The Impact of Export Opportunities on AI Innovation: Evidence from Indian Firms

Despoina Balouktsi* Laura Contreras-Portela† Priyam Verma‡

This version: September 7, 2026

Abstract

Developing artificial intelligence (AI) technologies requires access to data, yet little is known about the economic forces that shape firms' access to commercially valuable data. This paper studies whether access to foreign markets contributes to AI innovation by expanding firms' access to data generated through interactions with foreign customers. We combine firm-level financial information from CMIE Prowess, patent records from the Indian Patent Office, and newly developed crosswalks linking firms' products and services to detailed trade classifications. Exploiting plausibly exogenous variation in foreign demand through a shift-share design, we find that stronger export opportunities increase both the probability and intensity of AI patenting and are accompanied by higher expenditure on data-related activities. To examine whether commercially generated data constitute the underlying mechanism, we exploit the implementation of the European Union's General Data Protection Regulation (GDPR), which increased the costs of collecting and processing customer data while leaving firms' access to foreign markets largely unchanged. Firms with greater pre-existing exposure to the European Union experienced a decline in AI patenting intensity following the GDPR, despite showing little evidence of changes in exports or broader firm performance. Our findings indicate that commercially generated data are an important mechanism through which access to foreign markets fosters AI innovation.

Keywords: data, innovation, artificial intelligence, export.

*Joint Research Center. Email: despoina.BALOUKTSI@ec.europa.eu.
†Aix-Marseille School of Economics, Aix-Marseille Université. Email: laura.contreras-portela@univ-amu.fr.
‡Department of Economics, Ashoka University. Email: priyam.verma@ashoka.edu.in

1 Introduction

Artificial intelligence (AI) is rapidly reshaping production, innovation, and economic growth. At the heart of this transformation lies an increasingly valuable economic input: data. Recent advances in machine learning have dramatically increased the returns to collecting, processing, and exploiting data (Carrière-Swallow and Haksar, 2019). Unlike many previous technologies, AI systems improve through exposure to large, continuously updated datasets, making access to commercially valuable data a key determinant of frontier innovation (World Trade Organization, 2024; Agrawal et al., 2026; Jones and Tonetti, 2020). Yet, despite the growing importance of data, relatively little is known about the economic forces that determine how firms acquire the data needed to develop AI technologies.

One important source of commercially valuable data is firms' relationships with customers. Every commercial relationship generates more than revenue. Firms continuously interact with customers before, during, and after the delivery of goods and services, generating information about customer preferences, product performance, operational processes, and market-specific demand. Unlike physical capital or R&D expenditures, these data are jointly produced through commercial activity itself and, because they are non-rival and can often be reused across projects within the firm (Jones and Tonetti, 2020; Beraja et al., 2022), become a reusable stock of data that supports future AI development.

Export relationships extend firms' interactions beyond domestic markets, providing one important channel through which they accumulate commercially generated data from foreign customers. This suggests a different mechanism linking exports and innovation. While the existing literature emphasizes larger markets, stronger competitive pressures, learning-by-exporting, and knowledge spillovers as the main channels through which exports stimulate innovation (Bloom et al., 2016; Bombardini et al., 2017; Ahn et al., 2018; Autor et al., 2020; Akcigit and Melitz, 2022; Aghion et al., 2023, 2024), we ask whether exports also foster AI innovation by enabling firms to accumulate this reusable stock of consumer-generated data.

We answer this question using firm-level data from India, one of the world's largest exporters of ICT and digitally delivered services. We construct a novel dataset that combines financial information from the CMIE Prowess database, patent applications from the Indian Patent Office, and newly developed product and service crosswalks linking firms' commercial activities to internationally standardized trade classifications. The resulting dataset enables us to measure AI innovation at the firm level and to combine firms' innovation activities with detailed information on their export portfolios and destination markets.

Our empirical strategy combines two complementary sources of plausibly exogenous variation. First, we exploit firm-level export demand shocks, constructed from firms' predetermined export portfolios and changes in foreign demand across destination markets, to identify the effect of improved export opportunities on AI innovation. Second, we exploit the implementation of the European Union's General Data Protection Regulation (GDPR), which increased the costs of collecting and processing customer data while leaving firms' access to foreign markets largely unchanged. This second strategy allows us to test whether commer-

cially generated data are the mechanism through which export opportunities foster AI innovation.

Our analysis delivers three main findings. First, improved export opportunities increase firms' innovative activity, with particularly strong effects on AI innovation. A one-standard-deviation increase in the export demand shock raises the probability of filing an AI patent by 0.24 percentage points and increases firms' AI patent stock by approximately 5%. We also find positive effects on ICT and overall patent stocks, consistent with the broader literature documenting positive effects of exports on innovation (Aghion et al., 2024; Akcigit and Melitz, 2022).

Second, we examine how firms respond to improved export opportunities. The same foreign demand shocks are accompanied by greater investment in data-related activities. A one-standard-deviation increase in foreign demand raises the probability of exporting by 12.5 percentage points, export sales by 2.5%, and total sales by 2.9%. Most notably, firms increase expenditure on software, data centres, communications, web hosting, and other IT-enabled services by approximately 5.2%. These patterns suggest that firms respond to improved export opportunities not only by expanding production but also by investing in digital capabilities that are consistent with greater data collection, storage, and processing.

Third, we examine the proposed mechanism by exploiting the implementation of the European Union's General Data Protection Regulation (GDPR) in a differences-in-differences framework. We find that firms with greater pre-existing exposure to the European Union experienced a 0.9% decline in AI patenting intensity following the introduction of the regulation, despite showing little evidence of changes in other innovation activities, exports, sales, productivity, or broader firm performance. At the same time, these firms allocated a larger share of their total expenditures to data-related activities, indicating that data-related activities became relatively more costly after the GDPR. Because the regulation primarily governs the collection and processing of personal data rather than firms' access to foreign markets, this evidence is consistent with commercially generated customer data being an important mechanism linking exports to AI innovation. These findings underscore that the regulation of data may shape firms' innovative activity through channels that differ from conventional trade policies. Our results are robust to a wide range of alternative specifications, including placebo and falsification tests, alternative constructions of the export demand shock, controls for firms' initial export composition, balanced-panel specifications, and leave-one-out exercises.

Our findings also have broader policy implications. As governments increasingly regulate the collection, processing, and use of data while simultaneously promoting AI development, understanding the interaction between data governance and trade becomes increasingly important. Our results suggest that data regulation may influence firms' innovative activity through channels that are distinct from conventional trade barriers. These findings should not be interpreted as a welfare evaluation of privacy regulation, which provides important benefits by protecting consumers' personal information and strengthening trust in digital markets. Instead, they highlight that policies governing data can shape firms' incentives to develop frontier AI technologies even when access to foreign markets remains unchanged.

We acknowledge several limitations. First, our analysis focuses on AI innovation, as measured by patenting activity, rather than the adoption and diffusion of AI technologies within firms. While innovation determines the development of new AI capabilities, understanding how data availability shapes firms' incentives to adopt existing AI systems remains an important question for future research. Second, although the GDPR provides a useful source of variation in the costs of collecting and processing customer data, we do not directly observe the quantity, quality, or composition of the data firms acquire. Instead, we infer the role of data from firms' innovation responses and changes in data-related expenditure. Consequently, we cannot distinguish whether firms responded by investing more in compliance and digital infrastructure, relying more heavily on specialized providers of data services, purchasing third-party datasets, or shifting toward alternative sources of training data. Finally, our empirical strategy identifies the role of customer-generated data arising from firms' relationships with foreign customers. Whether supplier-generated data arising from firms' relationships with suppliers and production networks similarly contribute to AI innovation remains an important open question. Addressing this question would require detailed firm-level information on imported intermediate inputs and their shares in firms' production processes to construct plausibly exogenous import supply shocks. Because our data observe firms' outputs but not the composition of their imported inputs, we cannot credibly measure firms' exposure to foreign supply shocks or isolate the contribution of supplier-generated data.

Related literature. This paper contributes to three strands of literature.

First, we contribute to the literature on exports and innovation. A large body of work shows that firms engaged in international markets are more innovative and studies how exports foster innovation through larger market size, stronger competitive pressures, learning-by-exporting, and knowledge spillovers (Bloom et al., 2016; Bombardini et al., 2017; Ahn et al., 2018; Autor et al., 2020; Coelli et al., 2022; Aghion et al., 2023, 2024; Akcigit and Melitz, 2022). We complement this literature by identifying an additional channel through which export relationships affect innovation. Rather than focusing on how exports change firms' incentives or opportunities to innovate, we examine whether they enable firms to accumulate customer-generated data that subsequently support AI innovation. We also contribute to the emerging empirical literature on AI innovation, for which firm-level evidence remains limited, particularly in developing economies.

Second, the paper contributes to the growing literature that studies data as an economic input in production and innovation. Recent theoretical work incorporates data into models of production, innovation, and international trade, emphasizing its non-rival nature and its role in generating economic value (Jones and Tonetti, 2020; Farboodi and Veldkamp, 2021; Cong et al., 2021, 2022; Chang et al., 2023; Chen et al., 2021; Carrière-Swallow and Haksar, 2019). Recent empirical work has begun to examine how firms gain access to economically valuable data and how these data affect production and innovation. Closest to our paper, Beraja et al. (2022) show that firms developing facial-recognition AI obtain valuable training data through government procurement contracts, highlighting how access to data stimulates commercial AI innovation. We complement this literature by studying a different source of data. Rather than

arising from government procurement, the data in our setting are generated through firms' relationships with foreign customers, allowing us to examine how export relationships contribute to AI innovation through the accumulation of customer-generated data. Our paper is also closely related to Demirer et al. (2024), who exploit the implementation of the European Union's General Data Protection Regulation (GDPR) to show that stricter privacy regulation reduced firms' cloud data storage and processing, making firms less data-intensive. Building on this insight, we use the GDPR to isolate the role of customer-generated data in AI innovation. Our findings suggest that access to these data is an important channel through which export opportunities foster AI innovation.

Third, the paper contributes by developing new data resources for studying trade and innovation in India. We construct one of the first large-scale datasets linking firms in CMIE Prowess to patent applications from the Indian Patent Office, allowing us to measure AI and ICT innovation at the firm level. We also develop new product and service crosswalks linking firms' proprietary commercial activity descriptions to internationally standardized HS and EBOPS trade classifications. While recent work has successfully applied machine learning methods to classify standardized product descriptions into Harmonized System (HS) codes (De Artiñano et al., 2025), our setting requires mapping firms' proprietary descriptions of both goods and services into internationally comparable trade classifications. We address this challenge using a scalable pipeline that combines semantic retrieval, cross-encoder reranking, deterministic rules, and large language model adjudication. The resulting crosswalk enables the integration of CMIE with customs statistics, tariff schedules, foreign demand measures, and other internationally comparable datasets. In this paper, it is central to constructing product-level export demand shocks and linking firms' commercial activities to their innovation portfolios, while also providing a broadly applicable resource for future research on Indian firms.

