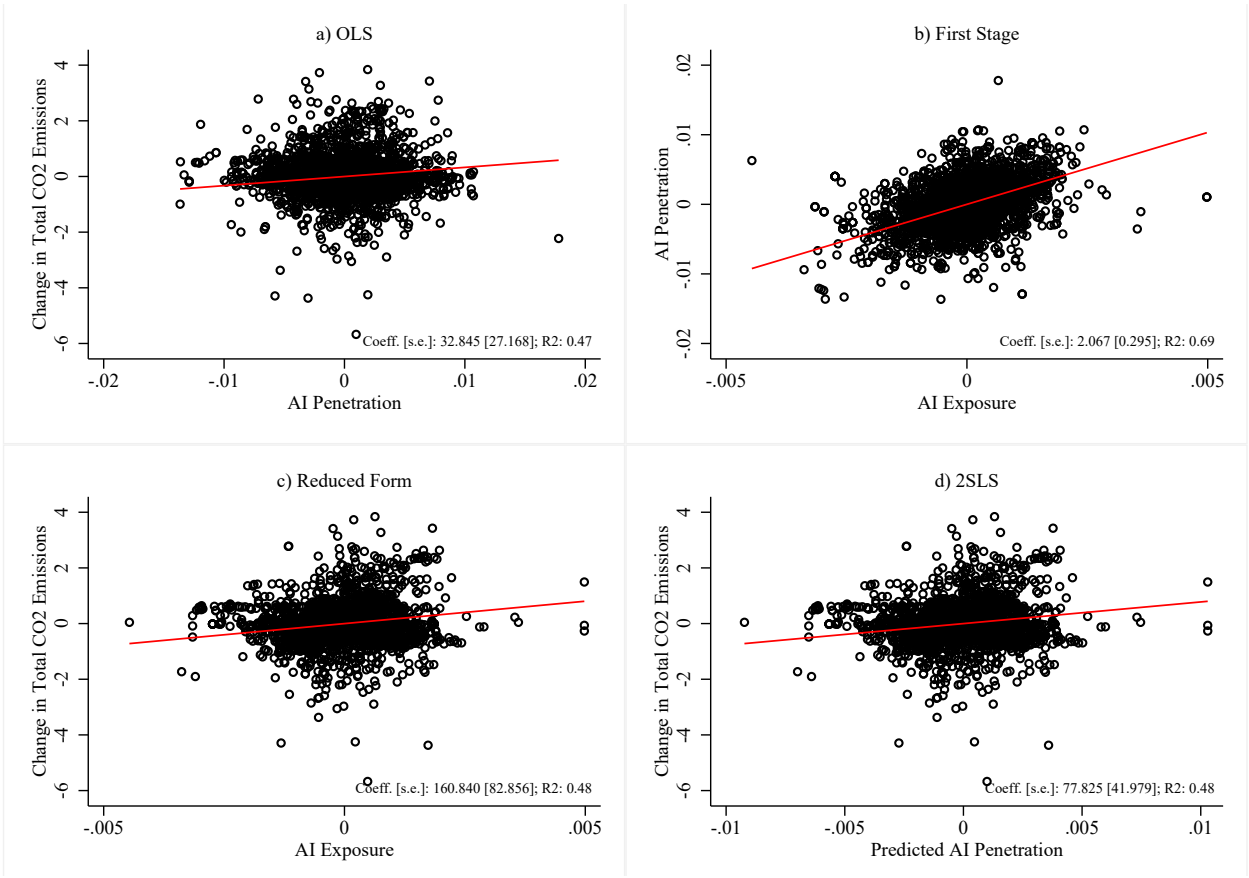
Figure 3 presents results from a parsimonious version of eq. (1). The dependent variable, ΔE_{ct}, is the change in total CO_2 emissions and the specification includes CZ and state$\times$year fixed effects only. Observations are weighted by the initial-period share of each CZ in total US population, and standard errors are clustered at the state-year level to allow for correlation in the residuals across all CZs within the same state in a given year. Each hollow circle corresponds to a CZ-period pair, for a total of 2,888 observations.

Plot a) displays the OLS regression of ΔE_{ct} on $AIpen_{ct}$, showing that CO_2 emissions and AI penetration are positively correlated in our sample. Plot b) plots the first-stage regression of $AIpen_{ct}$ on $AIexp_{ct}$. The instrument has strong predictive power in explaining AI penetration.[10] Plot c) illustrates a similarly strong and positive reduced-form relationship between ΔE_{ct} and $AIexp_{ct}$, suggesting that CO_2 emissions grow relatively faster in CZs with higher AI exposure. Finally, plot d) illustrates the relationship between ΔE_{ct} and $\widehat{AIpen}_{ct}$, the fitted value of AI penetration from the first-stage regression. The association between the two variables is positive and precisely estimated, suggesting that variation in AI penetration, driven by variation in exposure to AI, has a positive effect on CO_2 emissions. Interestingly, the 2SLS relationship in plot d) is stronger than the OLS relationship in plot a), consistent with demand shocks inducing a downward bias in the OLS estimates.

Table 3, column (1), reproduces the first-stage and 2SLS coefficients corresponding to plots b) and d) of Figure 3. The subsequent columns progressively enrich the specification by incorporating additional control variables. Column (2) adds controls for initial size and sectoral composition of the CZs interacted with period dummies. Columns (3) and (4) introduce labor market and weather

[10] The Kleibergen-Paap $F-$statistic, equal to 49, safely exceeds the value of 10 normally considered as a rule-of-thumb threshold for instrument relevance.

Figure 3: AI Penetration, AI Exposure and CO_2 Emissions in US Commuting Zones



Notes: The sample consists of 722 commuting zones observed over four five-year periods. In each plot, an observation is a CZ-period pair. *AI Penetration* and *AI Exposure* are defined in eq. (2) and (3), respectively. *Predicted AI Penetration* is the fitted value of *AI Penetration* from the first-stage regression in plot b). All variables are in deviations from CZ and state×year fixed effects.

shocks, respectively. Column (5) includes the interactions of period dummies with initial land use and topographical features of each CZ. Across all specifications, the results confirm that AI penetration has a positive and statistically significant effect on CO_2 emissions.

The table also presents standardized β-coefficients, which express the effect of AI penetration in units of standard deviation (s.d.). In the preferred specification (column 5), a one s.d. higher $AIpen_{ct}$ is associated with a 0.73 s.d. higher ΔE_{ct}. To illustrate the magnitude of this effect, consider the difference in $AIpen_{ct}$ between the CZ of Kansas City and that of New Orleans, roughly at the 75th and 25th percentiles, respectively, of average $AIpen_{ct}$ over the sample period. This difference accounts for approximately 41% of the observed difference in CO_2 emissions growth between the

Table 3: AI Penetration and CO_2 Emissions

	(1)	(2)	(3)	(4)	(5)
2nd Stage Regression					
$AIpen_{ct}$	77.825*	116.381**	112.643**	118.043***	132.279***
	(41.979)	(45.939)	(44.669)	(42.888)	(44.827)
1st Stage Regression					
$AIexp_{ct}$	2.067***	2.268***	2.106***	2.011***	1.969***
	(0.295)	(0.383)	(0.329)	(0.300)	(0.271)
Kleibergen-Paap $F-$Statistic	49.04	34.98	41.04	44.81	52.94
$\beta-$Coefficients					
$AIpen_{ct}$	0.43	0.64	0.62	0.65	0.73
CZ FE	Yes	Yes	Yes	Yes	Yes
State$\times$Year FE	Yes	Yes	Yes	Yes	Yes
Size and Economic Structure	No	Yes	Yes	Yes	Yes
Labor Market Shocks	No	No	Yes	Yes	Yes
Weather Shocks	No	No	No	Yes	Yes
Topography and Land Use	No	No	No	No	Yes
Obs.	2,888	2,888	2,888	2,888	2,888

Notes: The sample consists of 722 CZs observed over four five-year periods. The dependent variable is the change in CO_2 emissions. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects and state$\times$year fixed effects. Column (2) adds period dummies interacted with each CZ's initial population and initial manufacturing share. Column (3) further includes controls for exposure to Chinese import competition and industrial robots. Column (4) additionally controls for changes in climate variables, including specific humidity, wind speed, average temperature, temperature excursion and total precipitation. Column (5) adds period dummies interacted with the standard deviation of elevation, the difference between the lowest and highest elevation and six land use dummies. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

two CZs during the sample period. Moreover, the estimates imply that, had the CZ with the average $AIpen_{ct}$ value (0.002) experienced no AI penetration, its CO_2 emissions would have experienced a 37% greater decline; that is, they would have fallen by 34.3% compared to the average fall of 25% observed in the data. While these figures should not be interpreted as counterfactual exercises, since nation-wide effects are differenced out in our empirical strategy, they nonetheless suggest that AI penetration has significantly contributed to slowing down the reduction of CO_2 emissions in the US over the past two decades.

Table 4 extends the analysis by examining the impact of AI penetration on emissions of other air pollutants. In each column, eq. (1) is estimated using a different dependent variable, namely,

Table 4: AI Penetration and Emissions of Other Air Pollutants

	CO (1)	SO_2 (2)	NO_2 (3)	$PM_{1.0}$ (4)	$PM_{2.5}$ (5)	Pollution Index (6)
2nd Stage Regression						
$AIpen_{ct}$	2.397**	5.503	8.699	89.635	65.169***	19.033***
	(0.973)	(6.933)	(13.477)	(65.903)	(15.815)	(4.651)
1st Stage Regression						
$AIexp_{ct}$	1.969***	1.969***	1.969***	1.969***	1.969***	1.969***
	(0.271)	(0.271)	(0.271)	(0.271)	(0.271)	(0.271)
Kleibergen-Paap $F-$Statistic	52.94	52.94	52.94	52.94	52.94	52.94
$\beta-$Coefficients						
$AIpen_{ct}$	0.18	0.04	0.03	0.10	0.17	0.20
Obs.	2,886	2,886	2,886	2,886	2,888	2,886

Notes: The sample consists of 722 CZs observed over four five-year periods. The dependent variable is the change in emissions of the air pollutants listed in the column headers. The pollution index used in column (6) is the first principal component of the air pollutants in columns (1)-(5). The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state×year fixed effects and the control variables used in column (5) of Table 3. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

the change in emissions of the air pollutant indicated in the column header. Column (6) presents results using an aggregate "pollution index", constructed as the first principal component of the five individual pollutants considered in columns (1)-(5). The coefficient on $AIpen_{ct}$ is positive across all specifications, statistically significant for CO, $PM_{2.5}$ and the aggregate pollution index, and only marginally insignificant for $PM_{1.0}$. These results indicate that the environmental impact of AI penetration is not limited to CO_2 but extends to other air pollutants, especially those more closely related to energy consumption. The magnitude of these effects, however, is considerably smaller than for CO_2 emissions, as evidenced by the standardized β-coefficients reported in Table 4. This suggests that, while AI penetration contributes to broader environmental impacts, CO_2 emissions remain the primary channel through which it currently affects the environment.[11]

A key strength of our data is that they contain information not only on total CO_2 emissions but also on CO_2 emissions from all relevant sources ("sectors"). In Table 5, we study how the effects of AI penetration vary across sectors. Column (1) reproduces the baseline estimate for total CO_2 emissions (see column 5 of Table 3). Columns (2)-(5) estimate instead eq. (1) using sector-specific changes

[11]The number of observations is reduced to 2,886 because of missing data on pollutants in the gridded Environmental Impacts Frame for CZ 20402 in 2017 and 2022.

Table 5: AI Penetration and CO_2 Emissions by Sector

	Total (1)	Electric (2)	Business (3)	Residential (4)	Transportation (5)
2nd Stage Regression					
$AIpen_{ct}$	132.279***	55.677**	75.092*	-30.074	31.583***
	(44.827)	(23.667)	(39.709)	(35.842)	(8.237)
1st Stage Regression					
$AIexp_{ct}$	1.969***	1.969***	1.969***	1.969***	1.969***
	(0.271)	(0.271)	(0.271)	(0.271)	(0.271)
Kleibergen-Paap $F-$Statistic	52.94	52.94	52.94	52.94	52.94
$\beta-$Coefficients					
$AIpen_{ct}$	0.73	0.37	0.91	-0.67	0.83
Obs.	2,888	2,888	2,888	2,888	2,888

Notes: The sample consists of 722 CZs observed over four five-year periods. The dependent variable is the change in CO_2 emissions by a given sector, as indicated in the column headers. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state$\times$year fixed effects and the control variables used in column (5) of Table 3. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

in CO_2 emissions as the dependent variable. The coefficient on $AIpen_{ct}$ is positive and statistically significant in all cases, except for column (4). These results indicate that AI penetration is associated with increased CO_2 emissions not only in total but also across a broad range of sectors, the only exception being emissions from residential buildings.

5.2 Decomposition

As outlined in Section 2, the positive effect of AI penetration on emissions can be decomposed into three margins: the scale effect, the composition effect and the technique effect. In what follows, we examine the role played by each of these margins.

5.2.1 Scale Effect

AI penetration has the potential to stimulate economic activity, drawing more firms and workers into a given CZ. The resulting expansion in the scale of the local economy, in turn, could lead to an increase in emissions. To examine the role played by scale, in column (1) of Table 6, we start by studying how AI penetration affects the size of a CZ. To this purpose, we estimate eq. (1) using the log change in

Table 6: AI Penetration, CZ Size and Per-Capita CO_2 Emissions

	Population	CO_2 Emissions				
		Total	Electric	Business	Residential	Transportation
	(1)	(2)	(3)	(4)	(5)	(6)
2nd Stage Regression						
$AIpen_{ct}$	4.114***	0.020	0.123***	-0.028	-0.071*	-0.004
	(0.684)	(0.059)	(0.044)	(0.034)	(0.040)	(0.007)
1st Stage Regression						
$AIexp_{ct}$	1.969***	1.969***	1.969***	1.969***	1.969***	1.969***
	(0.271)	(0.271)	(0.271)	(0.271)	(0.271)	(0.271)
Kleibergen-Paap F-Statistic	52.94	52.94	52.94	52.94	52.94	52.94
β-Coefficients						
$AIpen_{ct}$	0.30	0.01	0.05	-0.04	-0.14	-0.02
Obs.	2,888	2,888	2,888	2,888	2,888	2,888

Notes: The sample consists of 722 CZs observed over four five-year periods. In column (1), the dependent variable is the log change in population. In columns (2)-(6), it is the change in CO_2 emissions from a given sector (indicated in the column headers) per 1,000 individuals aged 16 and older. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state×year fixed effects and the control variables used in column (5) of Table 3. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

CZ population as the dependent variable. The coefficient on $AIpen_{ct}$ is positive, sizable and precisely estimated. This result confirms that AI penetration does lead to a significant expansion in CZ size.