The rest of the paper is organized as follows. Section 2 provides institutional background and discusses why India offers a suitable setting for studying the relationship between international trade and AI innovation. Section 3 describes the data sources, the construction of the linked firm-level dataset, and the product and service crosswalk. Section 4 presents the empirical strategy. Section 5 reports the main results on the relationship between exports and AI innovation. Section 6 examines the role of commercially generated data using the implementation of the General Data Protection Regulation (GDPR) as a source of exogenous variation in firms' access to data. Section 7 discusses the implications and limitations of the findings. Section 8 concludes.

2 Background and Institutional Context

This section describes the institutional context underlying our analysis. We first provide background on India's economic transformation, emphasizing its integration into global markets, the expansion of knowledge-intensive service exports, and the emergence of a rapidly growing AI ecosystem. We then discuss the European Union's General Data Protection Regulation (GDPR), outlining its scope, compliance requirements, enforcement mechanisms, and extrater-

ritorial application.

2.1 The Indian Context: Global Integration and Innovation

Following a balance-of-payments crisis in 1991, the Indian government implemented a broad package of market-oriented reforms that reduced trade barriers, eased restrictions on foreign investment, and liberalized industrial licensing. These reforms accelerated the country's integration into global markets and contributed to the expansion of internationally oriented firms.

International trade has grown substantially over recent decades. The nominal value of Indian goods exports rose from USD 121 billion in 2002 to USD 461 billion in 2024, an increase of 280%. By 2024, the United States and the United Arab Emirates were India's two largest export destinations for goods, followed by China, the United Kingdom, and Germany. Notably, five European Union countries (Germany, the Netherlands, France, Italy, alongside the United Kingdom, which remained under GDPR through the transition period) appeared among India's top fifteen goods destinations, jointly accounting for over USD 60 billion in exports—a substantial shift from the 2011 pattern documented in prior work (Barrows and Ollivier, 2021), when Asian and Gulf economies dominated the top rankings. The transformation is even more pronounced in services trade, where India has emerged as a major global exporter of information technology, professional, and business services. The European Union has consistently absorbed between 24 and 29 percent of India's services exports over the past two decades. The United States accounted for approximately 21 percent of Indian services exports in 2024.

Another notable feature of India's development has been the expansion of the service sector. Unlike many emerging economies whose structural transformation was driven primarily by manufacturing, services became the largest contributor to GDP, foreign direct investment, and export growth (Mukherjee, 2013). Reforms in service industries have also been shown to generate important spillovers to downstream manufacturing firms by increasing product variety and improving firm performance (Mulyukova, 2025). Within this sector, information and communications technology (ICT), software, business process outsourcing, and other digitally delivered services have grown rapidly, making India one of the world's leading exporters of knowledge-intensive services.

India has also become an increasingly important source of AI research and innovation. According to Baruffaldi et al. (2020), India's share of global AI publications increased from 3.4% in 2006–08 to nearly 11% in 2016–18, placing it among the world's leading producers of AI research. During the same period, India ranked fourth globally in highly cited AI publications, surpassing several advanced economies, including the United Kingdom and France. AI-related patenting has also expanded rapidly: the contribution of Indian inventors to AI-related patent families increased from 0.2% in 2004–06 to 2.2% in 2014–16, and the share of AI patents in India's overall patent portfolio rose from 0.4% to around 6% over the same period. Additionally, India's patent system has expanded considerably over the sample period. Resident filings accounted for more than 55% of total patent applications in FY24, and India has become the world's sixth-largest patent office by filing volume (World Intellectual Prop-

erty Organization, 2025; NASSCOM, 2025). This institutional development coincides with the rapid expansion of AI-related patenting documented above.

These developments provide the institutional context for studying the relationship between international trade and AI innovation in India.

2.2 The General Data Protection Regulation (GDPR)

The General Data Protection Regulation (GDPR) is the European Union's comprehensive legal framework governing the collection, processing, storage, and transfer of personal data. Adopted in 2016 and enforceable from May 25, 2018, following a two-year transition period, the GDPR replaced the fragmented system established under Directive 95/46/EC with a harmonized set of rules directly applicable across all EU member states (Johnson, 2022; Demirer et al., 2024). Owing to its broad scope and influence, the GDPR is widely regarded as one of the most consequential privacy regulations adopted to date and has shaped data protection legislation well beyond Europe.

A defining feature of the GDPR is its extraterritorial scope. The regulation applies not only to organizations established within the European Union but also to firms located elsewhere whenever they offer goods or services to individuals in the EU or process their personal data (Johnson, 2022; Aridor et al., 2020). Consequently, many firms operating outside Europe may be required to comply with its provisions despite having no physical presence within the EU. Importantly, this extends well beyond the technology sector: manufacturers, retailers, business-to-business suppliers, exporters, and providers of professional services may all fall within the scope of the regulation if they process the personal data of individuals located in the European Union.

The GDPR adopts a data protection approach rather than a consumer protection approach, imposing affirmative obligations on firms that collect or process personal information (Boyne, 2018; Jones and Tonetti, 2020; Demirer et al., 2024). It defines personal data broadly to include any information relating to an identified or identifiable individual, ranging from names and email addresses to online identifiers and IP addresses. The regulation also strengthens the rights of data subjects by granting rights of access, rectification, erasure (the "right to be forgotten"), portability, and restrictions on data processing.

From the perspective of firms, compliance requires a wide range of organizational and technical measures. Depending on the nature of their activities, firms may be required to maintain detailed records of processing activities, conduct data protection impact assessments, implement appropriate technical and organizational safeguards, report qualifying data breaches, respond to requests from data subjects, and appoint dedicated Data Protection Officers (Demirer et al., 2024). Rather than prescribing a uniform set of compliance standards, the GDPR follows a risk-based approach under which firms are expected to adopt measures commensurate with the scale and sensitivity of their data processing activities (Hustinx, 2017; Gellert, 2018).

These requirements can generate substantial compliance costs. Survey evidence summarized by Demirer et al. (2024) suggests that firms incur expenditures on legal services, cyber-

security infrastructure, documentation systems, employee training, outside consulting, and specialized personnel, with costs varying according to firm size and the complexity of their data operations. Compliance is overseen by independent Data Protection Authorities (DPAs) in each EU member state, which may investigate complaints, conduct audits, issue warnings or corrective orders, and impose administrative sanctions. Under Article 83 of the GDPR, fines can reach up to 20 million euros or 4% of a firm's worldwide annual turnover, whichever is greater. Public enforcement actions show that these powers have been exercised across a broad range of industries and firm sizes, affecting both multinational corporations and smaller organizations (Demirer et al., 2024).

The GDPR is particularly relevant in the Indian context because many firms maintain commercial relationships with customers and business partners in the European Union. Through its extraterritorial application, the regulation affects not only software developers and information technology service providers but also exporters of manufactured goods and other firms engaged in cross-border commerce whenever they process the personal data of individuals located in Europe. The GDPR, therefore, changed the legal environment governing data processing for a broad set of Indian firms participating in international markets.

An important feature of the GDPR for this paper is that it regulates the collection, processing, storage, and use of personal data rather than international trade itself. The regulation does not impose tariffs, quotas, or restrictions on firms' ability to export goods or services to the European Union, nor does it prohibit cross-border commercial relationships. Instead, it establishes legal obligations and compliance requirements governing how firms manage personal data generated in the course of those activities. While complying with these requirements may increase the costs of serving foreign customers—for example, through investments in legal compliance, cybersecurity, or data governance—the primary object of the regulation is the use of data rather than firms' access to foreign markets. Existing evidence similarly characterizes the GDPR as a data governance regulation rather than a trade barrier (World Trade Organization, 2024).

This distinction is central to our analysis. The GDPR changed the legal framework governing the collection, processing, and storage of personal data without directly changing tariffs, market access conditions, or firms' legal ability to export to the European Union. This institutional setting allows us to examine whether changes in the regulatory environment governing commercially generated data are associated with changes in AI innovation. At the same time, compliance with the GDPR may have increased the costs of serving customers in the European Union, providing an alternative channel through which the regulation could affect firms' innovation decisions. As a first piece of descriptive evidence, aggregate Indian exports do not exhibit an obvious decline or reallocation across destination markets following the implementation of the GDPR (see Appendix A). Nevertheless, we explicitly examine the effects of the regulation on firms' export activity and broader economic performance to assess whether our results operate through changes in data processing or through changes in the costs of serving foreign markets.

3 Exporters and Innovators: Data and Descriptive Statistics

This section describes the data used in the analysis and presents descriptive evidence on the relationship between exports and AI innovation. We first introduce the different data sources and explain how they are combined to construct our firm-level panel. We then present summary statistics and stylized facts that motivate the empirical analysis.

3.1 Data

Our analysis combines production and balance-sheet information, patent records, product descriptions, and international trade data for Indian firms over the period 1990–2019. These data allow us to construct firm-level measures of innovation and detailed measures of firms' export activities. They also provide the information required to build firm-specific foreign demand shocks based on the products firms export and the destinations they serve.

To assemble the dataset, we make two main data contributions. First, we construct a new database linking Indian firms to their patenting activity and classify patents according to their technological content. Second, we develop a large-scale crosswalk that maps the proprietary product descriptions reported in the Centre for Monitoring Indian Economy (CMIE) to internationally standardized Harmonized System (HS6) and Extended Balance of Payments Services (EBOPS) classifications using a combination of deterministic rules and large language models (LLMs). These linkages allow us to connect firms' innovation activities with detailed information on the goods and services they produce and export.

Firm-level data: CMIE. Our primary source of firm-level information is the Prowess database compiled by the Centre for Monitoring the Indian Economy (CMIE), the leading source of microdata on India's organized corporate sector. We use annual data covering the period 1990–2024.

PROWESS combines audited financial statements for publicly listed firms with detailed surveys of privately held companies, providing broad coverage across both manufacturing and service industries. Although it does not capture the informal sector, the database accounts for more than 70% of economic activity in the organized industrial sector and approximately 75% (95%) of corporate (excise duty) tax revenues collected by the Indian government (Goldberg et al., 2010). For each firm-year, PROWESS reports detailed information from firms' balance sheets and income statements, including sales, exports, imports, value added, production costs, wages, capital expenditures, profits, assets, liabilities, and ownership characteristics.[1] As a result, it has become the standard data source for firm-level research on the Indian economy.

Several features of the database are particularly useful for our analysis. First, it records firms' export activity, which allows us to construct measures of exposure to foreign demand. Second, it provides detailed textual descriptions of the goods and services produced by each firm. We use these descriptions to map firms' products into internationally standardized Har-

[1] Monetary variables are reported in Indian Rupees (INR) and are deflated using the Wholesale Price Index (WPI), applying industry-specific deflators whenever available.

monized System (HS6) and Extended Balance of Payments Services (EBOPS) classifications, enabling us to link firms to product-level trade data. Finally, PROWESS reports expenditures on information and communications technology (ICT) services, including software development, internet services, data centers and web hosting, communications, satellite links, and other IT-enabled services.

Firm-level patent data. To measure technological innovation, we complement the PROWESS database with a new firm-level patent dataset constructed from records of the Indian Patent Office (IPO). Although patent records are publicly available, no existing source systematically links patent applicants to firms in PROWESS.

We compile the underlying data through a large-scale web-scraping exercise covering patent applications filed between 2000 and 2025. For each application, we collect information on applicant names, inventors, filing and publication dates, titles, abstracts, and International Patent Classification (IPC) codes. We then clean and standardize applicant names and match them to firms in Prowess using a combination of exact matching, fuzzy string matching, and manual verification for ambiguous cases. Starting from 286,594 patent applications, this procedure identifies 56,853 applications linked to CMIE firms. To ensure a high degree of accuracy, we retain only matches with a similarity score of at least 80% and manually review uncertain cases.

We classify patents using the methodology developed by Baruffaldi et al. (2020), which combines IPC codes with validated keyword-based criteria to identify artificial intelligence (AI) inventions. We apply the same approach to identify patents related to information and communications technologies (ICT), allowing us to distinguish AI-specific innovation from broader digital technologies. Appendix B.1 provides additional details on the data collection process, the matching algorithm, the patent classification procedure, and validation exercises.

Our main innovation measures are indicators equal to one if a firm files at least one AI-related or ICT-related patent application in a given year. We also construct cumulative patent stocks and indicators for whether a firm ever patents in each category. The annual filing indicators are our preferred outcomes because they most closely capture the timing of innovative activity.

CMIE products to HS6–EBOPS classifications. Our empirical strategy exploits product-level variation in foreign demand. Constructing firm-level foreign demand shocks, therefore, requires linking firms' products to internationally standardized trade classifications. However, the product-level information contained in PROWESS is not reported using such classifications. Instead, CMIE records highly detailed proprietary textual descriptions of the goods and services produced by each firm. While these descriptions provide a granular view of firms' production activities, they cannot be directly linked to Harmonized System (HS) product codes or Extended Balance of Payments Services (EBOPS) classifications. This limitation has made it difficult to combine PROWESS with customs records, tariff schedules, and other product-level international trade datasets.

To address this issue, we construct a new LLM-assisted crosswalk that maps CMIE's proprietary product descriptions to internationally standardized trade classifications: six-digit

Harmonized System (HS6) codes for goods and Extended Balance of Payments Services (EBOPS) categories for services. Merchandise trade is internationally classified at the highly disaggregated HS6 level, while trade in services is reported using the EBOPS classification, the international statistical standard for bilateral services trade. Our baseline uses the most detailed publicly available EBOPS level (Level 2), allowing us to consistently link firms' service activities to international trade data.

The classification procedure proceeds in four stages. We first standardize product descriptions by removing formatting inconsistencies, expanding common abbreviations, correcting spelling variations, and harmonizing punctuation and capitalization. Next, we use semantic retrieval to identify a small set of candidate HS6 or EBOPS codes whose official descriptions are most similar to the cleaned CMIE description. These candidates are then reranked using a cross-encoder model that jointly evaluates the product description, candidate trade codes, and industry context. Finally, a large language model selects the most appropriate classification and reports a confidence score together with an explanation. Observations with low confidence or ambiguous classifications are flagged for additional review. This approach is particularly well suited to CMIE data, where product descriptions frequently consist of abbreviations, accounting labels, firm-specific terminology, or incomplete phrases that are difficult to classify using conventional dictionary-based methods.

The resulting crosswalk achieves broad coverage while preserving economically meaningful product heterogeneity.[2] Approximately 65% of eligible goods descriptions and 80% of service descriptions receive internationally comparable classifications, with even higher effective coverage when observations are weighted by sales, indicating that unmatched products are concentrated among low-revenue activities (Table A4). The methodology also allows identical textual descriptions to receive different classifications depending on industry context while recognizing that different descriptions may refer to the same underlying traded product.

Compared with conventional industry-based concordances (Goldberg et al., 2010), the crosswalk recovers substantially richer product-level information. A standard NIC4-based mapping identifies only 47 distinct HS6 products across the sample, whereas our approach identifies 3,391 distinct HS6 products. For services, mapping firms' activities to EBOPS provides an internationally comparable classification that can be directly linked to bilateral services trade data, substantially improving upon the absence of any standardized product-level correspondence.

The resulting classifications enable us to merge PROWESS with CEPII BACI merchandise trade data and OECD–WTO BaTIS services trade data, construct firm-level foreign demand shocks based on the products and services firms produce, and place firms' products and patents into a common internationally comparable classification system. These linkages are central to our empirical strategy, allowing us to measure firms' exposure to changes in foreign demand across both goods and services and examine how export opportunities affect AI inno-

[2]The underlying data contain more than 2 million firm–year–product observations, corresponding to 181,439 unique product–industry combinations and approximately 103,000 distinct textual descriptions classified through our pipeline.

vation. Additional details on the methodology, validation exercises, and coverage diagnostics are provided in Appendix B.2.

Trade data and foreign demand shocks. We complement the firm-level information from PROWESS with bilateral merchandise and services trade data. For goods, we use the CEPII BACI database (Gaulier and Zignago, 2010), which reports annual bilateral trade flows by origin, destination, and six-digit Harmonized System (HS6) product. For services, we use the OECD–WTO Balanced Trade in Services (BaTIS) database, which reports bilateral trade flows by Extended Balance of Payments Services (EBOPS) category. Together, these datasets provide comprehensive coverage of India's international trade in both goods and services and allow us to track changes in foreign demand across products, service categories, and destination markets over time.

Combining these trade data with our CMIE–HS6/EBOPS crosswalk enables us to link firms' production portfolios to internationally comparable trade classifications and to measure their exposure to foreign markets at a highly disaggregated level. In particular, we construct firm-specific foreign demand shocks based on the goods and services firms produce and the destinations they serve.

The foreign demand shocks capture changes in export opportunities arising from conditions in foreign markets rather than from changes in Indian supply. For each HS6 product and EBOPS service category, we measure demand growth using imports from (or services supplied by) all exporting countries other than India, thereby excluding contemporaneous movements driven by Indian firms themselves. We then aggregate these product- and service-level demand measures to the firm level using predetermined weights based on firms' initial product portfolios and export composition. The resulting shocks therefore reflect changes in external demand interacting with firms' pre-existing specialization rather than endogenous adjustments in firms' production or destination choices.

These firm-level foreign demand shocks constitute the main source of identifying variation in our empirical analysis. Section 4 provides a formal description of their construction and explains how they are used to identify the effects of export opportunities on AI innovation.

3.2 Descriptive Statistics and Stylized Facts

Table A15 reports summary statistics, while Table A14 compares firms by innovation status. Several patterns emerge from the data that motivate the empirical analysis.

Fact 1: AI patenting is concentrated among a small group of firms. AI-related patenting is rare in our sample. Only 48 firms (approximately 0.1% of the sample) ever file an AI patent, indicating that AI patenting is concentrated among a small subset of Indian firms. These firms also differ from the rest of the sample along several observable dimensions. On average, AI innovators are larger, with mean sales of INR 5.6 billion compared with INR 0.9 billion for non-AI firms, are older (37 versus 30 years), and report substantially higher research and development expenditures.

Figure 12 further illustrates these differences by plotting the distribution of value added across groups. The distribution shifts progressively to the right when moving from domestic firms to exporters, ICT innovators, and AI innovators, showing that AI-patenting firms tend to be larger than other firms in the sample.[3]

Fact 2: AI innovators are more likely to export, particularly in services. Firms that file AI-related patents exhibit substantially higher rates of participation in export markets than the rest of the sample. Approximately 83% of AI innovators export, compared with only 40% of firms without AI patents, with the difference being especially pronounced for service exporters. By contrast, conditional on exporting, AI innovators do not appear to have systematically higher export intensity, suggesting that the association is stronger at the extensive margin than at the intensive margin.

Figure 10 further shows that, even within the set of AI innovators, exporting firms tend to be more productive than non-exporters, with the productivity distribution shifted to the right. This highlights the substantial heterogeneity among AI innovators and motivates the need for an empirical strategy that moves beyond simple cross-sectional comparisons.

These descriptive patterns are also reflected in conditional regressions. Table A20 shows that service exports are positively associated with AI innovation after controlling for firm fixed effects, industry-year-location effects, and baseline firm characteristics, whereas goods exports display no comparable relationship. Table A21 further indicates that this association is not purely contemporaneous: lagged service exports are also positively related to subsequent AI patenting, while we find no similar pattern for ICT innovation. Taken together, these results document a robust association between service export activity and AI innovation. However, they should not be interpreted causally, as firms self-select into export markets and unobserved factors may jointly influence exporting and innovation.

Fact 3: AI innovators report substantially higher data-related expenditures. Firms that patent in AI spend considerably more on data-related activities than other firms in the sample. On average, data-related expenditure is more than three times higher among AI innovators than among non-AI firms.

Figure 11 illustrates this pattern. The left panel shows that the distribution of data-related expenditure for AI innovators is shifted to the right relative to non-AI firms. The right panel further shows that, within the group of AI innovators, exporters report higher data-related expenditure than firms serving only the domestic market. These descriptive patterns indicate a close association between AI patenting, export participation, and firms' investments in data-related infrastructure and services.

Fact 4: Data-related expenditures, R&D, and service exports increase prior to AI patenting. Figure 14 traces firm outcomes around the filing of a firm's first AI patent. Data-related expenditure, R&D investment, and service exports all increase in the years leading up to the first AI patent filing and continue to rise thereafter. In contrast, we observe no comparable pat-

[3] AI patenting is also geographically and sectorally concentrated. A disproportionate share of AI patents originates from firms located in Karnataka, Maharashtra, Delhi, and Tamil Nadu, and from ICT-intensive industries. Additional descriptive evidence is provided in Appendix C.

tern for goods exports. These event-study profiles document descriptively that AI patenting is preceded by changes in several firm characteristics and investments.

Fact 5: AI patents are closely aligned with firms' observed product portfolios. To examine the relationship between firms' innovation and production activities, we map both patents and products into a common Harmonized System (HS) classification and measure the degree of patent–product alignment. AI innovators exhibit the highest degree of alignment, followed by ICT innovators and other firms. Approximately 67% of AI patents match firms' observed product portfolios at least at the HS2 level, compared with 61% for ICT patents and 52% for other patents, with a similar ordering at finer levels of aggregation (Table A9). These patterns indicate that AI patents are typically associated with technologies that are closely related to firms' existing products rather than being concentrated in unrelated activities. Additional details on the construction of the alignment measure are provided in Appendix B.3.

These stylized facts reveal several systematic differences between AI innovators and other firms in the sample. AI patenting is rare and concentrated among a small group of firms, which are more likely to export, particularly in services, and report higher data-related expenditures. In addition, data-related expenditures, R&D investment, and service exports tend to increase before firms' first AI patent filing, and AI patents are closely aligned with firms' observed product portfolios.

While these patterns are purely descriptive, they motivate the empirical analysis that follows. In particular, they are consistent with the possibility that firms' exposure to international markets and their investments in data-related activities are closely linked to AI innovation. The next section investigates this relationship using plausibly exogenous variation in firm-level export opportunities generated by foreign demand shocks.