Next, we investigate the extent to which the scale of the local economy mediates the impact of AI penetration on emissions. To this purpose, we estimate eq. (1) using changes in CO_2 emissions per 1,000 inhabitants as the dependent variables. If the scale effect was the only channel at play, emissions would rise proportionally with CZ size, and AI penetration would have no effect on per-capita emissions. Consistent with this, the coefficient on $AIpen_{ct}$ is small and statistically insignificant in all but one of the specifications reported in columns (2)-(6). The only exception is column (3), where the dependent variable is the change in per-capita CO_2 emissions from electricity generation: in this case, the coefficient on $AIpen_{ct}$ is positive and very precisely estimated.

These findings suggest that the scale effect plays a central role, fully accounting for the impact of AI penetration both on total CO_2 emissions and on CO_2 emissions from the business and transportation sectors. However, the evidence also points to the presence of additional mechanisms. In particular, the increase in CO_2 emissions from electricity generation cannot be fully attributed to the

scale effect, indicating that other channels are at play.

5.2.2 Composition Effect

The second margin through which AI penetration may influence emissions is by altering the economic composition of a CZ. If AI penetration leads to a shift toward less emission-intensive industries, the composition effect would contribute to a reduction in emissions. Conversely, if the shift favors more emission-intensive industries, the composition effect would drive emissions upward. Furthermore, because emission-intensive industries tend to rely heavily on electricity, such a reallocation could also help explain why AI penetration increases CO_2 emissions from electricity generation, even after accounting for changes in the overall scale of economic activity. Although composition effects are generally found to play a modest role in driving aggregate CO_2 emissions at the national level, we now investigate their relevance at the local level by analyzing how AI penetration reshapes the economic structure of individual CZs across industries with varying energy intensities.

We start by computing the average energy intensity of each industry over the sample period, $\overline{EI}_i$, defined as the average ratio of total energy spending in total output by industry i. We use data from the Production Account Tables of the US Bureau of Economic Analysis (BEA), covering 61 industries—19 in manufacturing and 42 in services and primary sectors. Next, we construct the following variable:

$$\Delta EI_{ct} = \sum_i \overline{EI}_i \times \Delta \left(\frac{L_{cit}}{L_{ct}} \right), \tag{4}$$

where $\Delta \left(\frac{L_{cit}}{L_{ct}} \right)$ denotes the change in the employment share of industry i within CZ c over period t. ΔEI_{ct} captures the change in the average energy intensity of production in CZ c over period t that is solely attributable to shifts in industrial composition, holding each industry's energy intensity fixed. We then estimate eq. (1) using ΔEI_{ct} as the dependent variable.

The results are presented in Table 7, column (1). The coefficient on $AIpen_{ct}$ is negative but statistically insignificant. AI penetration thus seems to exert no impact, or at most a modest negative effect, on the average energy intensity of a CZ. A possible reason for this result is that various industries with high rates of AI adoption are in the services sector, which is inherently less energy-intensive than manufacturing. To account for this, we reconstruct ΔEI_{ct} separately for manufacturing and non-manufacturing industries. In columns (2) and (3) of Table 7, we estimate eq. (1) using the sector-specific variables—ΔEI^M_{ct} for manufacturing and ΔEI^{NM}_{ct} for non-manufacturing industries—in place of the aggregate measure ΔEI_{ct}. For manufacturing, the coefficient on $AIpen_{ct}$ is positive, but it is small in magnitude and statistically insignificant. For non-manufacturing, the

Table 7: Composition Effect

	Average Energy Intensity			Employment Share of Energy-Intensive Industries		
	All Industries (1)	Manufacturing (2)	Non-Manufacturing (3)	All Industries (4)	Manufacturing (5)	Non-Manufacturing (6)
2nd Stage Regression						
$AIpen_{ct}$	-0.013	0.029	-0.011	-0.029***	0.028	-0.035***
	(0.009)	(0.018)	(0.009)	(0.007)	(0.079)	(0.009)
1st Stage Regression						
$AIexp_{ct}$	1.969***	1.969***	1.969***	1.969***	1.969***	1.969***
	(0.271)	(0.271)	(0.271)	(0.271)	(0.271)	(0.271)
Kleibergen-Paap F-Statistic	52.94	52.94	52.94	52.94	52.94	52.94
Obs.	2,888	2,888	2,888	2,888	2,888	2,888

Notes: The sample consists of 722 CZs observed over four five-year periods. In columns (1)–(3), the dependent variable is the change in average energy intensity, constructed as in eq. (4). In columns (4)–(6), the dependent variable is the change in the employment share of energy-intensive industries, defined as those with $\overline{EI}_i$ above the sample median. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state×year fixed effects and the control variables used in column (5) of Table 3. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

coefficient is negative and imprecisely estimated.

In columns (4)–(6), we repeat the previous regressions using the change in the employment share of high energy-intensive industries as the dependent variable. These industries are defined as those with $\overline{EI}_i$ above the sample median. The results now reveal a negative and statistically significant effect of $AIpen_{ct}$, primarily driven by non-manufacturing industries. Overall, these results suggest that the composition effect plays a limited role in mediating the impact of AI penetration on emissions. If anything, the changes in industrial composition induced by AI penetration are more consistent with a reduction in CO_2 emissions than with an increase, suggesting that these compositional shifts may slightly attenuate the overall effect of AI penetration.

5.3 Technique Effect

The third margin through which AI penetration may affect emissions is the technique effect, which captures changes in the emission intensity of production within individual industries. Besides being interesting in its own right, this mechanism could also help explain the effect of AI penetration on per-capita CO_2 emissions from electricity generation, assuming AI led firms to adopt more energy-intensive production methods.

To investigate the technique effect, we construct the following Bartik measure for changes in

industry-level energy intensity:

$$EIexp_{ct} = \sum_i \omega_{ci0} \times \Delta \log EI_{it}, \tag{5}$$

where $\Delta \log EI_{it}$ is the log change in the energy intensity of industry i over period t, and ω_{ci0} is the share of industry i in the total employment of CZ c in 2000. $EIexp_{ct}$ captures the idea that a given increase in the energy intensity of an industry is felt more strongly in CZs that have historically relied more heavily on it. Since the employment shares ω_{ci0} are held fixed over time, $EIexp_{ct}$ isolates the effect of within-industry changes in energy intensity, independent of variation in industrial composition.

In Table 8, we estimate eq. (1) including $EIexp_{ct}$ as an additional explanatory variable. From this point onward, we focus on our primary outcomes of interest—total CO_2 emissions and CO_2 emissions from electricity generation, both in absolute and per-capita terms—as they summarize the main results obtained thus far. The coefficient on $EIexp_{ct}$ is highly imprecisely estimated, and the inclusion of this variable leaves the coefficients on $AIpen_{ct}$ close to the baseline estimates.

In the subsequent columns, we use alternative Bartik measures for changes in other industry-level characteristics, namely, log employment ($Lexp_{ct}$), gross output per worker ($YLexp_{ct}$) and capital stock per worker ($KLexp_{ct}$).[12] These variables capture the possibility that increases in industry size, productivity or capital intensity may be associated with the use of techniques characterized by different energy intensities, even when this is not immediately reflected in energy expenditures. The coefficients on these variables are generally estimated with little precision and are not robust. Moreover, the coefficient on $AIpen_{ct}$ remains largely unchanged across specifications. These results provide no evidence that the effect of AI penetration on CO_2 emissions could be driven by within-industry changes in production techniques.

6 Robustness Checks and Threats to Identification

In this section, we present an extensive sensitivity analysis to assess the robustness of the baseline results. We then discuss the main threats to identification.

[12] See Appendix A.2.5 for data sources and variables definitions.

Table 8: Technique Effect

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Total (Level)					Electric (Level)				
2nd Stage Regression										
$AIpen_{ct}$	131.816**	171.539***	129.664**	154.746**	211.463***	55.821**	93.386***	55.164**	81.107***	133.781***
	(58.363)	(61.839)	(58.169)	(65.086)	(69.419)	(23.789)	(34.923)	(23.032)	(25.074)	(41.027)
$EIexp_{ct}$	4.013				4.437	-1.247				-1.623
	(4.774)				(4.551)	(1.863)				(2.079)
$Lexp_{ct}$		-19.318*			-23.424*		-18.555**			-22.135**
		(10.751)			(12.331)		(8.435)			(9.582)
$YLexp_{ct}$			-4.948		-23.318**			-0.970		-16.244*
			(5.435)		(10.113)			(4.582)		(8.359)
$KLexp_{ct}$				5.256	10.389**				5.949**	9.711**
				(3.199)	(4.823)				(2.312)	(3.848)
	Total (Per Capita)					Electric (Per Capita)				
2nd Stage Regression										
$AIpen_{ct}$	0.018	0.075	0.026	0.037	0.080	0.123**	0.175**	0.131***	0.160***	0.205**
	(0.065)	(0.084)	(0.065)	(0.074)	(0.099)	(0.049)	(0.071)	(0.047)	(0.052)	(0.082)
$EIexp_{ct}$	0.010				0.008	-0.001				-0.003
	(0.006)				(0.006)	(0.004)				(0.005)
$Lexp_{ct}$		-0.027			-0.022		-0.026			-0.022
		(0.022)			(0.024)		(0.021)			(0.023)
$YLexp_{ct}$			0.013		0.001			0.016*		0.003
			(0.012)		(0.017)			(0.008)		(0.015)
$KLexp_{ct}$				0.004	0.004				0.009**	0.008
				(0.005)	(0.007)				(0.004)	(0.006)
1st Stage Regression										
$AIexp_{ct}$	1.966***	1.867***	2.025***	1.826***	1.572***	1.966***	1.867***	2.025***	1.826***	1.572***
	(0.300)	(0.353)	(0.282)	(0.292)	(0.300)	(0.300)	(0.353)	(0.282)	(0.292)	(0.300)
Kleibergen-Paap $F-$Statistic	42.92	27.99	51.47	39.11	27.43	42.92	27.99	51.47	39.11	27.43
Obs.	2,888	2,888	2,888	2,888	2,888	2,888	2,888	2,888	2,888	2,888

Notes: The sample consists of 722 CZs observed over four five-year periods. The dependent variable is the change in CO_2 emissions, either total or from electricity generation, measured in levels or per 1,000 individuals aged 16 and older, as specified in the column headers. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state×year fixed effects and the control variables used in column (5) of Table 3. The variables $EIexp_{ct}$, $Lexp_{ct}$, $KLexp_{ct}$ and $YLexp_{ct}$ are Bartik measures for changes in log industry energy intensity, employment, gross output per worker and capital intensity (capital-to-labor ratio), respectively (see eq. (A.1)-(A.4)). All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

6.1 Robustness Checks

Appendix B presents a wide range of robustness checks, whose key conclusions are outlined below. In Appendix B.1, we examine the role of outliers and influential observations, as well as the influence of the COVID-19 pandemic. In Appendix B.2, we explore alternative methods for correcting the standard errors, including various adjustments for clustering, the inference approach proposed by Borusyak et al. (2022) for shift-share instruments and the spatial correlation correction method of Conley (1999) and Colella et al. (2023). Finally, in Appendix B.3, we explore alternative model specifications and different constructions of the main variables. Specifically, we: (i) construct $AIpen_{ct}$ and $AIexp_{ct}$ using alternative definitions of data-intensive occupations, focusing only on data scientists or employing the classifications of AI-related jobs proposed by Hanson (2022) and Eloundou et al. (2024); (ii) use alternative measures of CZ size when computing per-capita CO_2 emissions; and (iii) estimate eq. (1) using 10-year differences instead of quinquennial changes. Our main evidence remains robust across all these exercises.

6.2 Threats to Identification

Our identification strategy relies on the assumption that, conditional on the extensive set of fixed effects and covariates included in eq. (1), no unobserved factors remain that are correlated with the instrument and independently affect emissions across CZs. As discussed in Section 4.2, violations of the exclusion restriction may arise from two main types of confounders: (i) differential underlying trends and (ii) contemporaneous shocks that influence pollution independently of AI. We now use alternative approaches to accommodate the possible influence of such confounders and study how the coefficient β is affected.

We begin by dealing with contemporaneous shocks. In Table 9, we consider the possibility that CZs experiencing similar changes in CO_2 emissions or AI penetration may also be subject to similar unobserved shocks, either to technology and demand or related to the business cycle. To account for these shocks, we group CZs into deciles based on the average values of ΔE_{ct} and $AIpen_{ct}$ over the sample period. We then add to eq. (1) interactions between decile dummies and period dummies. These interactions absorb time-varying shocks common to CZs with similar trends in CO_2 emissions or AI penetration. The coefficient β is now identified from the remaining variation across CZs within the same decile-period cell. The results are largely unchanged.

In Table 10, we conduct a set of complementary exercises. Panel a) estimates eq. (1) after excluding CZs in the top 5% of the distribution of $AIpen_{ct}$ over the sample period. The main pattern of results is preserved, suggesting that our findings are not driven by shocks specific to a small number

Table 9: AI Penetration and CO_2 Emissions: CZ-Specific Shocks

	Total (Level) (1)	Electric (Level) (2)	Total (Per Capita) (3)	Electric (Per Capita) (4)
	a) Decile Bins of ΔE			
2nd Stage Regression				
$AIpen_{ct}$	113.377***	52.576**	-0.008	0.083**
	(42.760)	(20.383)	(0.059)	(0.039)
1st Stage Regression				
$AIexp_{ct}$	1.920***	1.942***	1.851***	1.921***
	(0.256)	(0.262)	(0.222)	(0.271)
Kleibergen-Paap $F-$Statistic	56.10	55.00	69.43	50.32
Obs.	2,888	2,888	2,888	2,888
	b) Decile Bins of $AIpen$			
2nd Stage Regression				
$AIpen_{ct}$	198.850**	116.477***	0.072	0.237***
	(79.043)	(38.202)	(0.104)	(0.069)
1st Stage Regression				
$AIexp_{ct}$	1.250***	1.250***	1.250***	1.250***
	(0.221)	(0.221)	(0.221)	(0.221)
Kleibergen-Paap $F-$Statistic	32.04	32.04	32.04	32.04
Obs.	2,888	2,888	2,888	2,888
	c) Decile Bins (ΔE & $AIpen$)			
2nd Stage Regression				
$AIpen_{ct}$	182.308**	105.977***	0.068	0.187***
	(79.739)	(36.239)	(0.111)	(0.064)
1st Stage Regression				
$AIexp_{ct}$	1.216***	1.257***	1.174***	1.253***
	(0.216)	(0.215)	(0.185)	(0.213)
Kleibergen-Paap $F-$Statistic	31.69	34.07	40.47	34.76
Obs.	2,888	2,888	2,888	2,888

Notes: The sample consists of 722 CZs observed over four five-year periods. The dependent variable is the change in CO_2 emissions, either total or from electricity generation, measured in levels or per 1,000 individuals aged 16 and older, as specified in the column headers. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state$\times$year fixed effects and the control variables used in column (5) of Table 3. The specification in panel a) controls for interactions between period dummies and dummies for deciles of the average change in CO_2 emission levels (ΔE_{ct}), either total or from electricity generation, over the sample period. The specification in panel b) controls for interactions between period dummies and dummies for deciles of the average $AIpen_{ct}$ over the sample period. The specification in panel c) jointly includes the two sets of interactions used in panel a) and b). All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table 10: AI Penetration and CO_2 Emissions: Industry-Specific Shocks

	Total (Level) (1)	Electric (Level) (2)	Total (Per Capita) (3)	Electric (Per Capita) (4)
	a) No CZs in Top 5% of $AIpen$			
2nd Stage Regression				
$AIpen_{ct}$	286.647***	40.162	0.039	0.317**
	(98.170)	(49.898)	(0.250)	(0.162)
1st Stage Regression				
$AIexp_{ct}$	0.835***	0.835***	0.835***	0.835***
	(0.167)	(0.167)	(0.167)	(0.167)
Kleibergen-Paap $F-$Statistic	24.90	24.90	24.90	24.90
Obs.	2,740	2,740	2,740	2,740
	b) Leave-One-Out $AIexp$			
2nd Stage Regression				
$AIpen_{ct}$	147.207***	60.320**	0.018	0.120***
	(47.113)	(25.016)	(0.060)	(0.045)
1st Stage Regression				
$AIexp_{ct}$	1.871***	1.871***	1.871***	1.871***
	(0.281)	(0.281)	(0.281)	(0.281)
Kleibergen-Paap $F-$Statistic	44.42	44.42	44.42	44.42
Obs.	2,888	2,888	2,888	2,888
	c) No AI-Developing Industries			
2nd Stage Regression				
$AIpen_{ct}$	283.547***	83.399*	0.004	0.224*
	(81.194)	(49.659)	(0.190)	(0.115)
1st Stage Regression				
$AIexp_{ct}$	1.197***	1.197***	1.197***	1.197***
	(0.261)	(0.261)	(0.261)	(0.261)
Kleibergen-Paap $F-$Statistic	20.95	20.95	20.95	20.95
Obs.	2,888	2,888	2,888	2,888

Notes: The sample consists of 722 CZs observed over four five-year periods. The dependent variable is the change in CO_2 emissions, either total or from electricity generation, measured in levels or per 1,000 individuals aged 16 and older, as specified in the column headers. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state$\times$year fixed effects and the control variables used in column (5) of Table 3. Panel a) excludes CZs that fall within the top 5% of the distribution of average $AIpen_{ct}$ over the sample period. In panel b), the variable $AIexp_{ct}$ is constructed by excluding the CZ to which it refers. Panel c) computes both $AIpen_{ct}$ and $AIexp_{ct}$ after excluding industries that are likely involved in the development of AI applications (see footnote 13). All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

of CZs with exceptionally high levels of AI penetration. A related concern is that the industry shifts used to construct the instrument may themselves be influenced by unobserved shocks affecting specific CZs. If such shocks also impacted CO_2 emissions, the exclusion restriction would be violated. We believe this concern to be mitigated by the small size of most CZs. Yet, we explicitly address the issue in panel b). We employ a "leave-one-out" instrument, in which industry-level shifts are calculated after excluding the CZ to which the instrument refers. The results remain robust. Finally, we deal with industry-level shocks. In panel c), we reconstruct both $AIpen_{ct}$ and $AIexp_{ct}$ after excluding industries likely involved in the development of AI applications.[13] This exercise helps address two concerns. First, it assuages the risk that our results are driven by the production of AI applications rather than their adoption. Second, since these industries tend to have high concentrations of AI-related occupations, it also mitigates the concern that our findings are shaped by shocks concentrated in just a few high-penetration industries. The main findings are confirmed.

We now turn to discussing underlying trends. In panel a) of Table 11, we augment eq. (1) by including a full set of interactions between period dummies and the initial level of CO_2 emissions in each CZ. These interactions flexibly account for heterogeneous trends across CZs with different baseline pollution levels. The main results are unchanged. In panel b), we conduct a falsification test by regressing changes in CO_2 emissions over the first sample period on AI penetration in the final period. If future AI penetration were to explain past changes in emissions, this would suggest the presence of pre-existing trends or persistent CZ-level characteristics that influence both AI and emissions. Reassuringly, the estimated coefficients on $AIpen_{ct}$ are consistently small and statistically insignificant. Finally, in panel c), we estimate eq. (1) including the lagged value of ΔE_{ct} as an additional control. In line with the falsification test, this variable has no bearing on the main results, further supporting their robustness to concerns about pre-trends.

As a final exercise, we adopt a different perspective and consider the possibility that the instrument may be correlated with the error term due to unobserved confounding factors. We then follow the approach of Conley et al. (2012) and evaluate how severe a violation of the exclusion restriction would need to be for inference on β to become uninformative about the causal effect of AI penetration (see Appendix C for details). Figure 4 presents the main results. It plots 90% confidence intervals for the coefficients β under varying degrees of violation of the exclusion restriction, characterized by the parameter δ. When $\delta = 0$, i.e., the benchmark case, the exclusion restriction holds. When $\delta > 0$, the exclusion restriction is violated. Specifically, a value of $\delta = x > 0$ corresponds to a violation such that a one s.d. change in $AIexp_{ct}$ has a direct effect on ΔE_{ct} equivalent to the effect of a x s.d.

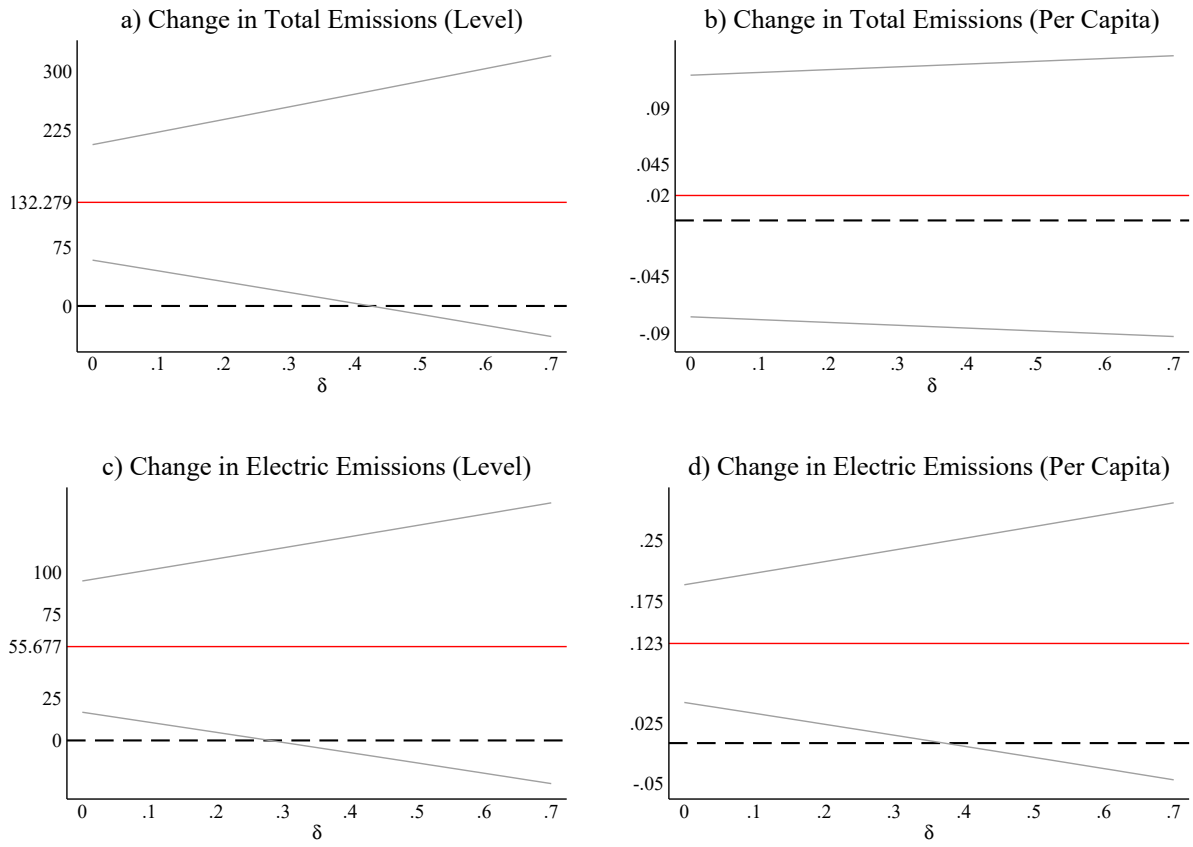
[13] Specifically, we exclude the following industries: "Computer and Related Equipment", "Computer and Data Processing Services", "Telecommunication Services" and "Communication, Audio and Video Equipment".

Table 11: AI Penetration and CO_2 Emissions: Underlying Trends

	Total (Level) (1)	Electric (Level) (2)	Total (Per Capita) (3)	Electric (Per Capita) (4)
	a) Initial Emissions×Period Dummies			
2nd Stage Regression				
$AIpen_{ct}$	123.388***	49.765***	-0.016	0.071*
	(41.629)	(19.019)	(0.059)	(0.037)
1st Stage Regression				
$AIexp_{ct}$	1.948***	1.967***	1.975***	1.972***
	(0.275)	(0.270)	(0.276)	(0.273)
Kleibergen-Paap $F-$Statistic	50.18	53.21	51.21	52.11
Obs.	2,888	2,888	2,888	2,888
	b) Placebo Test with Future $AIpen_{ct}$			
2nd Stage Regression				
$AIpen_{ct}$	-18.469	-20.888	0.047	0.017
	(39.370)	(14.542)	(0.046)	(0.016)
1st Stage Regression				
$AIexp_{ct}$	2.312***	2.312***	2.312***	2.312***
	(0.384)	(0.384)	(0.384)	(0.384)
Kleibergen-Paap $F-$Statistic	36.20	36.20	36.20	36.20
Obs.	722	722	722	722
	c) ΔE_{ct-1} as Control			
2nd Stage Regression				
$AIpen_{ct}$	313.599***	137.364***	0.113	0.165**
	(97.671)	(39.543)	(0.094)	(0.070)
1st Stage Regression				
$AIexp_{ct}$	1.554***	1.560***	1.560***	1.560***
	(0.282)	(0.281)	(0.281)	(0.281)
Kleibergen-Paap $F-$Statistic	30.34	30.83	30.82	30.82
Obs.	2,166	2,166	2,166	2,166

Notes: The sample consists of 722 CZs observed over four five-year periods, except in panel b), where it consists of a cross-section of 722 CZs. The dependent variable is the change in CO_2 emissions, either total or from electricity generation, measured in levels or per 1,000 individuals aged 16 and older, as specified in the column headers. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). The specifications in panels a) and c) include CZ fixed effects, state×year fixed effects and the control variables used in column (5) of Table 3. The specification in panel b) includes state fixed effects, along with demographic, economic and topographic controls measured in 2000. It also accounts for automation exposure, import competition, and weather shocks over the 2002–2007 period. Panel a) controls for interactions between period dummies and the initial level of CO_2 emissions. In panel b), the dependent variable is the change in CO_2 emissions over 2002-2007 and is regressed on AI penetration over 2017–2022. The specification in panel c) additionally controls for the lagged value of the dependent variable. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Figure 4: Sensitivity of Inference to Violations of Exclusion Restriction



Notes: The figure plots 90% confidence intervals around the baseline 2SLS coefficient on $AIpen_{ct}$ for different priors about a potential violation of the exclusion restriction. Priors are described by the parameter δ reported on the horizontal axis: $\delta = 0$ implies that the exclusion restriction is satisfied; $\delta = x > 0$ corresponds to a violation of the exclusion restriction such that a one standard deviation change in $AIexp_{ct}$ has a direct effect on CO_2 emissions equal to a x standard deviation change in $AIpen_{ct}$. The confidence intervals are based on standard errors clustered at the state-year level.

change in $AIpen_{ct}$.

As δ departs from zero, the confidence intervals around the estimated coefficients widen progressively. However, owing also to the strong predictive power of the instrument, the loss of precision occurs relatively slowly. Focusing on the coefficients in panels a), c) and d), which are precisely estimated under the benchmark case $\delta = 0$, the confidence intervals begin to include zero only for values of δ in the range $(0.3, 0.4)$. This implies that, for our estimates to become uninformative about the causal effect of AI, the direct effect of $AIexp_{ct}$ on ΔE_{ct} would need to be at least 30% to 40% as large as the effect of a comparable exogenous change in $AIpen_{ct}$. Such a change is approximately equal to the sample median of $AIpen_{ct}$ and roughly corresponds to the difference in average AI penetration between the CZs of Houston or Indianapolis and those of Lawrence or Madison County. These results suggest that our estimates would remain informative about the effects of AI even in the presence of non-trivial violations of the exclusion restriction.

7 Power Sources and Data Centers

Our findings so far indicate that AI penetration leads to higher emissions, primarily through the expansion of economic activity. A notable exception to this pattern is CO_2 emissions from electricity generation, which also rise on a per-capita basis with AI penetration. While this points to mechanisms beyond the scale effect, the result does not seem to be driven by changes in the composition of output at the CZ level or in techniques at the industry level. In this section, we examine the impact of AI penetration on CO_2 emissions from electricity generation more closely. Using highly detailed data on the location and characteristics of power plants and data centers, we document that AI penetration leads to a shift in the electricity generation mix toward more carbon-intensive sources, and that data centers play an important role in driving this change.