4 Empirical Strategy

To identify the effect of export opportunities on firms' AI innovation, we exploit plausibly exogenous variation in firm-level export demand shocks. We combine firm-level information from PROWESS with international trade data to construct product- and destination-specific measures of foreign demand, which are then aggregated to the firm level using firms' predetermined product portfolios. The resulting firm-level export demand shocks capture changes in export opportunities that originate in foreign markets rather than in Indian firms themselves and constitute the main source of identifying variation in the empirical analysis.

4.1 Foreign Demand Shocks

Our preferred identification strategy follows the shift-share approach developed in the international trade literature and exploits plausibly exogenous changes in foreign demand faced by Indian firms. The key intuition is that firms differ in their exposure to foreign markets because they produce different products and export to different destinations (Aghion et al., 2023, 2024). Consequently, changes in demand originating abroad affect firms differentially depending on

their predetermined product mix. The identification logic follows Borusyak et al. (2022): under quasi-random assignment of product-level shocks, a shift-share regressor captures the causal effect of trade exposure on firm outcomes even when the firm-level shares are not strictly exogenous.

Foreign demand at the product–destination level. We begin by constructing product-level foreign demand shocks separately for goods and services. For goods, we use bilateral merchandise trade flows from the CEPII BACI database at the six-digit Harmonized System (HS6) level. For services, we use bilateral trade flows from the OECD–WTO Balanced Trade in Services (BaTIS) database at the Extended Balance of Payments Services (EBOPS) classification. In both cases, foreign demand is measured using imports by destination country from all exporting countries excluding India, thereby removing mechanical correlations between Indian supply shocks and measured foreign demand. Let X_{dkt} denote imports of product (or service category) k by destination country d in year t, where k corresponds to either an HS6 product or an EBOPS category. Following Barrows and Ollivier (2021), annual foreign demand growth is measured using the Davis–Haltiwanger growth rate,

$$DH(X_{dkt}) = \frac{X_{dkt} - X_{dk,t-1}}{0.5\,(X_{dkt} + X_{dk,t-1})},$$

which approximates a log difference while remaining well defined in the presence of entry into and exit from export markets.

Product-level foreign demand shocks. For each product or service category k, the foreign demand shock is constructed as

$$\Delta FD_{kt} = \sum_{d \in \Omega_k} s_{dk}^{\text{pre}}\, DH(X_{dkt}),$$

where Ω_k denotes the set of destination countries importing product (or service category) k, and s_{dk}^{pre} are predetermined destination weights computed as the average share of destination d in global imports of product k during the pre-period. Using fixed destination weights prevents the measure from being affected by contemporaneous reallocations of exports across markets. The resulting shocks vary at the product-year level and capture changes in foreign demand originating in destination markets. Throughout the paper, product-level shocks are constructed separately for goods (using BACI and HS6 classifications) and services (using BaTIS and EBOPS classifications) before being aggregated to the firm level.

Firm-level export demand shocks. Many Indian firms export both goods and services, particularly in technology-intensive industries such as software, business services, and information technology. Restricting the analysis to either margin alone would therefore provide an incomplete measure of firms' exposure to international markets. We therefore construct a firm-level export demand shock that incorporates foreign demand changes for both goods and services according to firms' predetermined product portfolios.

Let k denote a product or service category, where k corresponds either to a six-digit Har-

monized System (HS6) code for goods or to an EBOPS category for services.[4] For each firm-product pair (i,k), we compute pre-period revenue shares using CMIE Prowess data over 2002–2004 and normalize these shares so that they sum to one across all classified goods and services produced by the firm:

$$w_{ik}^{\text{pre}} = \frac{\tilde{w}_{ik}^{\text{pre}}}{\sum_{k' \in K_i^{\text{class}}} \tilde{w}_{ik'}^{\text{pre}}}, \tag{1}$$

where K_i^{class} includes all classified goods and service categories produced by firm i. The firm-level export demand shock is then

$$Z_{it} = \sum_{k \in K_i^{\text{class}}} w_{ik}^{\text{pre}} \Delta FD_{kt}, \tag{2}$$

where ΔFD_{kt} denotes the foreign demand shock for product or service category k. Goods shocks are constructed from BACI using HS6 trade flows, while service shocks are constructed from BaTIS using EBOPS trade flows. In both cases, product-level shocks aggregate destination-specific demand growth using predetermined destination weights and exclude India's own exports from the calculation.

For interpretation, the firm-level shock can be decomposed into goods and services components:

$$Z_{it} = z_{it}^{goods} + z_{it}^{services}, \tag{3}$$

where

$$z_{it}^{goods} = \sum_{k \in K_i^{goods}} w_{ik}^{\text{pre}} \Delta FD_{kt}, \qquad z_{it}^{services} = \sum_{k \in K_i^{services}} w_{ik}^{\text{pre}} \Delta FD_{kt}. \tag{4}$$

Importantly, z_{it}^{goods} and $z_{it}^{services}$ are components of the same portfolio-weighted firm-level shock: the weights are normalized across all classified goods and services, not separately within each margin. Thus, the relative importance of goods and services is determined by the firm's pre-period portfolio composition. By construction, firms operating within the same four-digit industry may experience different trajectories of Z_{it} because they produce different combinations of goods and services and therefore face different changes in foreign demand. This within-industry variation in export demand exposure provides the identifying variation in our empirical analysis.

[4] Our baseline uses EBOPS Level 2, the most disaggregated classification available. As a robustness check, Appendix E reports results using EBOPS Level 1.

4.2 Empirical Specification

Our objective is to estimate the reduced-form effect of exogenous changes in foreign demand on firms' AI innovation rather than the causal effect of exports per se. One might instead use the foreign demand shock as an instrument for firms' exports in a two-stage least-squares framework. Such an approach, however, requires the exclusion restriction that foreign demand affects AI innovation solely through realized export activity. This assumption is difficult to justify in our setting because foreign demand shocks induce a broad set of firm responses beyond contemporaneous export volumes.[5] These responses constitute precisely the mechanisms through which foreign demand may foster AI innovation. Accordingly, we interpret Z_{it} as an exogenous demand shifter and estimate its reduced-form effect on AI innovation.

Our main estimating equation is

$$\mathrm{AI}_{it} = \alpha_i + \alpha_{kt} + \beta Z_{it} + \mathbf{X}'_{i0}\delta_t + \varepsilon_{it}, \tag{5}$$

where AI_{it} is an indicator equal to one if firm i files an AI-related patent in year t, Z_{it} is the firm-level export demand shock defined above, α_i are firm fixed effects, α_{kt} are four-digit industry-by-year fixed effects, and $\mathbf{X}_{i0}$ is a vector of baseline firm characteristics—including firm size, age, R&D intensity, export status, foreign ownership, leverage, and prior patenting history—interacted with year indicators. Standard errors are clustered at the two-digit industry-by-state level to allow for arbitrary correlation in unobservables across firms operating in the same broad sector and geographic area.

Firm fixed effects absorb all time-invariant firm characteristics, while industry-by-year fixed effects absorb shocks common to firms operating within the same narrowly defined industry. Consequently, identification comes from within-firm changes in export demand exposure over time, after netting out industry-wide shocks. Because firms have different predetermined product portfolios, changes in foreign demand translate into different trajectories of Z_{it} even for firms operating within the same four-digit industry.[6]

Identification assumption. The exogeneity of shift-share can stem from the exogeneity in either the shares or the shocks (Borusyak et al., 2022). In our case, we argue that identification of β comes from the exogeneity of the shocks. Hence, it requires that, conditional on the included fixed effects and controls, fluctuations in foreign demand ΔFD_{pt} are quasi-randomly assigned across HS6 products. Formally, following Borusyak et al. (2022), the identifying restriction is

$$\mathbb{E}\left[\Delta FD_{pt} \cdot \tilde{\varepsilon}_{ipt} \,\middle|\, \boldsymbol{X}_{i0}, \alpha_i, \alpha_{kt}\right] = 0,$$

where $\tilde{\varepsilon}_{ipt}$ is the firm-product-year structural residual implicit in ε_{it}. The assumption is plausible in our setting because Indian firms in our sample account for a small share of global trade

[5] These responses include changes in revenues, profits, investment, and expectations about future market opportunities, all of which may affect firms' incentives to innovate (Akcigit and Melitz, 2022).

[6] For example, two firms classified in the same four-digit industry may specialize in different products or export different services, implying different exposure to changes in foreign demand over time.

at the HS6 level, so foreign demand growth for any individual HS6 product reflects global factors (demand shocks in importing countries, third-country supply responses, exchange rate movements) largely exogenous to individual Indian firms' decisions. Given that foreign demand is measured using imports from all countries *excluding* India, and because firm exposure is determined by predetermined product shares, changes in Z_{it} plausibly capture exogenous variation in export opportunities rather than endogenous adjustments in firms' production or innovation decisions. Put differently, firms producing different HS6 products experience differential changes in foreign demand for reasons unrelated to their own AI activities, allowing us to identify the causal effect of trade exposure on innovation.

Given the quasi-random assignment of foreign demand shocks, the primary threat to identification is that unobserved firm characteristics may both predict future growth and be correlated with exposure to growing foreign markets (Barrows and Ollivier, 2021). For example, firms with more sophisticated market research capabilities or superior managerial quality may systematically enter markets experiencing rapid demand expansion while also exhibiting stronger underlying performance. In that case, estimated effects would reflect endogenous market selection rather than exogenous changes in foreign demand. To assess this concern, we follow the diagnostic approach by Aghion et al. (2024) and Barrows and Ollivier (2021) and run falsification tests of correlations between current firm outcomes and future foreign demand shocks. Specifically, we regress contemporaneous values of our dependent variables on one- and two-year leads of the shift–share instrument. Under the identifying assumptions, future shocks should not explain current outcomes. A significant relationship would indicate that firms subsequently exposed to more favorable foreign demand shocks already differed in their outcomes before those shocks materialized, raising concerns about the instrument's exogeneity (see Appendix F.2).

In addition to shock exogeneity, the validity of the shift–share design requires a sufficiently large and heterogeneous set of underlying shocks (Borusyak et al., 2022). Our setting satisfies this condition. We construct the instrument using 91,151 HS6-by-year foreign demand growth rates, ΔFD_{pt}, covering 5,479 distinct six-digit product codes over 18 years. This large set of shocks provides substantial identifying variation across firms. Moreover, the distribution of foreign demand growth rates exhibits considerable dispersion. As documented in Appendix D, both the raw and residualized shocks display meaningful variation across products, even after removing product and industry-by-year fixed effects, indicating that the identifying variation is not driven by a small subset of products or common sectoral trends.

5 Results

This section presents the main empirical findings. We first estimate the effect of exogenous changes in export opportunities on firms' AI innovation. We then examine the mechanisms underlying this relationship, focusing on firms' accumulation of data-related capabilities and the differential role of goods and services.

5.1 Export Demand, Firm Outcomes, and AI Innovation

We begin by estimating the effect of firm-level export demand shocks on firms' innovation outcomes. Table 1 reports the baseline estimates using our preferred measure of export demand exposure, which incorporates foreign demand changes across the goods and services produced by each firm.