7.1 Sources of Electricity Generation

To quantify the sources of electricity generation within CZs, we resort to novel and highly detailed micro data on power plants operating in the US. These data are sourced from the Emissions and Generation Resource Integrated Database (eGRID), a comprehensive dataset covering nearly all US electricity-generating facilities with at least 1MW capacity.[14] The dataset is assembled and maintained by the US Environmental Protection Agency (EPA), which integrates information from the Energy Information Administration (EIA) and EPA Clean Air Markets.

[14]Capacity is the instantaneous power output of a generator.

Over the sample period, the total number of power plants with non-negative net electricity generation operating inland in the US is 13,154. Most power plants operate in all sample years while some may start or cease operations over time; on average, in a given year, the number of active power plants is 7,803. The database eGRID contains the name, nominal capacity (in MW) and exact geographical location (latitude and longitude) of all power plants, allowing us to attribute each of them to a CZ. As shown in Appendix Table A.4 and Appendix Figure A.2, the number of CZs hosting at least one power plant over the sample period is 662, for an average number of 13 power plants per CZ. The database also contains the total annual net generation of electricity (in MWh) by each power plant, as well as its partition into renewable and non-renewable sources of generation.[15]

For each CZ c, we compute the change in total net electricity generation from all power plants over period t, denoted ΔEG_{ct}, both in absolute terms and per 1,000 inhabitants to account for differences in scale. We also calculate the change in the share of net electricity generated from non-renewable sources, $\Delta \frac{EG^{NR}_{ct}}{EG_{ct}}$. Equation (1) is then estimated using each of these variables as the dependent variable in turn. The results, reported in Table 12, show that AI penetration has no statistically significant effect on total net electricity generation, whether measured in levels or per capita (columns 1 and 2). While AI penetration does not appear to increase overall electricity production within a CZ, it does significantly alter the energy mix: as shown in column (3), higher AI penetration is associated with a greater reliance on non-renewable energy sources. The estimated coefficient suggests that a one s.d. higher $AIpen_{ct}$ corresponds to a 0.15 s.d. (3 percentage points) higher share of electricity generated from non-renewable sources. The estimate also implies that the difference in $AIpen_{ct}$ corresponding to the interquartile range of the distribution accounts for roughly 6.4% of the observed difference in the share of electricity generated from non-renewable sources in the corresponding CZs. Columns (4)–(6) replicate the analysis restricting the sample to CZs hosting large power plants (capacity above 10 MW). The results remain virtually unchanged, indicating that the observed pattern is not driven by smaller generation units, which tend to rely more heavily on less clean energy sources.

These findings indicate that AI penetration leads to a shift in the composition of electricity generation, increasing reliance on non-renewable relative to renewable sources. Besides being more carbon-intensive, non-renewables differ from renewables in other dimensions. In particular, they offer greater reliability, dispatchability and round-the-clock availability, making them particularly suited to satisfy technologies that require a stable electricity supply. In contrast, renewable sources like solar and wind are still intermittent and often dependent on weather conditions and storage infrastructure, which can limit their ability to match continuous load profiles. In the next section, we show that a key driver of

[15]The former relates to hydro, biomass, wind, solar and geothermal generation. The latter includes nuclear, gas, oil, coal and other fossil fuels.

Table 12: AI Penetration and Electricity Generation

	Net Electricity Generation (All Power Plants)			Net Electricity Generation (High-Capacity)		
	Total	Per Capita	Non-Renewables Share	Total	Per Capita	Non-Renewables Share
	(1)	(2)	(3)	(4)	(5)	(6)
2nd Stage Regression						
$AIpen_{ct}$	-87.526	-0.198	7.512***	-80.047	-0.141	7.242***
	(83.425)	(0.158)	(2.001)	(85.368)	(0.161)	(2.000)
1st Stage Regression						
$AIexp_{ct}$	1.967***	1.967***	1.978***	1.967***	1.967***	1.967***
	(0.280)	(0.279)	(0.280)	(0.283)	(0.283)	(0.285)
Kleibergen-Paap $F-$Statistic	49.26	49.26	50.19	48.48	48.48	47.74
Obs.	2,431	2,431	2,386	2,267	2,267	2,240

Notes: The sample includes 646 CZs with power plant activity observed for at least one year in columns (1)–(3), and 611 CZs with high-capacity power plant activity observed for at least one year in columns (4)–(6). The smaller number of observations in columns (3) and (6) reflects CZ–year pairs in which power plants did not produce electricity. The dependent variable is the change in net electricity generation (measured in TWh), expressed in levels in columns (1) and (4) and per 1,000 individuals aged 16 and older in columns (2) and (5). In columns (3) and (6), the dependent variable is the change in the share of net electricity generated from non-renewable sources. The variables $AIpen_{ct}$ and $AIexp_{ct}$ capture AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state×year fixed effects, and the control variables used in column (5) of Table 3. High-capacity power plants are defined as those with a nominal capacity exceeding 10 MW. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

this shift is the activity of data centers, which require a stable and reliable energy supply.

7.2 The Role of Data Centers

A power plant's operational decisions are influenced by nearby sources of electricity demand, especially large, continuous consumers that operate around the clock and require a stable, high-capacity electricity supply. The reason is that, because electricity cannot be stored at scale without significant infrastructure, the grid must continuously balance supply and demand in real time. In line with this argument, we now show that data centers exert a substantial impact on the operations of power plants, especially their fuel choices, and consequently affect their CO_2 emissions.

To examine these effects, we exploit our micro-level data on the geographic location of data centers and power plants to estimate regressions of the following form:

$$Y_{p\tau} = \alpha_{d\tau} + \beta \ln D_p + \mathbf{X}'_{p\tau}\gamma + \varepsilon_{p\tau}, \tag{6}$$

where $Y_{p\tau}$ denotes an outcome of interest for power plant p in year τ, and D_p is the average distance (in

kilometers) between power plant p and the data centers located within its own CZ or in a contiguous CZ. The specification includes Census Division$\times$year fixed effects, $\alpha_{d\tau}$, to account for time-varying shocks within the same geographic area. The vector $\mathbf{X}_{p\tau}$ includes control variables, namely, the log distance of the power plant from the nearest city and the population of that city. The coefficient of interest, β, captures how a given outcome varies across power plants that are located within the same Census Division, share similar proximity to, and size of, the nearest city, but differ in their distance from data centers.

A concern with the OLS estimates of β is that the location of data centers is unlikely to be exogenous. Rather than being randomly assigned, data centers tend to select locations based on various characteristics, some of which may also influence the operations of power plants. For instance, data centers may prefer less urbanized areas due to lower rental and land costs, whereas power plants are frequently located closer to urban centers to meet concentrated energy demand. This selection would imply a downward bias in the OLS estimate of β.

Accordingly, we also report 2SLS estimates of eq. (6). Our instrument captures variation in D_p arising from differences in the AI exposure of the areas surrounding each power plant. The instrument is constructed as follows:

$$AIexp_p = \frac{1}{C_p} \sum_{c=1}^{C_p} \left(\frac{\overline{AIexp}_c}{D_{pc}} \right), \tag{7}$$

where $\overline{AIexp}_c$ is the average AI exposure of CZ c over the sample period, D_{pc} is the distance (in thousands of kilometers) between power plant p and the centroid of CZ c, and C_p is the number of CZs adjacent to the one in which plant p is located. In words, $AIexp_p$ is a distance-weighted average of the AI exposure of CZs neighboring power plant p's CZ.[16] The intuition behind this instrument is as follows. As shown in Fang and Greenstein (2025), the presence of data centers is significantly predicted by local demand from information-intensive industries. At the same time, AI exposure is unlikely to be directly correlated with contemporaneous shocks affecting power plant operations, as $\overline{AIexp}_c$ is constructed from aggregate industry-level AI shifts and historical industrial compositions.

The results are reported in Table 13. Each column presents estimates from equation (6) using a different outcome variable, as indicated in the column headings. Standard errors are clustered at the power-plant level to account for serial correlation in the residuals. The OLS estimates of β in columns (1)–(3) indicate that power plants located closer to data centers emit significantly more CO_2, generate greater amounts of electricity, and rely more heavily on non-renewable energy sources.

Columns (4)–(6) present the corresponding 2SLS estimates. The first-stage coefficient on $AIexp_p$

[16] We exclude power plant p's own CZ because the distance between a plant and its own CZ's centroid is inherently influenced by the plant's location choice.

Table 13: Distance to Data Centers and Power Plant Activities

	OLS			2SLS		
		Net Electricity Generation			Net Electricity Generation	
	CO_2 Emissions	Total	Non-Renewables Share	CO_2 Emissions	Total	Non-Renewables Share
	(1)	(2)	(3)	(4)	(5)	(6)
2nd Stage Regression						
D_p	-0.188*	-0.152**	-0.026***	-3.942***	-0.357	-0.376***
	(0.113)	(0.061)	(0.008)	(1.004)	(0.433)	(0.069)
1st Stage Regression						
$AIexp_p$				-12.950***	-12.875***	-12.994***
				(1.300)	(1.043)	(1.059)
Kleibergen-Paap $F-$Statistic				99.28	152.35	150.59
Obs.	22,746	29,953	28,305	22,746	29,953	28,305

Notes: The sample consists of 11,500 power plants with non-negative electricity generation and five time periods. The smaller number of observations in columns (3) and (6) relative to columns (2) and (5) reflects power plant–year pairs with no recorded electricity production. The dependent variable is the inverse hyperbolic sine of CO_2 emissions in columns (1) and (4), and of total net electricity generation in columns (2) and (5); in columns (3) and (6), it corresponds to the share of total electricity generated from non-renewable sources. The variable D_p measures the log average distance to all data centers located within the same or contiguous CZs. All specifications control for the log distance to the nearest city, the log population of that city, and include Census Division$\times$year fixed effects. Standard errors, shown in parentheses, are clustered at the power plant level. * Significant at the 10% level; ** at the 5% level; *** at the 1% level.

is negative, statistically significant, and sizable, indicating that areas with greater AI exposure are indeed associated with data centers locating closer to power plants. The second-stage estimates of β maintain the same sign as their OLS counterparts. The results confirm that power plants situated nearer to data centers exhibit significantly higher CO_2 emissions. These higher emission levels are not driven by increased electricity production, as shown by the coefficient in column (5), which becomes imprecisely estimated once the endogenous location of data centers is accounted for. Instead, the effect stems from a markedly different energy mix: as shown in column (6), power plants located closer to data centers generate a greater share of electricity from non-renewable sources. In magnitude, the 2SLS coefficients are larger than their OLS counterparts, suggesting that the endogenous siting of data centers biases the OLS estimates downward.[17]

These findings suggest that data centers significantly affect the operations of nearby power plants, particularly by shifting electricity generation toward non-renewable sources. This aligns with the

[17]Restricting the sample to large power plants (capacity above 10 MW) yields virtually identical results, available upon request.

idea that data centers require a stable and reliable electricity supply, which non-renewable sources are generally better positioned to provide. We now show that, through this channel, data centers play a pivotal role in explaining the positive effect of AI penetration on per capita CO_2 emissions from electricity generation. To do so, we augment equation (1) as follows:

$$\Delta E_{ct} = \alpha_c + \alpha_{st} + \beta_1 AIpen_{ct} + \beta_2 \left(AIpen_{ct} \times DC_c\right) + \mathbf{X'}_{ct}\gamma + \varepsilon_{ct}, \tag{8}$$

where DC_c measures the importance of data centers in CZ c, proxied by their total size.[18] In this specification, β_1 captures the effect of AI penetration in CZs without data centers, while β_2 reflects how this effect varies with the size of local data centers. Consistent with the model and the empirical findings in Section 5.2, we expect $\beta_1 \approx 0$, indicating that AI penetration alone has no impact in areas without data centers, where it primarily reflects AI adoption within firms. In contrast, we expect $\beta_2 > 0$, implying that the impact of AI penetration is driven by data center activity and increases with their local presence. We instrument both $AIpen_{ct}$ and the interaction term $AIpen_{ct} \times DC_c$ using $AIexp_{ct}$ and its interaction with DC_c.

The results are presented in Table 14. The dependent variable is the change in CO_2 emissions from the electric sector, measured either in levels or per 1,000 inhabitants. Panel a) reports estimates from the baseline specification, while panels b)–d) present a series of robustness checks. Across all specifications, the coefficient β_2 is positive and highly statistically significant, indicating that the effect of $AIpen_{ct}$ on CO_2 emissions from the electric sector is stronger in areas with a larger data center presence. In contrast, the coefficient β_1 is consistently small and statistically insignificant, suggesting that AI adoption by firms alone does not significantly affect CO_2 emissions in the absence of data centers. These results highlight the central role of data centers in explaining the positive effect of AI penetration and CO_2 emissions across US CZs.

8 Conclusions

This paper has provided systematic evidence on the environmental consequences of AI. Leveraging a novel dataset that combines measures of AI penetration, the location of data centers and power plants, and CO_2 emissions across US CZs, we have documented four main findings. First, AI penetration has significantly increased emissions, with differences in adoption explaining a sizable share of the observed variation in emissions growth across CZs. Second, decomposition analysis shows that scale

[18] The linear term in DC_c is absorbed by the CZ fixed effects. DC_c is missing for 22 CZs, resulting in a slightly smaller sample size for these specifications.

Table 14: AI Penetration and CO_2 Emissions: The Role of Data Centers

	Electric (Level) (1)	Electric (Per Capita) (2)	Electric (Level) (3)	Electric (Per Capita) (4)
	a) Baseline		b) Decile Bins (ΔE & $AIpen$)	
2nd Stage Regression				
$AIpen_{ct}$	-9.447	0.091*	-10.789	0.076
	(25.961)	(0.054)	(38.323)	(0.080)
$AIpen_{ct} \times DC_c$	5.603***	0.003**	6.043***	0.005***
	(1.574)	(0.001)	(1.638)	(0.002)
Shea Partial R^2 ($AIpen_{ct}$)	0.161	0.161	0.064	0.062
Shea Partial R^2 ($AIpen_{ct} \times DC_c$)	0.647	0.647	0.553	0.557
Obs.	2,800	2,800	2,800	2,800
	c) Leave-One-Out $AIexp_{ct}$		d) Underlying Trends	
2nd Stage Regression				
$AIpen_{ct}$	-7.279	0.084	-15.446	0.032
	(26.830)	(0.055)	(23.635)	(0.047)
$AIpen_{ct} \times DC_c$	5.661***	0.003**	5.536***	0.004**
	(1.585)	(0.001)	(1.469)	(0.001)
Shea Partial R^2 ($AIpen_{ct}$)	0.147	0.147	0.161	0.159
Shea Partial R^2 ($AIpen_{ct} \times DC_c$)	0.633	0.633	0.642	0.649
Obs.	2,800	2,800	2,800	2,800

Notes: The sample consists of 700 CZs observed over four five-year periods. The dependent variable is the change in CO_2 emissions from electricity generation, measured either in levels or per 1,000 individuals aged 16 and older, as indicated in the column headers. The variables $AIpen_{ct}$ and $AIexp_{ct}$ capture AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). The variable DC_c represents total data center space, measured in million square feet. All specifications include CZ fixed effects, state×year fixed effects and the control variables used in column (5) of Table 3. The specification in panel b) includes interactions between period dummies and dummies for deciles of the average change in CO_2 emissions from electricity generation and for deciles of the average $AIpen_{ct}$ over the sample period. In panel c), $AIexp_{ct}$ is constructed by excluding the CZ to which it refers. The specification in panel d) adds interactions between period dummies and the initial level of CO_2 emissions from electricity generation. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

effects dominate, while shifts in industrial composition provide at most a weak mitigating influence; yet, electricity generation becomes more carbon intensive. Third, AI penetration has increased reliance on non-renewable electricity. Fourth, proximity to data centers drives this effect, as nearby power plants shift toward greater fossil fuel use.

Our findings challenge the perception of AI as a clean general-purpose technology. While AI has the potential to enhance energy efficiency in sectors such as industry, transport and buildings, these gains tend to diffuse slowly and can be outweighed by the surge in electricity demand. Our results suggest that, absent rapid decarbonization of power generation, the diffusion of AI is likely to exacerbate environmental externalities. They further highlight the importance of aligning digital and climate policies; for example, by incentivizing renewable capacity in regions that attract data centers and AI-intensive activity.

While our analysis provides novel evidence on the environmental impact of AI, several limitations remain. First, our measure of AI penetration may not fully capture the most recent advances, such as large language models, whose energy requirements are potentially larger and more heterogeneous. Second, our empirical strategy focuses on local effects within CZs and does not account for general equilibrium adjustments at the national or global level. Third, the analysis abstracts from dynamic responses, including investment in renewable capacity, improvements in data center efficiency or regulatory interventions that could mitigate emissions over time. Incorporating these effects into the analysis could be important avenues for future work.

References

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

Alekseeva, L., J. Azar, M. Giné, S. Samila, and B. Taska (2021). The demand for AI skills in the labor market. *Labour Economics 71*, 102002.

Autor, D., D. Dorn, and G. Hanson (2019). When work disappears: Manufacturing decline and the falling marriage market value of young men. *American Economic Review: Insights 1*(2), 161–178.

Autor, D. H. and D. Dorn (2013). The growth of low-skill service jobs and the polarization of the US labor market. *American Economic Review 103*(5), 1553–1597.

Autor, D. H., D. Dorn, and G. H. Hanson (2013). The China syndrome: Local labor market effects of import competition in the United States. *American Economic Review 103*(6), 2121–68.

Bartik, T. J. (1991). Who benefits from state and local economic development policies? *WE Upjohn Institute for Employment Research*.

Benetton, M., G. Compiani, and A. Morse (2023). When cryptomining comes to town: High electricity-use spillovers to the local economy. Working Paper 31312, National Bureau of Economic Research.

Blanchard, O. J. and Katz (1992). Regional evolutions. *Brookings Papers on Economic Activity 1992*(1), 1–75.

Bonfiglioli, A., G. Gancia, I. Papadakis, and R. Crinò (2025). Artificial Intelligence and jobs: Evidence from US commuting zones. *Economic Policy 40*(121), 145–194.

Borusyak, K., P. Hull, and X. Jaravel (2022). Quasi-experimental shift-share research designs. *The Review of Economic Studies 89*(1), 181–213.

Colella, F., R. Lalive, S. O. Sakalli, and M. Thoenig (2023). acreg: Arbitrary correlation regression. *The Stata Journal 23*(1), 119–147.

Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics 92*(1), 1–45.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012). Plausibly exogenous. *Review of Economics and Statistics 94*(1), 260–272.

Davenport, C., C. Singer, N. Mehta, B. Lee, and J. Mackay (2024). AI, data centers and the coming US power demand surge. *Goldman Sachs 26*.

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

EPRI (2024). Powering data centers: U.S. energy system and emissions impacts of growing loads. *EPRI Report 3002031198*.

Fang, T. P. and S. Greenstein (2025). Where the cloud rests: The economic geography of data centers. *Strategy Science* (forthcoming).

Feher, A., E. Garcia-Appendini, and R. Mihet (2025). Is AI trained on public money? Evidence from U.S. data centers. *CEPR Discussion Paper No. 20758*.

Gaulier, G. and S. Zignago (2010). Baci: International trade database at the product-level (The 1994-2007 version). *CEPII Working Paper 2010-23*.

Gurney, K., P. Dass, A. Kato, B. Gawuc, H. Aslam, and H. Sun (2025). Vulcan version 4.0 high-resolution annual carbon dioxide emissions in the US for the 2010-2022 time period. *Mimeo*.

Gurney, K., Y. Zhou, S. Geethakumar, A. Godbole, D. Mendoza, M. Vaidhyanathan, and N. Sahni (2009). The Vulcan Project: Recent advances and emissions estimation for the NACP mid-continent intensive campaign region. In *AGU Fall Meeting Abstracts*, Volume 2009, pp. B51E–0337.

Hanson, G. (2022). Immigration and regional specialization in AI. In *Robots and AI: A New Economic Era*, pp. 180–231. Routledge.

Knittel, C. R., J. R. L. Senga, and S. Wang (2025). Flexible data centers and the grid: Lower costs, higher emissions? *National Bureau of Economic Research 34065*.

Lange, S., J. Pohl, and T. Santarius (2020). Digitalization and energy consumption. Does ICT reduce energy demand? *Ecological Economics 176*, 106760.

Levinson, A. (2009). Technology, international trade, and pollution from US manufacturing. *American Economic Review 99*(5), 2177–2192.

Luccioni, A. S., S. Viguier, and A.-L. Ligozat (2023). Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research 24*(253), 1–15.

Ruggles, S., S. Flood, M. Sobek, D. Backman, A. Chen, G. Cooper, S. Richards, R. Rogers, and M. Schouweiler (2023). *IPUMS USA: Version 14.0.* Minneapolis, MN: IPUMS, 2023.

Taddy, M. (2018). The technological elements of artificial intelligence. Technical Report 24301, National Bureau of Economic Research.

Tolbert, C. M. and M. Sizer (1996). US commuting zones and labor market areas: A 1990 update. Technical Report 278812, United States Department of Agriculture, Economic Research Service.

A Data Appendix

In this Appendix, we provide detailed information about data sources and variable definitions. Our sample includes 722 CZs and spans from 2002 to 2022. The main specifications are estimated by stacking four five-year differences corresponding to the periods 2002-2007, 2007-2012, 2012-2017 and 2017-2022.[1]

A.1 Main Variables

A.1.1 Data-Intensive Occupations, AI Penetration and AI Exposure

To measure employment in data-intensive occupations within each CZ, we follow Bonfiglioli et al. (2025) and exploit the novel "Hot Technologie" section of the O*NET database. This section reports the software requirements most frequently included in all current employer job postings in the US, separately for each occupation in the 2018 Standard Occupational Classification (SOC). The original list contains 172 software titles, ranging from general-purpose applications such as Microsoft Excel to advanced programming languages like Python and C++. We ask GPT-5 to analyze the description of each software (as provided in O*NET) and to identify keywords related to machine learning and big data management.[2] We then select the software titles whose descriptions contain at least one of these keywords, yielding a list of 31 tools reported in Table A.1. This list covers tools for data processing, storage and large-scale data management, as well as frameworks that support machine learning and AI, either by directly executing and adapting algorithms or by generating and managing data used as inputs to AI systems. Based on this refined software list, we rely on the "Hot Technologie" section of O*NET to identify occupations for which each tool is considered "in demand". This procedure produces an initial set of 82 occupations whose job postings typically require knowledge of at least one of these tools.

We refine the list through two sequential filters. First, we exclude occupations for which only one software tool is listed as "in demand". This step removes 37 occupations that rely on a single specialized tool in their daily activities, such as "Geographers" (who only use SQL) or "Sales Engineers" (who only use Amazon Web Services). Second, we apply the crosswalk between the SOC and the Classification of Instructional Programs (CIP) to retain only those occupations requiring skills in do-

[1] When data for a variable are missing at an endpoint of an interval, we use the closest available year and express the resulting change as a five-year equivalent.

[2] The list of keywords identified by GPT-5 is: analytics, intelligence, clustering, expert system, imaging, OCR, pattern, speech, voice recognition, text to speech, synthesizer, information retrieval, database, mining, compression, conversion, query, metadata, workflow, ERP, CRM, enterprise management and storage networking.

Table A.1: Software Used to Identify the Data-Intensive Occupations

Adobe After Effects	Canva	Oracle Database
Adobe Creative Cloud Software	Extensible Markup Language XML	PyTorch
Adobe Illustrator	Google Analytics	SAP Software
Adobe Photoshop	Graph QL	Service Now
Alteryx Software	Informatica Software	Splunk Enterprise
Amazon Redshift	Jenkins CI	Structured Query Language SQL
Amazon Simple Storage Service S3	Microsoft Access	Tableau
Amazon Web Services AWS Software	Microsoft Power BI	Workday Software
Ansible Software	Microsoft SQL Server	Yardi Software
Apache Spark	Microsoft SQL Server Integration Service	
Atlassian JIRA	MySQL	

mains typically associated with AI. Specifically, we focus on non-administrative, non-supportive and non-educational occupations whose skill requirements align with academic programs in computer and information sciences.[3] This step excludes occupations such as "Economists", for whom SQL, Tableau and Microsoft Power BI are "in demand", but whose core competencies lie in the social sciences. The final list of data-intensive occupations comprises 13 titles (see Table 1).

We combine the list of data-intensive occupations with micro-level employment data from the US Census (for the year 2000) and the American Community Survey (ACS, for the years 2007, 2012, 2017 and 2022), using the information on each worker's SOC occupation available in both datasets (Ruggles et al., 2023).[4] To track the data-intensive occupations back in time, across the revisions of the SOC occurred over the sample period, we use correspondence tables from the US Bureau of Labor Statistics (BLS). We construct both total employment and employment in data-intensive occupations in each CZ using sample weights, considering individuals aged 16+, who are not unpaid family workers, do not reside in institutional group quarters and have reported being employed over the year prior to the interview.

With these employment data in hand, we construct our proxies for AI penetration, $AIpen_{ct}$, and AI exposure, $AIexp_{ct}$, as in eq. (2) and (3). To construct $AIexp_{ct}$, we use data from the 1990 US

[3]The SOC–CIP crosswalk is provided by the National Center for Education Statistics of the US Department of Education. The relevant CIP codes among the occupations that pass the first filter are 110101, 110201 and 110701.

[4]The ACS is a 1% random sample of the US population. To increase sample size, we follow Autor et al. (2013) and Acemoglu and Restrepo (2020) by using pooled 3-year ACS 2007 data for 2005–2007 and pooled 5-year ACS 2012, 2017 and 2022 for the periods 2008–2012, 2013–2017 and 2018–2022, respectively. We rely on 3-year pooled data for 2005–2007 because 5-year pooled data are not available prior to 2009. Both the US Census and the ACS are representative at the level of micro-regions called Public Use Microdata Areas (PUMAs). We map PUMAs to CZs using the crosswalks developed by Autor and Dorn (2013). For the year 2002, we use data from the 2000 US Census, as the ACS does not report PUMA identifiers prior to 2004.

Table A.2: Summary Statistics on AI Penetration and AI Exposure

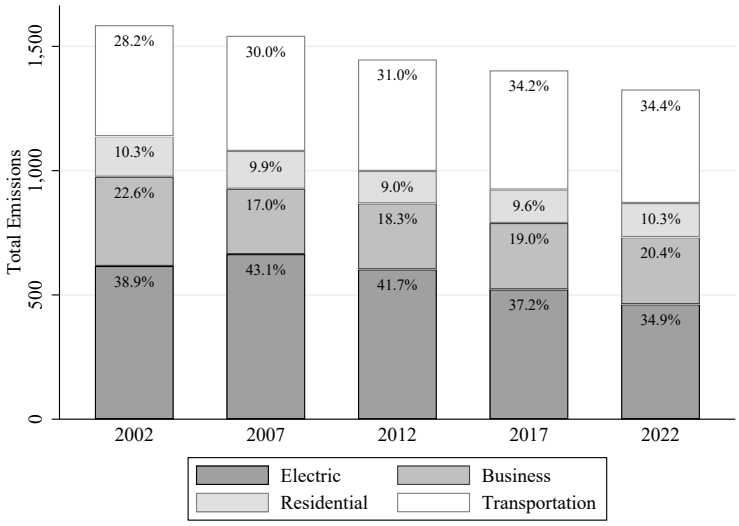
	Obs.	Mean	SD
a) AI Penetration			
$AIpen_{ct}$	2,888	0.002	0.004
$AIpen_{ct}$ (Data Scientists)	2,888	0.000	0.001
$AIpen_{ct}$ (Hanson, 2022)	2,888	0.001	0.004
$AIpen_{ct}$ (Eloundou et al., 2024)	2,888	-0.002	0.010
$AIpen_{ct}$ (No Top & Bottom CZs by $AIpen_{ct}$)	2,880	0.002	0.004
$AIpen_{ct}$ (No CZs in Top 5% of $AIpen_{ct}$)	2,740	0.002	0.003
$AIpen_{ct}$ (No AI-Developing Industries)	2,888	0.001	0.003
$AIpen_{ct}$ (No Covid-19)	2,888	0.001	0.003
$AIpen_{ct}$ (Winsorized 1% Tails)	2,888	0.002	0.003
$AIpen_{ct}$ (Winsorized 2SD from Mean)	2,888	0.002	0.004
$AIpen_{ct}$ (10-Year Changes)	1,444	0.004	0.006
b) AI Exposure			
$AIexp_{ct}$	2,888	0.002	0.002
$AIexp_{ct}$ (Data Scientists)	2,888	0.000	0.001
$AIexp_{ct}$ (Hanson, 2022)	2,888	0.002	0.003
$AIexp_{ct}$ (Eloundou et al., 2024)	2,888	-0.002	0.004
$AIexp_{ct}$ (No Top & Bottom CZs by $AIpen_{ct}$)	2,880	0.002	0.002
$AIexp_{ct}$ (No CZs in Top 5% of $AIpen_{ct}$)	2,740	0.002	0.002
$AIexp_{ct}$ (No AI-Developing Industries)	2,888	0.001	0.002
$AIexp_{ct}$ (No Covid-19)	2,888	0.001	0.001
$AIexp_{ct}$ (Leave-One-Out)	2,888	0.002	0.002
$AIexp_{ct}$ (10-Year Changes)	1,444	0.004	0.003

Notes: The sample consists of 722 CZs observed over four five-year periods. For variables measured in ten-year changes, each CZ is observed over two ten-year periods.