Table 1: Foreign Demand and Innovation

	Patent indicator (dummy)			Patent stock		
	(1) AI	(2) ICT	(3) Any	(4) AI	(5) ICT	(6) Any
Z_{it}	0.244*	0.0514	0.0605	0.0538***	0.101***	0.0941*
	(2.36)	(0.72)	(0.98)	(4.02)	(7.53)	(2.02)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
NIC4D×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls×Year	Yes	Yes	Yes	Yes	Yes	Yes
N	1306	2449	14093	1597	6451	20435

Notes: The table reports estimates from the baseline specification using the firm-level export demand shock constructed from firms' predetermined goods and services portfolios. Columns (1)–(3) estimate patent indicators using Poisson models, while columns (4)–(6) estimate the logarithm of cumulative patent stocks using OLS. All specifications include firm fixed effects, four-digit industry-by-year fixed effects, and baseline firm characteristics (including firm size, sales, R&D expenditure, data-related expenditure, export status, foreign ownership, leverage, and prior patenting history) interacted with year indicators. Standard errors are clustered at the two-digit industry-by-location level. Robust t-statistics are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The results show that stronger export demand increases AI innovation. Since the export demand shock is standardized, the coefficient in column (1) implies that a one-standard-deviation increase in export demand raises the probability of filing an AI patent by 0.244 percentage points. Column (4) shows that the same increase raises firms' AI patent stock by approximately 5%, indicating that export demand not only increases the likelihood of entering AI innovation but also stimulates the accumulation of AI-related knowledge over time. These effects are economically meaningful despite the rarity of AI innovation in our sample. AI patenting is observed for only a small fraction of firms, making statistically significant responses difficult to detect. Nevertheless, we find economically meaningful and statistically significant effects on both the extensive and intensive margins, suggesting that improved export opportunities play an important role in stimulating frontier AI innovation.

The effects are weaker for broader innovation outcomes. We find no statistically significant effect on the probability of filing ICT or other patents, despite these forms of innovation being considerably more common than AI patenting. Patent stocks, however, increase for ICT innovation and, to a lesser extent, for overall patenting activity. The positive effects on ICT and overall patent stocks are also consistent with the broader literature documenting the beneficial effects of export demand on innovation (Aghion et al., 2024; Akcigit and Melitz, 2022). Our results extend this evidence by showing that, while stronger export demand stimulates firms' innovative activity more broadly, its relationship with frontier AI innovation is particularly pronounced.

To examine how firms respond to these export opportunities, Table 2 reports the effects of

export demand shocks on export activity, firm performance, and data-related expenditure. As expected, the results indicate that positive export demand shocks translate into economically meaningful changes in firms' international activities. A one-standard-deviation increase in export demand raises the probability of exporting by approximately 12.5% and increases export sales by 2.5%. These changes are accompanied by a 2.9% increase in total sales, suggesting that improved foreign demand leads firms to expand their overall economic activity.

Table 2: Export Demand Shocks and Firm Outcomes

	(1) Exporter	(2) Log exports	(3) Log sales	(4) Log data exp.
Z_{it}	0.1178***	0.0249***	0.0283*	0.0503***
	(5.03)	(2.20)	(2.50)	(8.96)
Firm FE	Yes	Yes	Yes	Yes
NIC4D×Year FE	Yes	Yes	Yes	Yes
Baseline controls×Year	Yes	Yes	Yes	Yes
Estimator	Poisson	Poisson	OLS	OLS
N	252804	98218	390973	440002

Notes: The table reports estimates from the baseline specification using the firm-level export demand shock constructed from firms' predetermined goods and services portfolios. Column (1) is estimated using a Poisson model, while columns (2)–(4) are estimated by OLS with the dependent variables expressed in logarithms. All specifications include firm fixed effects, four-digit industry-by-year fixed effects, and baseline firm characteristics (including firm size, sales, R&D expenditure, data-related expenditure, export status, foreign ownership, leverage, and prior patenting history) interacted with year indicators. Standard errors are clustered at the two-digit industry-by-location level. Robust *t*-statistics are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The largest response is observed for data-related expenditure. A one-standard-deviation increase in export demand raises firms' expenditure on software, data centres, communications, web hosting, and other IT-enabled services by approximately 5.2%. This finding suggests that firms do not respond to stronger export opportunities solely by increasing production, but also by investing in the digital infrastructure required to generate, store, and process commercially valuable data.

These findings point to a consistent pattern of adjustment. Firms experiencing stronger export demand become more integrated into international markets, expand their sales, and invest more heavily in data-related activities. While these estimates do not, by themselves, establish that customer-generated data are the mechanism linking exports to AI innovation, they are consistent with the hypothesis that export opportunities encourage firms to invest in the digital capabilities required to acquire and manage commercially valuable data. As discussed in Section 4, our preferred specification combines firms' exposure to foreign demand for both goods and services into a single firm-level export demand shock. This measure captures overall export opportunities while allowing the relative contribution of each component to be determined by firms' predetermined product portfolios. For completeness, Appendix F.1 reports specifications in which the goods (HS6) and services (EBOPS) components are entered separately. The results show that the positive relationship between export demand and AI innovation is primarily driven by the services component, while the goods component has lit-

tle explanatory power once both margins are considered jointly. These findings are consistent with our emphasis on the role of data-intensive interactions with foreign clients in fostering AI innovation.

The next section provides a more direct test of this mechanism by exploiting the implementation of the GDPR as an exogenous increase in the cost of processing personal data.

6 Evidence on the Role of Data

The previous section showed that positive export demand shocks increase AI innovation and are accompanied by greater investment in data-related activities. While these findings are consistent with commercially generated data playing an important role in AI innovation, they do not establish whether data constitute the mechanism through which export opportunities translate into AI innovation. To examine this question, we exploit the implementation of the European Union's General Data Protection Regulation (GDPR), which introduced substantially stricter requirements governing the collection, processing, storage, and transfer of personal data.

The GDPR provides a useful setting because it directly changed the regulatory environment governing the use of commercially generated data. Through its extraterritorial application, firms processing the personal data of EU residents became subject to stricter compliance requirements and substantially higher penalties for non-compliance (Voigt and von dem Bussche, 2017). Recent evidence confirms that these changes affected firms' use of data-related digital inputs. Using detailed cloud-computing records from a global cloud provider, Demirer et al. (2024) show that the GDPR reduced firms' cloud data storage by 26% and cloud data processing by 15%, making firms significantly less data-intensive.

If commercially generated data are an important input into AI innovation, firms more exposed to the GDPR should experience a larger decline in AI innovative activity following the introduction of the regulation. At the same time, the GDPR may have affected firms through channels other than data use, for example by increasing the costs of serving customers in the European Union. We therefore examine not only AI innovation, but also firms' export activity, broader measures of firm performance, and other innovation outcomes. This allows us to assess whether the observed responses are specific to AI innovation or reflect more general changes in firms' international activities.

6.1 Identification Strategy

To estimate the effects of the GDPR, we need to identify which firms were more likely to be affected by the regulation. While firms exporting to the European Union were directly exposed to the new data protection requirements, CMIE Prowess reports firms' total exports but not their export destinations. We therefore construct a firm-level measure of pre-existing exposure to the European Union by combining firms' pre-GDPR product portfolios with product-level destination shares. Specifically, for each HS6 product (goods) and EBOPS Level 2 category (ser-

vices), we compute the share of developed-market exports destined for the European Union using BACI and BaTIS trade data averaged over the pre-GDPR period (2013–2015). We then aggregate these product-level shares to the firm level using firms' predetermined product portfolios. Firms whose products were more heavily oriented toward European Union markets before 2018 are therefore assigned greater exposure to the GDPR. Appendix G.1 provides the complete details of the construction of the exposure measure, including the definition of the developed-market comparison group and the treatment assignment.

We estimate the effects of the regulation using a difference-in-differences design that compares firms with high and low pre-GDPR exposure to the European Union before and after its implementation. Our baseline specification defines treatment using the median of the pre-GDPR exposure distribution: firms with above-median exposure are classified as treated, while those below the median serve as the control group. Intuitively, the design compares firms that are similarly integrated into developed export markets but differ in the extent to which their products are oriented toward the European Union, where the GDPR was introduced. To improve comparability between treated and control firms, we combine the difference-in-differences design with propensity score matching based on pre-treatment firm characteristics, including firm size, sales, R&D expenditure, data-related expenditure, export status, firm age, and ownership characteristics.[7]

Our baseline specification is

$$Y_{it} = \beta \left(D_i \times Post_t \right) + \alpha_i + \alpha_t + \varepsilon_{it}, \tag{6}$$

where Y_{it} denotes the outcome of interest for firm i in year t, D_i identifies firms with high pre-GDPR exposure to the European Union, and $Post_t$ equals one for years from 2018 onward. All specifications include firm and year fixed effects, and standard errors are clustered at the two-digit industry-by-location level.

The coefficient of interest, β, captures the differential change in outcomes for firms more exposed to the GDPR following its implementation. Because exposure is constructed entirely from pre-GDPR information, firms cannot adjust their treatment status in response to the regulation. Moreover, by measuring exposure relative to other developed export destinations, the comparison is between firms serving broadly similar foreign markets while differing in their orientation toward the European Union. Identification therefore relies on the standard parallel trends assumption that, absent the GDPR, firms with high and low pre-existing exposure to the European Union would have followed similar trends. Appendix G.2 reports additional details on the matching procedure.

To assess the parallel trends assumption and examine the timing of the response, we estimate the following event-study specification:

[7] Appendix Table A26 reports balance diagnostics for the propensity score matching procedure. Before matching, firms with high EU exposure differ from the control group along several observable dimensions, particularly data-related expenditure. After kernel matching, these differences become economically small across all matching covariates. Figure 21 shows substantial overlap in the propensity score distributions, supporting the validity of the matching design.

$$Y_{it} = \sum_{\tau \neq -1} \delta_\tau \left(D_i \times \mathbf{1}\{t - 2018 = \tau\} \right) + \alpha_i + \alpha_t + \varepsilon_{it}, \tag{7}$$

where $\mathbf{1}\{t - 2018 = \tau\}$ is an indicator for event time τ relative to the implementation of the GDPR in 2018. The year immediately preceding the regulation, 2017, is omitted and serves as the reference period. The coefficients δ_τ trace the differential evolution of outcomes for firms with high pre-GDPR exposure to the European Union relative to less exposed firms. Estimates for $\tau < 0$ provide a direct test of the parallel trends assumption, while estimates for $\tau \geq 0$ characterize the dynamic response following the implementation of the GDPR.

6.2 Results

Table A27 reports the effects of the GDPR on firms' innovation outcomes using the matched difference-in-differences design. We find no statistically significant effect on the probability of filing an AI patent, suggesting that exposure to the GDPR did not affect firms' entry into AI innovation. In contrast, column (2) shows a negative and statistically significant effect on the logarithm of AI patent stocks, indicating that firms more exposed to the regulation experienced a decline in the intensity of AI innovation following the implementation of the GDPR. The estimated coefficient implies that high-exposure firms accumulated AI patent stocks approximately 0.9% lower than otherwise comparable firms with lower exposure to the European Union.

Table 3: GDPR Effects on Innovation

	Patent Entry			Patent Intensity (Log Stock)		
	(1) AI	(2) ICT	(3) Any	(4) AI	(5) ICT	(6) Any
High EU Exposure × Post	0.002	0.024**	-0.009	-0.009*	-0.100	-2.518
	(0.003)	(0.009)	(0.018)	(0.005)	(0.092)	(2.194)
Estimator	Poisson	Poisson	Poisson	OLS	OLS	OLS
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Kernel PSM Weights	Yes	Yes	Yes	Yes	Yes	Yes
N	4,133	4,133	4,133	4,133	4,133	4,133

Notes: Difference-in-differences estimates using kernel propensity-score matching weights. High EU Exposure equals one for firms with above-median pre-GDPR exposure to the European Union, measured as the EU share within developed-country export exposure. Columns (1)–(3) report Poisson estimates of patent entry, while columns (4)–(6) report OLS estimates for the logarithm of patent stocks. All specifications include firm and year fixed effects. Standard errors are clustered at the two-digit industry × location level. Matching is performed using an Epanechnikov kernel with common support based on firms' 2016 characteristics, including assets, sales, employment, age, R&D intensity, exporter status, goods exports, services exports, and data-related expenditure.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The remaining outcomes provide little evidence of a broader slowdown in innovative activity. We find no significant effects on overall patenting, either at the extensive or intensive margin (columns 5 and 6). Interestingly, the probability of filing an ICT patent increases following the GDPR (column 3), although we detect no corresponding increase in ICT patent stocks (column 4). This contrast is notable because AI patenting is substantially rarer than ICT

patenting in our sample. Despite the low incidence of AI innovation, exposure to the GDPR leads to a measurable decline in the stock of AI patents, whereas innovative activity in the broader ICT sector remains largely unaffected. One possible interpretation is that firms partially reallocated innovative effort toward complementary digital technologies that rely less heavily on commercially generated customer data than frontier AI applications.

A natural concern is that the GDPR may have affected AI innovation indirectly by increasing the cost of serving customers in the European Union, effectively acting as an EU-specific trade cost rather than a data-related regulatory shock. Under this interpretation, the estimated decline in AI patenting would reflect reduced access to European markets rather than restrictions on firms' ability to collect and process commercially generated data.

Table 4: GDPR Effects on Trade and Firm Performance

	(1) Sales	(2) Goods Exports	(3) Services Exports	(4) Total exports	(5) TFP	(6) Exporter
High EU Exposure × Post	0.048	-0.141	0.915	-0.039	0.021	0.080
	(0.233)	(0.444)	(0.608)	(0.248)	(0.159)	(0.076)
Estimator	OLS	OLS	OLS	OLS	OLS	Poisson
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Kernel PSM Weights	Yes	Yes	Yes	Yes	Yes	Yes
N	3,373	2,893	1,817	3,221	3,059	4,133

Notes: Difference-in-differences estimates for firm performance and trade outcomes using kernel propensity-score matching weights. High EU Exposure equals one for firms with above-median pre-GDPR exposure to the European Union, measured as the EU share within developed-country export exposure. Columns (1)–(5) are estimated by OLS and column (6) by Poisson. All specifications include firm and year fixed effects. Standard errors are clustered at the two-digit industry × location level. Matching is performed using an Epanechnikov kernel with common support based on firms' 2016 characteristics, including log assets, log sales, log employment, firm age, R&D intensity, exporter status, log goods exports, log services exports, and log data expenditure.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A28 provides an initial assessment of this alternative explanation. We find no statistically significant effects on sales, goods exports, services exports, total exports, total factor productivity, or the probability of exporting, providing little evidence that firms more exposed to the GDPR experienced a deterioration in their overall business activity or export performance. While stable total exports do not by themselves rule out an EU-specific trade shock—firms may redirect exports toward alternative destinations in response to higher trade costs (Meng and Wang, 2023)—additional analyses reported in Appendix G.4 provide little support for this interpretation. Using product-level destination regressions, aggregate India–EU trade flows and within-EU heterogeneity, we find no evidence that the GDPR reduced Indian firms' exports to the European Union or induced a reallocation of trade away from EU markets.

Overall, the results indicate that the GDPR is associated with a decline in the intensity of AI innovation, while leaving firms' trade performance and broader economic activity largely unchanged. We also find no evidence that the regulation reduced Indian firms' access to European markets or induced a reallocation of trade away from the European Union. Although these results do not by themselves establish the underlying mechanism, they are consistent with the interpretation that the GDPR primarily affected firms through changes in the regulatory environment governing the collection and processing of commercially generated data, rather than through reduced export opportunities.

Event-Study Evidence The identifying assumption underlying our difference-in-differences design is that, absent the GDPR, firms with high and low pre-existing exposure to European markets would have followed parallel trends in innovation outcomes. We assess this assumption by estimating the dynamic event-study specification described in Section 6.1.

Figure 1 reports the event-study estimates for AI patent stocks. The pre-treatment coefficients are small and statistically indistinguishable from zero, providing support for the parallel trends assumption underlying the difference-in-differences design. Following the implementation of the GDPR, the estimated effects become negative and remain so during the first three post-treatment years, consistent with a decline in AI patenting intensity among firms more exposed to European data regulation. The estimated effects become statistically insignificant in the later post-treatment years. One possible interpretation is that firms gradually adapted to the new regulatory environment by adjusting their data management and compliance practices. Overall, the dynamic pattern is consistent with the negative average treatment effect reported in Table A27.

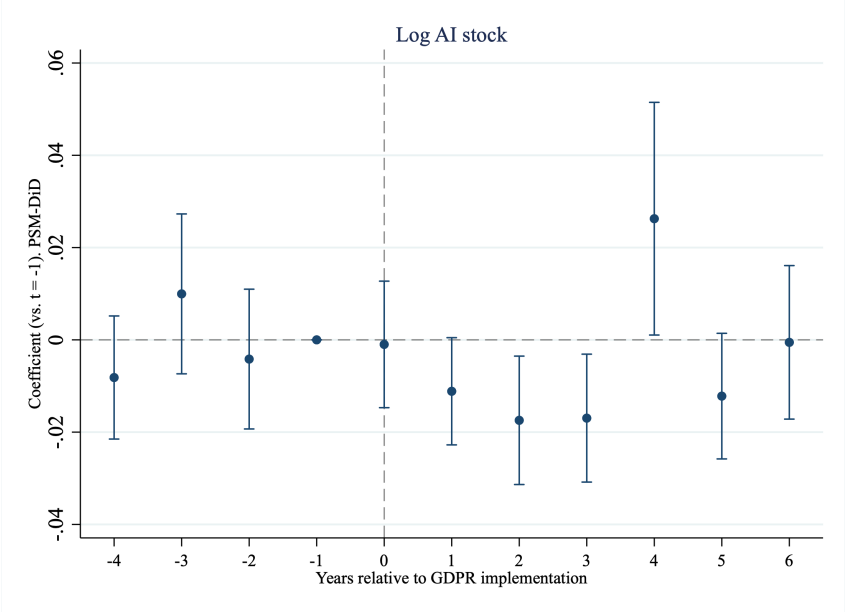


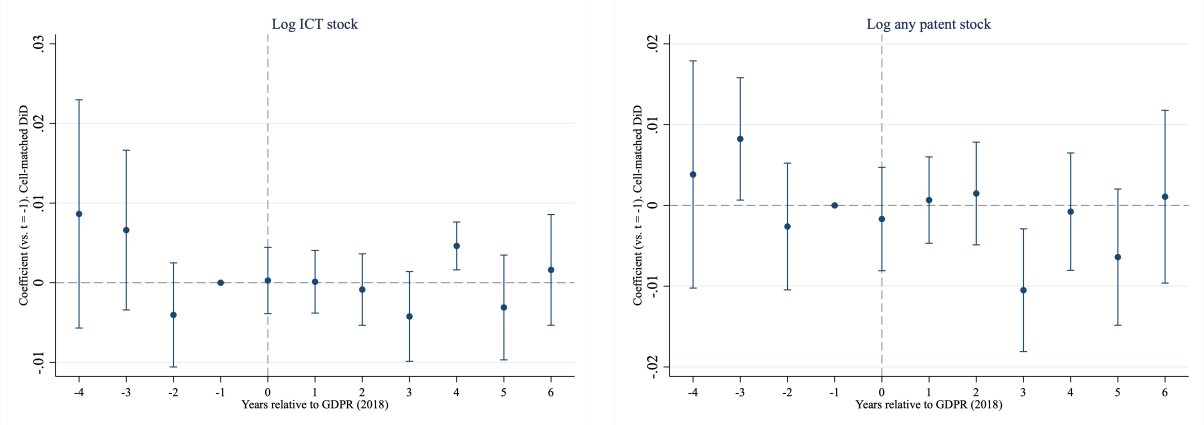
Figure 1: Event-Study Estimates: GDPR and AI Patent Stock

Notes: The figure reports dynamic difference-in-differences estimates of the effect of the GDPR on the logarithm of AI patent stock. Coefficients are estimated relative to the year immediately preceding the implementation of the GDPR (2017), using the kernel propensity-score matched sample. Vertical bars denote 95% confidence intervals. All specifications include firm and NIC4D×year fixed effects, and standard errors are clustered at the two-digit industry × location level.

The event-study estimates for other innovation outcomes and broader measures of firm performance reinforce this interpretation. Across all specifications, the pre-treatment coefficients are economically small and statistically indistinguishable from zero, providing no evidence of differential pre-trends between treated and control firms. Following the implementation of the GDPR, the estimated effects remain close to zero and are generally statistically insignificant throughout the post-treatment period.

Figure 2 shows little evidence that the regulation affected ICT patenting or broader innovation activity beyond AI. Similarly, Figure 23 indicates no meaningful effects on firms' export

participation, export sales, total sales, or productivity. The absence of systematic responses in these outcomes is important for interpreting the main results. If the GDPR had primarily reduced firms' access to foreign markets or weakened their overall economic performance, the decline in AI patenting could simply reflect lower export activity or a general contraction in innovative effort. Instead, the evidence points to a much more specific effect. The regulation reduced AI patenting intensity while leaving firms' participation in international markets and broader economic performance largely unchanged, consistent with an increase in the costs of collecting and processing commercially generated data rather than a deterioration in firms' export opportunities. This interpretation is also consistent with the growing evidence that AI innovation relies more heavily on access to data than other forms of technological change (Centre et al., 2020; Bessen et al., 2020; Beraja et al., 2022).



(a) ICT patent stock (b) Total patent stock

Figure 2: Event-Study Estimates: ICT and Overall Innovation

Notes: The figure reports event-study estimates for the logarithm of ICT patent stock (Panel A) and total patent stock (Panel B). Coefficients are estimated relative to the year immediately preceding the implementation of the GDPR (2017), which is omitted from the regressions. The sample is weighted using Epanechnikov kernel propensity-score matching weights with common support. All specifications include firm and NIC4D×year fixed effects. Standard errors are clustered at the two-digit industry × location level. Vertical bars denote 95% confidence intervals.