Census to calculate the share of industry i in total employment within CZ c in 1990, covering 210 industries across all sectors of the economy. The industry-level AI shifts are derived from micro-level data from the 2000 US Census and the ACS.[5] Descriptive statistics on both $AIpen_{ct}$ and $AIexp_{ct}$ are reported in Table A.2.

[5]We use the industry crosswalk developed by Autor et al. (2019) to map between *ind1990* and *ind1990dd* industry codes.

Figure A.1: CO_2 Emissions in the US



Notes: The figure shows the evolution of total CO_2 emissions (in million metric tonnes of carbon) in the US, along with the contributions from different sectors.

A.1.2 CO_2 and Other Air Pollutants

Data on CO_2 emissions come from the Vulcan Project database. We draw information on CO_2 emissions for the year 2002 from version 2.2 (Gurney et al., 2009) and for the period 2010-2021 from version 4.0 (Gurney et al., 2025).[6] The annual emissions data are reported at a high spatial resolution, using 1 km $\times$ 1 km grid, but are also available as aggregated CO_2 emission totals at the county level. We use the county-level information and obtain the total amount of CO_2 emissions by CZ and year using the crosswalk developed by Autor and Dorn (2013). The Vulcan dataset categorizes emissions across a wide range of economic sectors, which we group into the following categories: electric (electricity production); business (including industrial facilities, commercial buildings and cement production); residential (emissions from residential buildings); and transportation (covering both on-road and non-road sources). For each aggregate sector, we include total emissions across all relevant fuel types (coal, natural gas and petroleum products) from both point sources, like power plants and industrial facilities, and non-point sources, i.e., spatially diffuse activities including residential heating and commercial energy use. Emissions are expressed in million metric tonnes of carbon. Figure A.1 shows the evolution of total CO_2 emissions in the US and the contributions of the four sectors over the sample period.

[6] See https://vulcan.rc.nau.edu/.

To calculate per-capita CO_2 emissions, we use data on total population in each CZ and year. Total population is computed based on micro-level data from the US Census and the ACS, using sample weights and considering working-age individuals (aged 16+) who are not unpaid family workers and do not reside in institutional group quarters.

We combine data on CO_2 emissions with information on other air pollutants from various sources. Annual satellite-derived data on $PM_{2.5}$ for the years 2002, 2007, 2012, 2017 and 2022 come from the Atmospheric Composition Analysis Group (V5.NA.04.02 dataset).[7] The other pollutants (CO, SO_2, NO_2, $PM_{1.0}$) are obtained from the Gridded Environmental Impact Frame Database (EIF) for the years 2002, 2007, 2012, 2017 and 2020.[8] All pollutant concentration estimates are provided at a high spatial resolution based on 0.01-degree grid cells. We combine these data with the shapefile of US CZs to obtain annual average concentration levels by CZ. The aggregate pollution index is constructed using principal component analysis on the five individual pollutants and retaining the first factor only. Descriptive statistics on the outcome variables are presented in Table A.3.

A.1.3 Data Centers

Information on data centers is obtained by scraping the website of *Datacenters.com*.[9] This is a platform that helps businesses find and compare colocations, bare metal servers and cloud solutions from various providers worldwide, connecting users with data center providers. *Datacenters.com* collects information on all data centers from hundreds of providers, covering more than 6,300 facilities across 108 countries. We focus on the 2,194 data centers located in the US. For each data center, we retrieve its name and address. For a subset of data centers (1,445), *Datacenters.com* also provides information on their physical size—floor space, measured in square feet. We geolocalize each data center using forward geocoding through Stata's "*opencagegeo*" package. With the geographical coordinates in hand, we finally match data centers to CZs. As shown in Table A.4, data centers are located in 169 CZs, with roughly 13 data centers per CZ and an average size of 126 thousand square feet.[10]

A.1.4 Power Plants

Information on power plants comes from the Emissions and Generation Resource Integrated Database (eGRID) and is available for the years 2000, 2007, 2012, 2016 and 2022. eGRID provides detailed micro-level data on all power plants with at least 1MW capacity operating in the US. The dataset

[7] See https://sites.wustl.edu/acag/datasets/surface-pm2-5/.
[8] See https://www.census.gov/data/experimental-data-products/gridded-eif.html.
[9] See https://www.datacenters.com/locations/united_states.
[10] The CZs with available information on data center size are 147 out of 169.

Table A.3: Summary Statistics on Outcomes

	Obs.	Mean	SD
a) CO_2 Emissions (Level)			
ΔE_{ct} (Total)	2,888	-0.140	0.728
ΔE_{ct} (Electric)	2,888	-0.096	0.602
ΔE_{ct} (Business)	2,888	-0.017	0.329
ΔE_{ct} (Residential)	2,888	-0.019	0.179
ΔE_{ct} (Transportation)	2,888	-0.008	0.153
$\Delta PM2.5_{ct}$	2,888	-0.901	1.543
$\Delta PM1.0_{ct}$	2,886	-1.627	3.435
ΔCO_{ct}	2,886	-0.055	0.054
$\Delta NO2_{ct}$	2,886	-1.090	1.109
$\Delta SO2_{ct}$	2,886	-0.463	0.513
$\Delta PollutionIndex_{ct}$	2,886	-0.565	0.374
ΔE_{ct} (Total, No Top & Bottom CZs by ΔE_{ct})	2,880	-0.138	0.708
ΔE_{ct} (Electric, No Top & Bottom CZs by ΔE_{ct})	2,880	-0.093	0.579
ΔE_{ct} (Total, No Top & Bottom CZs by $AIpen_{ct}$)	2,880	-0.140	0.726
ΔE_{ct} (Electric, No Top & Bottom CZs by $AIpen_{ct}$)	2,880	-0.096	0.600
ΔE_{ct} (Total, No CZs in Top 5% of $AIpen_{ct}$)	2,740	-0.139	0.698
ΔE_{ct} (Electric, No CZs in Top 5% of $AIpen_{ct}$)	2,740	-0.094	0.584
ΔE_{ct} (Total, Winsorized 1% Tails)	2,888	-0.133	0.627
ΔE_{ct} (Electric, Winsorized 1% Tails)	2,888	-0.088	0.490
ΔE_{ct} (Total, Windsorized 2SD from Mean)	2,888	-0.096	0.467
ΔE_{ct} (Electric, Winsorized 2SD from Mean)	2,888	-0.063	0.373
ΔE_{ct} (Total, No Covid-19)	2,888	-0.123	0.803
ΔE_{ct} (Electric, No Covid-19)	2,888	-0.103	0.689
ΔE_{ct} (Total, 10-Year Changes)	1,444	-0.189	0.996
ΔE_{ct} (Electric, 10-Year Changes)	1,444	-0.117	0.724
b) CO_2 Emissions (Per 1,000 inhabitants)			
ΔE_{ct} (Total)	2,888	-0.001	0.011
ΔE_{ct} (Electric)	2,888	-0.001	0.010
ΔE_{ct} (Business)	2,888	-0.000	0.003
ΔE_{ct} (Residential)	2,888	-0.000	0.002
ΔE_{ct} (Transportation)	2,888	0.000	0001
ΔE_{ct} (Total, No Top & Bottom CZs by ΔE_{ct})	2,880	-0.001	0.011
ΔE_{ct} (Electric, No Top & Bottom CZs by ΔE_{ct})	2,880	-0.001	0.010
ΔE_{ct} (Total, No Top & Bottom CZs by $AIpen_{ct}$)	2,880	-0.001	0.009
ΔE_{ct} (Electric, No Top & Bottom CZs by $AIpen_{ct}$)	2,880	-0.001	0.008
ΔE_{ct} (Total, No CZs in Top 5% of $AIpen_{ct}$)	2,596	-0.001	0.011
ΔE_{ct} (Electric, No CZs in Top 5% of $AIpen_{ct}$)	2,596	-0.001	0.010
ΔE_{ct} (Total, Winsorized 1% Tails)	2,888	-0.001	0.006
ΔE_{ct} (Electric, Winsorized 1% Tails)	2,888	-0.001	0.004
ΔE_{ct} (Total, Winsorized 2SD from Mean)	2,888	-0.001	0.005
ΔE_{ct} (Electric, Winsorized 2SD from Mean)	2,888	-0.001	0.004
ΔE_{ct} (Total, Population 16-65)	2,888	-0.001	0.012
ΔE_{ct} (Electric, Population 16-65)	2,888	-0.001	0.011
ΔE_{ct} (Total, Labor Force)	2,888	-0.002	0.013
ΔE_{ct} (Electric, Labor Force)	2,888	-0.001	0.012
ΔE_{ct} (Total, No Covid-19)	2,888	-0.001	0.010
ΔE_{ct} (Electric, No Covid-19)	2,888	-0.001	0.009
ΔE_{ct} (Total, 10-Year Changes)	1,444	-0.002	0.010
ΔE_{ct} (Electric, 10-Year Changes)	1,444	-0.001	0.010

Notes: The sample consists of 722 CZs observed over four five-year periods. For variables measured in ten-year changes, each CZ is observed over two ten-year periods.

Table A.4: Summary Statistics on Data Centers and Power Plants

	Total	CZs Where Present	Average Number per CZ	Average Size per CZ
Data Centers	2,194	169	13.0	126.0
Power Plants	13,154	662	12.9	187.9

Notes: Data center size is measured by floor space, expressed in thousands of square feet. Power plant size is measured by nominal capacity, expressed in MW.

is managed by the US Environmental Protection Agency (EPA), which combines information from the Energy Information Administration (EIA) and EPA Clean Air Markets. We focus on the 13,154 power plants operating in the inland US with non-negative net electricity production; on average, 7,803 power plants are active in a given year. For each power plant, we have information on its name, capacity and geographic location (latitude and longitude), along with the total annual net generation of electricity and its sources: renewables (hydro, biomass, wind, solar and geothermal) and non-renewables (coal, oil, gas, nuclear and other fossils). We use the geographic coordinates to assign each power plant to a CZ. As reported in Table A.4, the number of CZs hosting at least one power plant is 662, with an average number of 13 power plants per CZ and an average capacity of 188 MW. Figure A.2 illustrates the distribution of power plants and their average capacity across CZs.

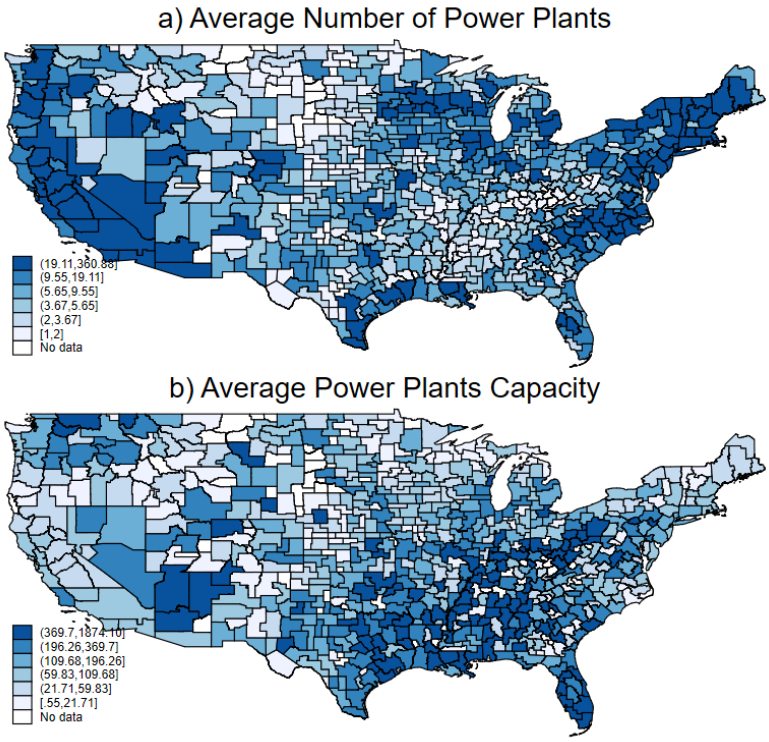
A.2 Control variables

The set of control variables included in our specifications comprise several CZ-level characteristics, encompassing population size, industrial composition and land use; automation, trade and weather shocks; and Bartik measures for changes in various industry characteristics.

A.2.1 Population, Labor Force and Industrial Composition

CZs' population and the employment share of manufacturing are computed using micro-level data from the 2000 US Census and the ACS, using sample weights and the same criteria adopted to compute AI-related and total employment (see Section A.1.1). The same data sources are used to compute working age population (individuals aged 16-65) and labor force (defined as working age population minus inactive individuals) to conduct robustness checks based on alternative normalizations of the outcome variables (see Table B.4). Finally, our plant-level regressions (see Table 13) control for the log population of the city closest to each power plant. We retrieve information on city population

Figure A.2: Power Plants across US CZs



Notes: Plot a) shows the average number of power plants per CZ, while plot b) presents their average capacity in MW. 60 CZs have no operating power plants with non-negative net electricity production during the period of analysis.

from the Intercensual Estimates of the resident total population for incorporated places and minor civil divisions, available on a yearly basis from 2000 to 2019.