The Role of Data. The previous section showed that positive export demand shocks are associated with higher data-related expenditure. A natural question is how firms adjust these expenditures when the use of commercially generated data becomes more costly. To explore this margin, we examine the share of data-related expenditure in total firm costs. Although this measure does not capture the quantity of data firms collect or process, it provides information on the importance of data-related activities within firms' overall spending.

Figure 3 reports the corresponding event-study estimates. The pre-treatment coefficients are small and statistically indistinguishable from zero, providing no evidence of differential pre-trends between treated and control firms. Following the implementation of the GDPR, the estimated coefficients become positive and remain so throughout the post-treatment period, suggesting that firms more exposed to the regulation allocated a larger share of their total expenditures to data-related activities.

This pattern should not be interpreted as evidence that firms increased their use of data.

Rather, it indicates that data-related activities became relatively more important within firms' overall cost structure after the introduction of the GDPR. This finding complements the evidence in Demirer et al. (2024), who show that the GDPR reduced firms' cloud data storage and processing. The results suggest that privacy regulation affected not only firms' use of data but also the way they allocated resources to data-related activities.

The precise adjustment margin remains an open question. Firms may have responded by increasing investments in compliance, data governance, or digital infrastructure, or by relying more heavily on externally provided data services. Distinguishing between these responses is beyond the scope of our data and represents a promising direction for future research.

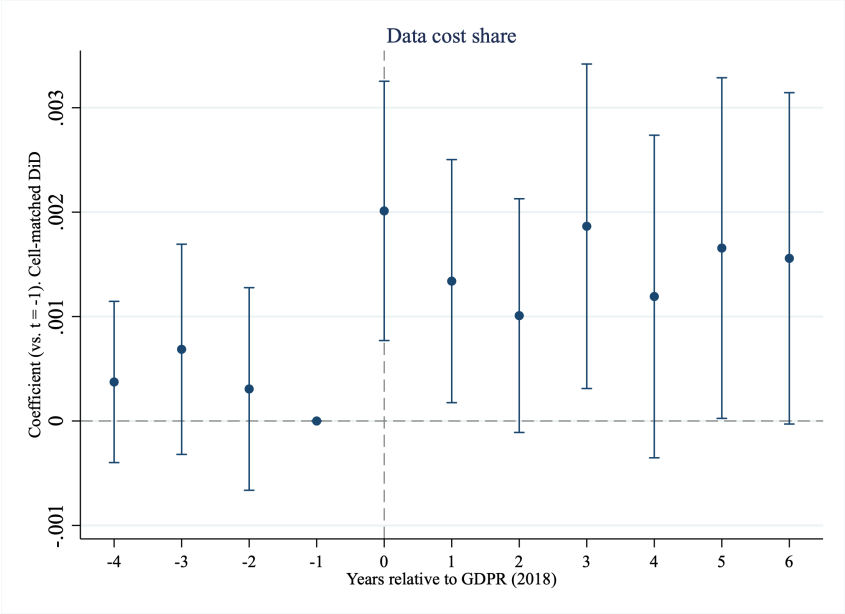


Figure 3: Event-study estimates of the GDPR effect on the share of data-related expenditure in total firm costs

Notes: The figure reports event-study estimates from the matched difference-in-differences specification using the share of data-related expenditure in total firm costs as the outcome. Coefficients are estimated relative to the year immediately preceding the implementation of the GDPR (2017), which is omitted from the regression. The sample is weighted using Epanechnikov kernel propensity-score matching weights with common support. All specifications include firm and year fixed effects. Standard errors are clustered at the two-digit industry $\times$ location level. Vertical bars denote 95% confidence intervals.

6.3 Robustness and Validation

This section assesses the robustness of our findings and the credibility of the identification strategy. We begin with placebo and falsification exercises that evaluate the identifying assumptions underlying the shift–share design. We then show that the results are robust to alternative constructions of the foreign demand shock and to flexible controls for firms' initial export composition. For the GDPR analysis, we additionally use propensity score matching to improve comparability between firms with high and low exposure to European markets. Finally, we examine whether the estimates are sensitive to sample composition or driven by a small number of influential observations.

Placebo and falsification tests. The identifying assumption underlying the shift–share design is that future foreign demand shocks should not predict current firm outcomes. Following Borusyak et al. (2022) and Barrows and Ollivier (2021), we estimate placebo regressions replacing the contemporaneous Bartik shock with future values of the instrument. We find no evidence that future foreign demand shocks predict current AI innovation, supporting the exogeneity of the identifying variation (see Appendix F.2).

Alternative construction of the foreign demand shock. Our baseline specification combines goods- and services-based foreign demand shocks into a single portfolio-weighted measure. As a robustness check, we estimate specifications entering the goods (HS6) and services (EBOPS) Bartik shocks separately. While the estimated effects are concentrated in services, the combined measure remains our preferred specification because it captures firms' overall exposure to foreign demand. We further show that the results are robust to alternative levels of product aggregation (HS4 and HS2) and to alternative weighting schemes based on firms' pre-period export portfolios (see Appendix F.1).

Initial export composition. A potential concern is that the Bartik measure captures persistent differences in firms' initial export specialization rather than exogenous variation in foreign demand. To address this issue, we interact firms' baseline export shares with year fixed effects, allowing firms with different initial export portfolios to follow flexible differential trends. The estimated effects remain quantitatively similar, indicating that firms' initial export composition does not drive the results.

Propensity score matching. Firms with greater pre-GDPR exposure to European markets differ from less exposed firms along several observable dimensions, raising the concern that pre-existing differences between the two groups may drive the estimated GDPR effects. As an additional robustness check, we combine propensity score matching with our difference-in-differences design. We match firms based on pre-GDPR characteristics, including size, employment, age, R&D expenditure, export activity, and data-related expenditure, and re-estimate the GDPR specification using kernel matching weights. Matching substantially improves covariate balance between high- and low-exposure firms. The event-study estimates using the matched sample show no evidence of differential pre-trends in AI innovation and reproduce the decline following GDPR implementation, consistent with the baseline results (see Appendix G.2).

7 Discussion, Policy Implications, and Limitations

This paper argues that international trade may contribute to AI innovation through a channel that has received relatively little attention in the trade literature: the generation of commercially valuable data. While previous work has emphasized mechanisms such as market expansion, competition, and learning-by-exporting, our findings suggest that interactions with foreign customers may also generate data that become an important input into AI development. The combination of the foreign demand and GDPR evidence is consistent with this

interpretation. Positive export demand shocks increase AI innovation and data-related expenditure, while a regulatory increase in the expected costs of using customer data reduces AI patenting intensity without materially affecting firms' trade performance.

These findings also contribute to the growing literature on data as a factor of production. Existing research has documented the importance of data for AI systems and the effects of privacy regulation on firms' data usage (Beraja et al., 2022; Demirer et al., 2024; Igna and Venturini, 2023). Our results connect these insights to international trade, suggesting that trade may influence firms' innovative capacity not only through revenues or knowledge diffusion, but also through access to commercially generated data. In this sense, data represent an additional mechanism linking globalization and frontier innovation.

Policy implications. Our findings suggest that policies governing cross-border data flows may influence firms' innovative activity in ways that are distinct from conventional trade policies. Unlike tariffs or other barriers to market access, privacy regulations primarily affect the conditions under which firms can collect, process, and use commercially generated data. Our results therefore highlight that trade policy and data governance should not be viewed as independent policy domains. At the same time, these findings should not be interpreted as a welfare evaluation of privacy regulation. Privacy laws generate important benefits by protecting consumers' personal information and strengthening trust in digital markets. Rather, our evidence suggests that these benefits may coexist with changes in firms' incentives and opportunities to develop AI technologies, illustrating the need to better understand the trade-offs between privacy protection and innovation.

Limitations. Several limitations remain. First, our analysis focuses on AI innovation, as measured by patenting activity, rather than the adoption and diffusion of AI technologies within firms. AI patenting is a relatively rare activity, capturing firms at the technological frontier rather than the broader population of AI users. While innovation determines the development of new AI capabilities, understanding how data availability shapes firms' incentives to adopt and deploy existing AI systems remains an equally important question for future research.

Second, although the GDPR provides a useful source of variation in the costs of using commercially generated data, our data do not allow us to directly observe the quantity, quality, or composition of the data firms collect. Instead, we infer the role of data from firms' innovation responses and changes in data-related expenditure. As a result, we cannot distinguish whether firms adjusted by investing more in compliance and digital infrastructure, relying more heavily on specialized providers of data services, purchasing third-party datasets, or shifting toward alternative sources of training data.

Third, our empirical strategy identifies the role of customer-generated data arising from firms' relationships with foreign customers. Whether commercially generated data generated through supplier relationships and production networks similarly contribute to AI innovation remains an important open question. Studying this channel would require detailed firm-level information on imported intermediate inputs and their shares in production to construct plausibly exogenous import supply shocks. Because our data observe firms' outputs but not the

composition of their imported inputs, we cannot credibly measure firms' exposure to foreign supply shocks or isolate the contribution of supplier-generated data.

Finally, our analysis focuses on innovation generated within firms. Yet AI innovation often emerges from interactions between firms, universities, and other research institutions. Our data do not observe these collaborations, preventing us from examining whether customer-generated data acquired through commercial relationships also stimulate innovation outside the firm or whether knowledge created in academic institutions feeds back into firms' AI innovation. Understanding these knowledge and data spillovers remains an important direction for future research.

More broadly, our findings highlight the growing importance of data as an economic input in the AI era. As countries increasingly adopt privacy regulations and data-governance frameworks, understanding how firms generate, access, and use commercially valuable data will become central to explaining patterns of technological progress and international competitiveness. Future research combining detailed information on firms' data acquisition, storage, processing, AI adoption, and AI development, together with richer information on firms' customer and supplier relationships, would help quantify these mechanisms more directly and provide a richer understanding of the relationship between trade, data, and technological change.

8 Conclusion

Artificial intelligence is becoming an increasingly important source of technological progress, yet relatively little is known about the economic forces that enable firms to develop AI technologies. This paper studies whether export relationships contribute to AI innovation by enabling firms to accumulate commercially valuable customer data generated through interactions with foreign markets.

To answer this question, we construct a novel firm-level dataset combining financial information from CMIE Prowess, patent records from the Indian Patent Office, and newly developed crosswalks linking firms' products and services to detailed trade classifications. Exploiting plausibly exogenous variation in foreign demand through a shift-share design, we find that stronger export opportunities increase firms' innovative activity, with particularly strong effects on AI innovation. These effects are accompanied by greater investment in data-related activities, suggesting that firms expand their digital capabilities as export opportunities improve.

We then examine the proposed mechanism using the implementation of the European Union's General Data Protection Regulation (GDPR). Firms with greater pre-existing exposure to the European Union experienced a decline in AI patenting intensity despite showing little evidence of changes in exports, sales, productivity, or broader innovation activities. At the same time, these firms allocated a larger share of their expenditures to data-related activities, indicating that such activities became relatively more costly following the GDPR. These

findings are consistent with customer-generated data being an important mechanism linking exports to AI innovation.

More broadly, our results suggest that export relationships create not only market opportunities but also informational assets that support frontier innovation. As AI becomes increasingly dependent on access to high-quality data, understanding how firms accumulate these data—and how policies shape their collection, use, and governance—will become increasingly important for explaining technological change, firm competitiveness, and economic growth.