A.2.2 Automation and Chinese Import Competition

To account for automation shocks, we compute a robot exposure measure following the approach proposed by Acemoglu and Restrepo (2020). In particular, exposure to industrial robots is a Bartik for changes in robot density (the number of installed robots per worker) and is constructed using data on robot installments in 17 industries sourced from the International Federation of Robotics (IFR) for the years 2000, 2007, 2012, 2017 and 2020.

To measure import competition from China at the CZ level, we use yearly product-level data defined at the 6-digit level of the Harmonized System (HS) classification and provided in the BACI database (Gaulier and Zignago, 2010) for the years 2002, 2007, 2012, 2017 and 2022.[11] We obtain a Bartik for changes in Chinese imports per worker across 82 industries in primary sectors and manufacturing, following Autor et al. (2013).

A.2.3 Weather Shocks

We obtain county-level data on average, maximum and minimum temperature in Fahrenheit degrees, and on total precipitations in inches for the years 2002, 2007, 2012, 2017 and 2020 from the National Center of Environmental Information (NCEI) of the National Oceanic and Atmospheric Administration (NOOA).[12] The section "Climate at a glance" provides real-time analysis of monthly temperature and precipitations data across the contiguous US for the study of climate variability. Starting from county-level data for each year, we compute the average temperature, the average temperature excursion (the difference between maximum and minimum temperature) and the total amount of precipitations at the CZ-year level.

Data on specific humidity and wind speed come from the Global Monitoring Laboratory (GML) database over the period 2002-2020.[13] Specific humidity is measured as the mass of water vapor (in grams) in one kilogram of dry air. Wind speed (in m/s) is computed as the square root of the sum of the square values of eastward and northward wind components. GML provides gridded data at the monthly level. Using these data, we compute the average specific humidity and wind speed in each CZ-year pair.

[11] See https://www.cepii.fr/CEPII/.
[12] See https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping/.
[13] See https://gml.noaa.gov/data.

A.2.4 Elevation and Land Use

We obtain elevation data from the GEBCO Gridded Bathymetry Data, which provides information on elevation, in meters, on a 15 arc-second interval geographic grid.[14] In our specification, we use the standard deviation of elevation and the difference between maximum and minimum elevation across all centroids of grid cells corresponding to a given CZ.

Land use data come from the Conterminous US Land Cover Projections (CONUS) database for the year 2000. CONUS provides information on land use at a 250-meter spatial resolution, with land use classified into 17 land cover classes. For each CZ, we identify the most common land use type across all grid points corresponding to that CZ. Then, we construct dummy variables for six aggregate categories of land use: (1) water; (2) developed land; (3) forest; (4) grassland; (5) cropland, hay/pasture land; and (6) herbaceous/woody wetland.

A.2.5 Bartik Measures for Changes in Industry Characteristics

The Bartik measures used in Section 5.3 are constructed as follows:

$$EIexp_{ct} = \sum_i \omega_{ci0} \times \Delta \log EI_{it}, \tag{A.1}$$

$$Lexp_{ct} = \sum_i \omega_{ci0} \times \Delta logL_{it}. \tag{A.2}$$

$$YLexp_{ct} = \sum_i \omega_{ci0} \times \Delta log\left(\frac{GO_{it}}{L_{it}}\right). \tag{A.3}$$

$$KLexp_{ct} = \sum_i \omega_{ci0} \times \Delta log\left(\frac{K_{it}}{L_{it}}\right). \tag{A.4}$$

where $\Delta logEI_{it}$, $\Delta logL_{it}$, $\Delta log\left(\frac{GO_{it}}{L_{it}}\right)$ and $\Delta log\left(\frac{K_{it}}{L_{it}}\right)$ denote the log change in energy intensity (total energy consumption over gross output), employment, gross output per worker and capital stock per worker, respectively, in industry i over period t. ω_{ci0} denotes the share of industry i in the total employment of CZ c in 2000, and is constructed using data from the US Census. Industry employment ($\Delta logL_{it}$) is computed using micro-level data from the US Census and the ACS. Yearly data on energy consumption, gross output and the capital-labor ratio for 61 industries—19 in manufacturing and 42 in services and primary sectors—are obtained from the Production Account Tables of the US Bureau of Economic Analysis (BEA) for the period 2002-2021.

[14] See https://www.gebco.net/data_and_products/gridded_bathymetry_data.

B Robustness Checks

In this Appendix, we assess the robustness of the baseline results (Table 5, columns 1 and 2, and Table 6, columns 2 and 3) along three dimensions: (i) the influence of potential outliers; (ii) alternative corrections for the standard errors; and (iii) the use of different model specifications and alternative definitions of key variables.

B.1 Outliers

Table B.1 presents 2SLS estimates of β obtained by estimating eq. (1) on various subsamples that exclude extreme observations in the key variables. In panels a) and b), we exclude CZs with the highest and lowest average values of ΔE_{ct} and $AIpen_{ct}$, respectively. The results are robust across both subsamples. Panel c) assesses the sensitivity of the results to the exclusion of the Covid-19 period. Specifically, we omit the years 2020-2022 from the construction of ΔE_{ct}, $AIpen_{ct}$ and $AIexp_{ct}$.[15] Although the Covid-19 pandemic may have influenced both emissions and AI penetration, the main results are qualitatively unchanged. Finally, Table B.2 reports results based on an alternative method for handling potential outliers. We winsorize extreme values of ΔE_{ct} (panels a) and b)) and $AIpen_{ct}$ (panels c) and d)), defined as observations that either (i) fall in the top or bottom 1% of the distribution or (ii) differ from the mean by more than two standard deviations. The main findings are confirmed.

B.2 Inference

Figure B.1 plots the baseline 2SLS estimate of β along with confidence intervals constructed using a variety of standard error corrections. Confidence interval (1) is the baseline one, corresponding to standard errors corrected for clustering at the state-year level. Confidence intervals (2) and (3) are corrected for clustering at the CZ level and state level, respectively, to account for within-CZ correlation over time and for cross-CZ correlation within states. Confidence interval (4) is corrected for two-way clustering at the CZ and year level, while confidence interval (5) for two-way clustering at the CZ and state-year level. These confidence intervals allow for residual correlation both within CZs over time and across CZs within a given year or state-year group. Confidence interval (6) applies the inference procedure proposed by Borusyak et al. (2022), which accounts for potential correlation in the residuals across observations with similar industry shares in shift-share designs. All alternative confidence intervals are similar to the baseline one.

[15] We use pooled ACS data from 2015-2019 in place of 2018-2022 to construct the final observations of $AIpen_{ct}$ and $AIexp_{ct}$ in this specification.

Table B.1: AI Penetration and CO_2 Emissions: Outliers

	Total (Level) (1)	Electric (Level) (2)	Total (Per Capita) (3)	Electric (Per Capita) (4)
	a) No Top & Bottom CZs in ΔE_{ct}			
2nd Stage Regression				
$AIpen_{ct}$	169.046***	49.484**	0.026	0.109**
	(47.118)	(22.539)	(0.068)	(0.043)
1st Stage Regression				
$AIexp_{ct}$	1.841***	1.972***	1.841***	1.972***
	(0.254)	(0.272)	(0.254)	(0.273)
Kleibergen-Paap $F-$Statistic	52.52	52.38	52.52	52.38
Obs.	2,880	2,880	2,880	2,880
	b) No Top & Bottom CZs in $AIpen_{ct}$			
2nd Stage Regression				
$AIpen_{ct}$	130.672***	56.202**	0.023	0.125***
	(44.388)	(25.323)	(0.060)	(0.046)
1st Stage Regression				
$AIexp_{ct}$	1.957***	1.957***	1.957***	1.957***
	(0.254)	(0.254)	(0.254)	(0.254)
Kleibergen-Paap $F-$Statistic	59.21	59.21	59.21	59.21
Obs.	2,880	2,880	2,880	2,880
	c) Excluding Covid-19 Pandemic			
2nd Stage Regression				
$AIpen_{ct}$	218.278**	49.290	-0.057	0.152**
	(93.924)	(42.136)	(0.110)	(0.075)
1st Stage Regression				
$AIexp_{ct}$	1.606***	1.606***	1.606***	1.606***
	(0.320)	(0.320)	(0.320)	(0.320)
Kleibergen-Paap $F-$Statistic	25.20	25.20	25.20	25.20
Obs.	2,888	2,888	2,888	2,888

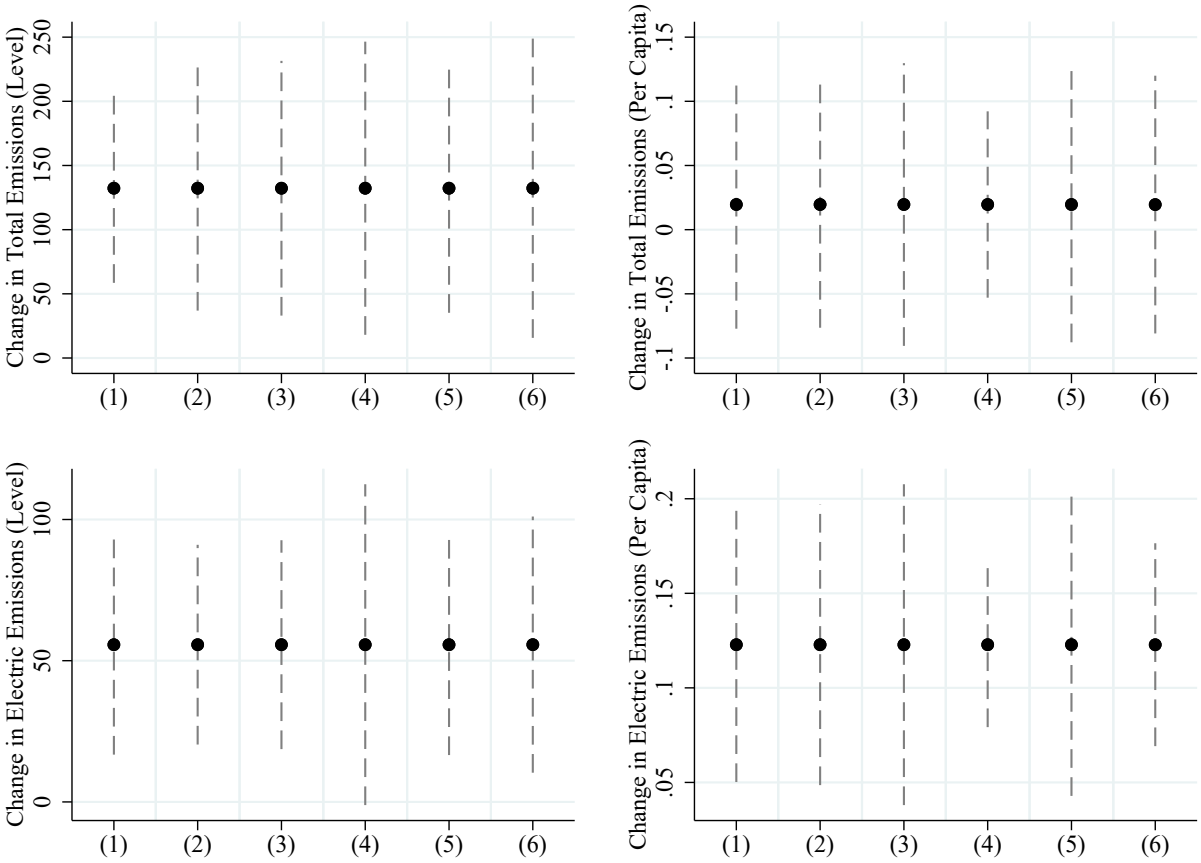
Notes: The sample consists of 722 CZs observed over four five-year periods. The dependent variable is the change in CO_2 emissions, either total or from electricity generation, measured in levels or per 1,000 individuals aged 16 and older, as specified in the column headers. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state×year fixed effects and the control variables used in column (5) of Table 3. The specifications in panel a) and b) exclude the top and bottom CZs in terms of the average values of ΔE_{ct} and $AIpen_{ct}$, respectively. The specification in panel c) uses pooled 5-year ACS data for the year 2019 in place of pooled 5-year ACS data for 2022. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table B.2: AI Penetration and CO_2 Emissions: Winsorization

	Total (Level) (1)	Electric (Level) (2)	Total (Per Capita) (3)	Electric (Per Capita) (4)	Total (Level) (5)	Electric (Level) (6)	Total (Per Capita) (7)	Electric (Per Capita) (8)
	a) 1% Tails of ΔE_{ct}				b) 2SD from the Mean of ΔE_{ct}			
2nd Stage Regression								
$AIpen_{ct}$	103.876***	31.960*	0.040	0.112***	51.833***	19.473	0.037	0.103***
	(33.592)	(17.740)	(0.042)	(0.037)	(19.514)	(14.373)	(0.041)	(0.034)
1st Stage Regression								
$AIexp_{ct}$	1.969***	1.969***	1.969***	1.969***	1.969***	1.969***	1.969***	1.969***
	(0.271)	(0.271)	(0.271)	(0.271)	(0.271)	(0.271)	(0.271)	(0.271)
Kleibergen-Paap $F-$Statistic	52.94	52.94	52.94	52.94	52.94	52.94	52.94	52.94
Obs.	2,888	2,888	2,888	2,888	2,888	2,888	2,888	2,888
	c) 1% Tails of $AIpen_{ct}$				d) 2SD from the Mean of $AIpen_{ct}$			
2nd Stage Regression								
$AIpen_{ct}$	176.104***	74.123**	0.026	0.163***	260.329***	109.574**	0.039	0.242***
	(58.100)	(31.601)	(0.079)	(0.060)	(91.370)	(49.552)	(0.117)	(0.091)
1st Stage Regression								
$AIexp_{ct}$	1.479***	1.479***	1.479***	1.479***	1.001***	1.001***	1.001***	1.001***
	(0.186)	(0.186)	(0.186)	(0.186)	(0.161)	(0.161)	(0.161)	(0.161)
Kleibergen-Paap $F-$Statistic	63.07	63.07	63.07	63.07	38.46	38.46	38.46	38.46
Obs.	2,888	2,888	2,888	2,888	2,888	2,888	2,888	2,888

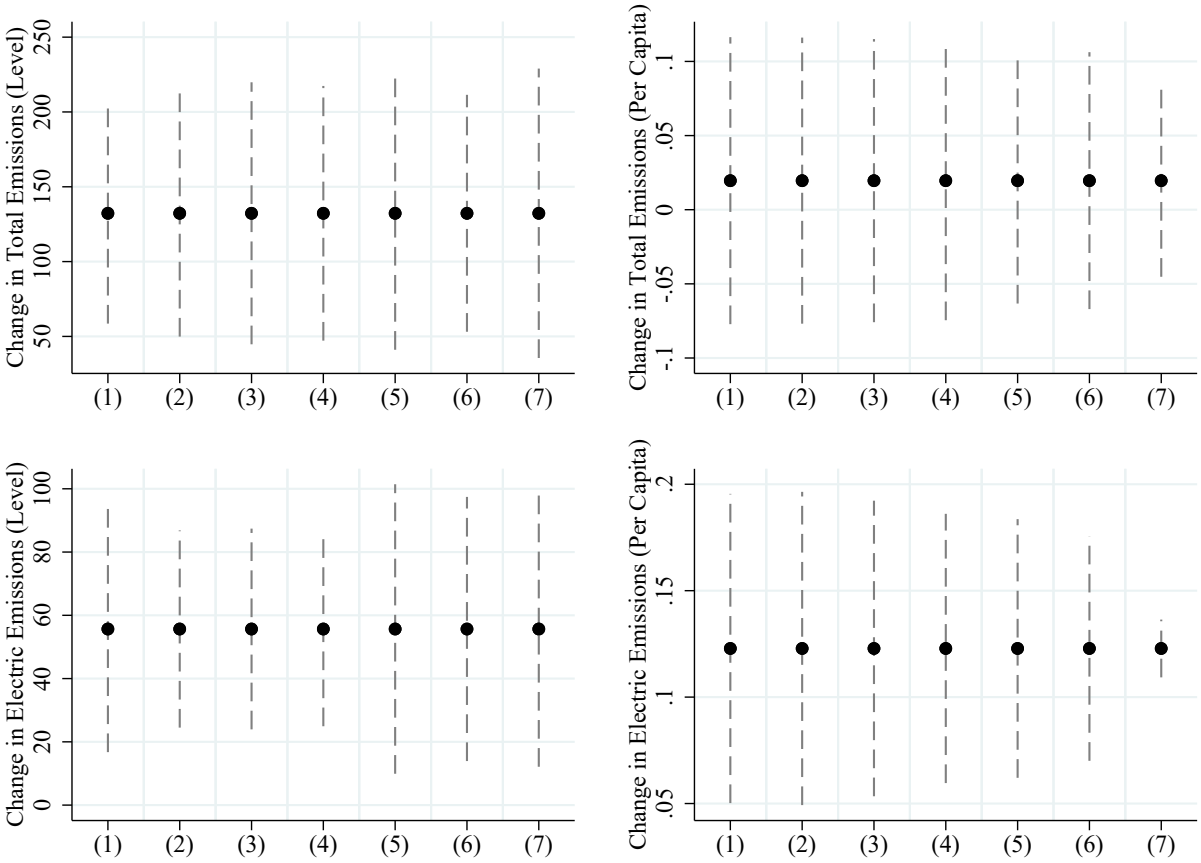
Notes: The sample consists of 722 CZs observed over four five-year periods. The dependent variable is the change in CO_2 emissions, either total or from electricity generation, measured in levels or per 1,000 individuals aged 16 and older, as specified in the column headers. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). The specifications in panels a) and b) winsorize observations of ΔE_{ct} that fall in the 1% tails of the distribution or deviate from the mean by more than two standard deviations. Panels c) and d) apply the same winsorization procedure to observations of $AIpen_{ct}$. All specifications include CZ fixed effects, state×year fixed effects and the control variables used in column (5) of Table 3. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Figure B.1: AI Penetration and CO_2 Emissions: Alternative Clusters



Notes: Circles correspond to the baseline 2SLS coefficient on $AIpen_{ct}$ (see Table 5, columns 1 and 2, and Table 6, columns 2 and 3). Dashed lines indicate 90% confidence intervals based on standard errors computed under alternative clustering schemes: (1) baseline (state-year level); (2) CZ level; (3) state level; (4) two-way clustering by CZ and year; (5) two-way clustering by CZ and state-year. Confidence interval (6) is based on the inference approach proposed by Borusyak et al. (2022).

Figure B.2: AI Penetration and CO_2 Emissions: Spatial Correlation



Notes: Circles correspond to the baseline 2SLS coefficient on $AIpen_{ct}$ (see Table 5, columns 1 and 2, and Table 6, columns 2 and 3). Dashed lines indicate 90% confidence intervals based on standard errors corrected for arbitrary spatial correlation following Colella et al. (2023). The spatial cluster of a given CZ is defined as the set of CZs that belong to the top 5th (2), 10th (3), 25th (4), 50th (5), 75th (6) or 90th (7) percentiles of the bilateral distance distribution. Confidence interval (1) is the baseline one, based on standard errors corrected for clustering by state-year.

Figure B.2 displays confidence intervals that account for arbitrary spatial correlation in the residuals across CZs within the same geographic cluster. These confidence intervals are computed using the method proposed by Conley (1999) and its extension to 2SLS settings developed by Colella et al. (2023). For each CZ, the spatial cluster is defined as the set of CZs that belong to the top 5th, 10th, 25th, 50th, 75th or 90th percentiles of the bilateral distance distribution. Reassuringly, our main conclusions remain unchanged across all spatial correlation thresholds.

B.3 Variable Definitions and Model Specifications

In Table B.3, we use alternative definitions of data-intensive occupations in the construction of $AIpen_{ct}$ and $AIexp_{ct}$. We begin by addressing the concern that our baseline definition may be overly broad. To address this, panel a) restricts the definition to a single occupation, "Data Scientists", whose tasks are closely aligned with the core domains of AI.[16] This definition is likely to largely understate the true extent of AI penetration, yet the resulting patterns are qualitatively consistent with those based on the broader baseline definition.

In panels b) and c), we adopt two entirely different classifications of data-intensive occupations, following Hanson (2022) and Eloundou et al. (2024). The former identifies five AI-related occupations: "Computer Scientists and System Analysts", "Computer Software Engineers", "Network Systems and Data Communication Analysts", "Statisticians" and "Computer Hardware Engineers".[17] The latter uses a specific rubric, coupled with human expertise and GPT-4 classifications, to measure occupations' alignment with large language models. We identify as data-intensive those occupations falling in the top decile of this indicator. The results are confirmed in both cases.

In Table B.4, we explore the robustness of the results to alternative definitions of the outcome variables; we redefine per-capita CO_2 emissions using working-age population and labor force instead of total population. This addresses the concern that total population may not adequately capture relevant demographic trends. In both cases, the results are largely unchanged. Table B.5 examines the sensitivity of the results to the use of a different time aggregation. Specifically, we estimate eq. (1) using two stacked differences based on 10-year changes (2002-2012 and 2012-2022), rather than four

[16] According to the SOC, Data Scientists "develop and implement a set of techniques or analytics applications to transform raw data into meaningful information using data-oriented programming languages and visualization software datasets. Visualize, interpret and report data findings. May create dynamic data reports".

[17] Hanson (2022) defines AI-related occupations as those in STEM fields that are not administrative or supportive in nature, are not tied to scientific disciplines unrelated to AI, and whose Census-defined job titles contain at least one term from the set "designer/design", "researcher/research", "scientist" or "statistician/statistical", and one from the set "computer", "data" or "software". The corresponding occupational codes are: 1510XX, 151030, 151081, 152041 and 172061.

Table B.3: AI Penetration and CO_2 Emissions: Alternative Definitions of Data-Intensive Occupations

	Total (Level) (1)	Electric (Level) (2)	Total (Per Capita) (3)	Electric (Per Capita) (4)
	a) Data Scientists			
2nd Stage Regression				
$AIpen_{ct}$	399.047**	229.284**	0.289	0.216
	(162.700)	(100.709)	(0.195)	(0.199)
1st Stage Regression				
$AIexp_{ct}$	3.155***	3.155***	3.155***	3.155***
	(0.276)	(0.276)	(0.276)	(0.276)
Kleibergen-Paap $F-$Statistic	130.51	130.51	130.51	130.51
Obs.	2,888	2,888	2,888	2,888
	b) Hanson (2022)			
2nd Stage Regression				
$AIpen_{ct}$	93.802***	49.468**	0.051	0.073**
	(28.981)	(19.403)	(0.034)	(0.036)
1st Stage Regression				
$AIexp_{ct}$	2.305***	2.305***	2.305***	2.305***
	(0.233)	(0.233)	(0.233)	(0.233)
Kleibergen-Paap $F-$Statistic	98.13	98.13	98.13	98.13
Obs.	2,888	2,888	2,888	2,888
	c) Eloundou et al. (2024)			
2nd Stage Regression				
$AIpen_{ct}$	91.613**	49.937**	-0.005	0.084**
	(35.446)	(22.212)	(0.070)	(0.037)
1st Stage Regression				
$AIexp_{ct}$	1.283***	1.283***	1.283***	1.283***
	(0.243)	(0.243)	(0.243)	(0.243)
Kleibergen-Paap $F-$Statistic	27.86	27.86	27.86	27.86
Obs.	2,888	2,888	2,888	2,888

Notes: The sample consists of 722 CZs observed over four five-year periods. The dependent variable is the change in CO_2 emissions, either total or from electricity generation, measured in levels or per 1,000 individuals aged 16 and older, as specified in the column headers. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state$\times$year fixed effects and the control variables used in column (5) of Table 3. Data-intensive occupations are defined as: 'Data Scientists' in panel a); the five AI-related occupations from Hanson (2022) in panel b); and occupations falling in the top decile of the continuous measure proposed by Eloundou et al. (2024) in panel c). All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table B.4: AI Penetration and CO_2 Emission: Alternative Normalizations of Per-Capita Emissions

	Total (Per Capita)	Electric (Per Capita)	Total (Per Capita)	Electric (Per Capita)
	Population 16-65		Labor Force	
	(1)	(2)	(3)	(4)
2nd Stage Regression				
$AIpen_{ct}$	0.018	0.127***	-0.016	0.114**
	(0.062)	(0.046)	(0.067)	(0.049)
1st Stage Regression				
$AIexp_{ct}$	1.969***	1.969***	1.969***	1.969***
	(0.271)	(0.271)	(0.271)	(0.271)
Kleibergen-Paap $F-$Statistic	52.94	52.94	52.94	52.94
Obs.	2,888	2,888	2,888	2,888

Notes: The sample consists of 722 CZs observed over four five-year periods. The dependent variable is the change in CO_2 emissions, either total or from electricity generation, relative to working-age population (1,000 individuals aged 16-65) in columns (1)-(2) and 1,000 individuals in the labor force in columns (3)-(4). The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state×year fixed effects and the control variables used in column (5) of Table 3. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

stacked five-year differences. If anything, the estimates based on longer differences are even stronger.

C Sensitivity of Inference to Violations of Exclusion Restriction

In this Appendix, we briefly describe how the approach of Conley et al. (2012) can be implemented within our empirical framework. Consider the following version of eq. (1):

$$\Delta E_{ct} = \alpha_c + \alpha_{st} + \beta\, AIpen_{ct} + \mathbf{X}'_{ct}\gamma + \lambda\, AIexp_{ct} + \varepsilon_{ct}, \tag{C.1}$$

where λ captures the extent to which the exclusion restriction is violated. The baseline results in the main text rely on the standard IV assumption that $\lambda = 0$. However, if the exclusion restriction is not satisfied, i.e., if $\lambda \neq 0$, it is still possible to conduct inference on β by introducing priors on λ and conditioning on its value. This is achieved by estimating the following specification using 2SLS:

$$\Delta E_{ct} - \lambda\, AIexp_{ct} = \alpha_c + \alpha_{st} + \beta\, AIpen_{ct} + \mathbf{X}'_{ct}\gamma + \varepsilon_{ct}, \tag{C.2}$$

Table B.5: AI Penetration and CO_2 Emission: 10-Year Differences

	Total (Level) (1)	Electric (Level) (2)	Total (Per Capita) (3)	Electric (Per Capita) (4)
2nd Stage Regression				
$AIpen_{ct}$	207.729***	48.755***	0.082**	0.127***
	(51.820)	(14.551)	(0.038)	(0.031)
1st Stage Regression				
$AIexp_{ct}$	2.145***	2.145***	2.145***	2.145***
	(0.364)	(0.364)	(0.364)	(0.364)
Kleibergen-Paap $F-$Statistic	34.63	34.63	34.63	34.63
Obs.	1,444	1,444	1,444	1,444

Notes: The sample consists of 722 CZs observed over two ten-year periods. The dependent variable is the change in CO_2 emissions, either total or from electricity generation, measured in levels or per 1,000 individuals aged 16 and older, as specified in the column headers. The variables $AIpen_{ct}$ and $AIexp_{ct}$ measure AI penetration and AI exposure, respectively, as defined in eq. (2) and (3). All specifications include CZ fixed effects, state×year fixed effects and the control variables used in column (5) of Table 3. All regressions are weighted by each CZ's initial-period share of the total population. Standard errors, reported in parentheses, are clustered at the state-year level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

where $AIpen_{ct}$ is instrumented with $AIexp_{ct}$. By varying the prior about λ, one can assess how sensitive the estimated effect of AI penetration (β) is to different degrees of violation of the exclusion restriction. Because the sensitivity of the 2SLS estimator to such violations is inversely related to the strength of the instrument, a stronger first-stage relationship implies a smaller loss of precision for a given value of λ.

We set λ to be a function of a parameter δ, which we incrementally increase (in steps of 0.01, starting from 0) to generate progressively larger violations of the exclusion restriction. We set $\lambda = \beta \times 2.2 \times \delta$, where β is the baseline 2SLS estimate and the factor 2.2 reflects the ratio of the standard deviation of $AIpen_{ct}$ to that of $AIexp_{ct}$. For each value of λ, we compute confidence intervals for β corresponding to both endpoints of the support $[-\lambda, \lambda]$. The resulting confidence interval is then constructed as the union of the two confidence intervals.[18]

[18] In addition to this "union of confidence intervals" approach, Conley et al. (2012) discuss alternative methods that incorporate stronger priors about λ. These methods typically impose additional parametric restrictions, yielding narrower and thus less conservative confidence intervals.