References

Aghion, Philippe, Antonin Bergeaud, Timothee Gigout, Matthieu Lequien, and Marc Malitz (2023), "Exporting ideas: Knowledge flows from expanding trade in goods."

Aghion, Philippe, Antonin Bergeaud, Matthieu Lequien, and Marc J Melitz (2024), "The heterogeneous impact of market size on innovation: Evidence from french firm-level exports." *Review of Economics and Statistics*, 106, 608–626.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2026), "The economics of artificial intelligence." URL https://press.uchicago.edu/ucp/books/book/chicago/E/bo35780726.html.

Ahn, JaeBin, Hyoungmin Han, and Yi Huang (2018), "Trade with benefits: New insights on competition and innovation." Technical report, Graduate Institute of International and Development Studies Working Paper.

Akcigit, Ufuk and Marc J. Melitz (2022), "International trade and innovation." In *Handbook of International Economics* (Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, eds.), volume 5, chapter 6, 377–404, Elsevier.

Aridor, Guy, Yeon-Koo Che, and Tobias Salz (2020), "The economic consequences of data privacy regulation: Empirical evidence from gdpr." Technical Report 26900, NBER.

Autor, David, David Dorn, Gordon H Hanson, Gary Pisano, and Pian Shu (2020), "Foreign competition and domestic innovation: Evidence from us patents." *American Economic Review: Insights*, 2, 357–374.

Barrows, Geoffrey and Hélène Ollivier (2021), "Foreign demand, developing country exports, and co2 emissions: Firm-level evidence from india." *Journal of Development Economics*, 149, 102587, URL https://www.sciencedirect.com/science/article/pii/S0304387820301620.

Baruffaldi, Stefano, Brigitte van Beuzekom, Hélène Dernis, Dietmar Harhoff, Nandan Rao, David Rosenfeld, and Mariagrazia Squicciarini (2020), "Identifying and measuring developments in artificial intelligence: Making the impossible possible."

Beraja, Martin, David Y Yang, and Noam Yuchtman (2022), "Data-intensive Innovation and the State: Evidence from AI Firms in China." *The Review of Economic Studies*, 90, 1701–1723, URL https://doi.org/10.1093/restud/rdac056.

Bergeaud, Antonin, Philippe Aghion, Timothée Gigout-Magiorani, Matthieu Lequien, and Marc J. Melitz (2024), "Exporting ideas: Knowledge flows from expanding trade in goods." Discussion Paper 19260, CEPR.

Bessen, James, Stephen Michael Impink, Lydia Reichensperger, and Robert Seamans (2020), "Gdpr and the importance of data to ai startups." Technical report, SSRN Working Paper.

Bloom, Nicholas, Mirko Draca, and John Van Reenen (2016), "Trade induced technical change? the impact of chinese imports on innovation, it and productivity." *The review of economic studies*, 83, 87–117.

Bombardini, Matilde, Bingjing Li, and Ruoying Wang (2017), "Import competition and innovation: Evidence from china." *Researchpaper: http://www7. econ. hitu. ac. jp/cces/trade_conference_2017/paper/matilde_bombardini*.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel (2022), "Quasi-Experimental Shift-Share research designs." *The Review of Economic Studies*, 89, 181–213, URL https://doi.org/10.1093/restud/rdab030.

Boyne, Shawn Marie (2018), "Data protection in the United States." *The American Journal of Comparative Law*, 66, 299–343, URL https://doi.org/10.1093/ajcl/avy016.

Carrière-Swallow, Yan and Vikram Haksar (2019), *The Economics and Implications of Data: An Integrated Perspective*. URL https://econpapers.repec.org/RePEc:imf:imfdps:2019/013.

Centre, Digital Curation, Trilateral Research, Jeni Tennison, and Maja Bogataj Jančič (2020), "The role of data in AI." Technical report, URL https://www.gpai.ai/projects/data-governance/role-of-data-in-ai.pdf.

Chang, Qing, Lin William Cong, Liyong Wang, and Longtian Zhang (2023), "Production, trade, and cross-border data flows." *Available at SSRN 4672185*.

Chen, Yongmin, Xinyu Hua, and Keith E Maskus (2021), "International protection of consumer data." *Journal of International Economics*, 132, 103517.

Coelli, Federica, Andreas Moxnes, and Karen Helene Ulltveit-Moe (2022), "Better, faster, stronger: Global innovation and trade liberalization." *Review of Economics and Statistics*, 104, 205–216.

Cong, Lin William, Wenshi Wei, Danxia Xie, and Longtian Zhang (2022), "Endogenous growth under multiple uses of data." *Journal of Economic Dynamics and Control*, 141, 104395.

Cong, Lin William, Danxia Xie, and Longtian Zhang (2021), "Knowledge accumulation, privacy, and growth in a data economy." *Management Science*, 67, 6480–6492.

De Artiñano, Ignacio Marra, Franco Riottini Depetris, and Christian Volpe Martincus (2025), "Automatic product classification in international trade: machine learning and large language models." *Review of International Economics*, 34, 3–19, URL https://doi.org/10.1111/roie.70009.

Demirer, Mert, Diego J. Jiménez Hernández, Dean Li, and Sida Peng (2024), "Data, Privacy Laws and Firm Production: Evidence from the GDPR." Technical report, URL https://www.nber.org/papers/w32146.

Farboodi, Maryam and Laura Veldkamp (2021), "A model of the data economy." Technical report, National Bureau of Economic Research Cambridge, MA, USA.

Gaulier, Guillaume and Soledad Zignago (2010), "Baci: International trade database at the product-level. the 1994-2007 version." Working Papers 2010-23, CEPII, URL https://www.cepii.fr/CEPII/fr/publications/wp/abstract.asp?NoDoc=2726.

Gellert, Raphaël (2018), "Understanding the notion of risk in the General Data Protection Regulation." *Computer law security review*, 34, 279–288, URL https://doi.org/10.1016/j.clsr.2017.12.003.

Goldberg, P. K., A. K. Khandelwal, N. Pavcnik, and P. Topalova (2010), "Imported Intermediate Inputs and Domestic Product Growth: Evidence from India." *The Quarterly Journal of Economics*, 125, 1727–1767, URL https://doi.org/10.1162/qjec.2010.125.4.1727.

Hustinx, Peter (2017), *EU Data protection Law: the review of Directive 95/46/EC and the General Data Protection Regulation*. URL https://doi.org/10.1093/acprof:oso/9780198807216.003.0005.

Igna, Ioana and Francesco Venturini (2023), "The determinants of ai innovation across european firms." *Research Policy*, 52, 104661.

Johnson, Garrett (2022), "Economic research on privacy regulation: Lessons from the gdpr and beyond." Technical Report 30705, NBER.

Jones, Charles I and Christopher Tonetti (2020), "Nonrivalry and the economics of data." *American Economic Review*, 110, 2819–2858.

Lybbert, Travis J. and Nikolas J. Zolas (2014), "Getting patents and economic data to speak to each other: An "algorithmic links with probabilities" approach for joint analyses of patenting and economic activity." *Research Policy*, 43, 530–542.

Melitz, Marc J (2003), "The impact of trade on intra-industry reallocations and aggregate industry productivity." *econometrica*, 71, 1695–1725.

Meng, Ning and Feicheng Wang (2023), "Navigating Trade Policy Shocks: How firms reallocate exports in third markets." *SSRN Electronic Journal*, URL https://doi.org/10.2139/ssrn.4636011.

Mukherjee, Arpita (2013), "The service sector in india." Technical Report 352, Asian Development Bank, Manila.

Mulyukova, Alina (2025), "Services liberalization and product variety of manufacturing firms: Evidence from india." Technical report, Kiel Institute for the World Economy and University of Göttingen.

NASSCOM (2025), "Patent pulse 2025: Decoding india's ascent in the ai patent landscape." Technical report, National Association of Software and Service Companies, URL https://nasscom.in/knowledge-center/publications/patent-pulse-2025-decoding-indias-ascent-ai-patent-landscape.

United Nations Statistics Division (2002), "Classification by broad economic categories, rev. 4." URL https://unstats.un.org/unsd/classifications/Family/Detail/10. Series M, No. 53, Rev. 4. United Nations, New York.

Voigt, Paul and Axel von dem Bussche (2017), "The eu general data protection regulation (gdpr): A practical guide."

World Intellectual Property Organization (2025), "World intellectual property indicators 2025." Technical report, WIPO, URL https://www.wipo.int/en/web/ip-indicators.

World Trade Organization (2024), "Trading with Intelligence: How AI shapes and is shaped by international trade." Technical report.

Supplementary Material

A The General Data Protection Regulation (GDPR)

This appendix provides additional institutional detail on the GDPR and the main compliance requirements imposed on firms. The discussion focuses on the provisions governing the collection, processing, storage, and protection of personal data. The section is compiled from the GDPR legal text as well as Demirer et al. (2024), Johnson (2022), Voigt and von dem Bussche (2017), and Aridor et al. (2020).

A.1 Main Compliance Requirements

The GDPR introduced a broad set of obligations for organizations that process personal data. The requirements most relevant to firms are summarized below.

Definition of Controller and Processor (Article 4). The GDPR distinguishes between data controllers and data processors. A controller determines the purposes and means of processing personal data, while a processor processes personal data on behalf of a controller. A firm may act as a controller when it determines how personal data are collected and used for its own business purposes, or as a processor when it processes personal data on behalf of another organization, such as a client. The GDPR establishes distinct obligations for each role. In particular, controllers are responsible for ensuring that processors provide sufficient guarantees to implement appropriate technical and organizational measures for the protection of personal data, and the relationship between the two parties must be governed by a written contract.

Records of Processing Activities (Article 30). Controllers and processors must maintain records describing their processing activities, including the purposes of processing, the categories of personal data collected, recipients of the data, retention periods, and the technical and organizational measures implemented to protect personal data.

Technical and Organizational Measures (Article 32). Organizations are required to implement appropriate technical and organizational measures to ensure the security of personal data. The GDPR follows a risk-based approach and does not prescribe specific technologies, instead requiring firms to adopt safeguards appropriate to the nature of their processing activities.

Data Subject Rights (Articles 15–21). Individuals are granted rights over their personal data, including the rights to access, rectify, erase, restrict, and transfer their personal information. Firms must establish procedures allowing them to respond to these requests within the deadlines established by the regulation.

Data Protection Impact Assessments (Article 35). Processing activities that are likely to result in high risks to the rights and freedoms of individuals require a Data Protection Impact Assessment before implementation. These assessments document the purpose of the processing activity, evaluate potential risks, and identify measures adopted to mitigate them.

Data Breach Notification (Article 33). Organizations must notify supervisory authorities of qualifying personal data breaches within 72 hours of becoming aware of the incident. Where breaches are likely to create high risks for affected individuals, those individuals must also be informed without undue delay.

Administrative Penalties (Article 83). The GDPR authorizes supervisory authorities to impose administrative fines of up to 20 million or 4% of a firm's worldwide annual turnover, whichever is higher, depending on the nature and severity of the infringement.

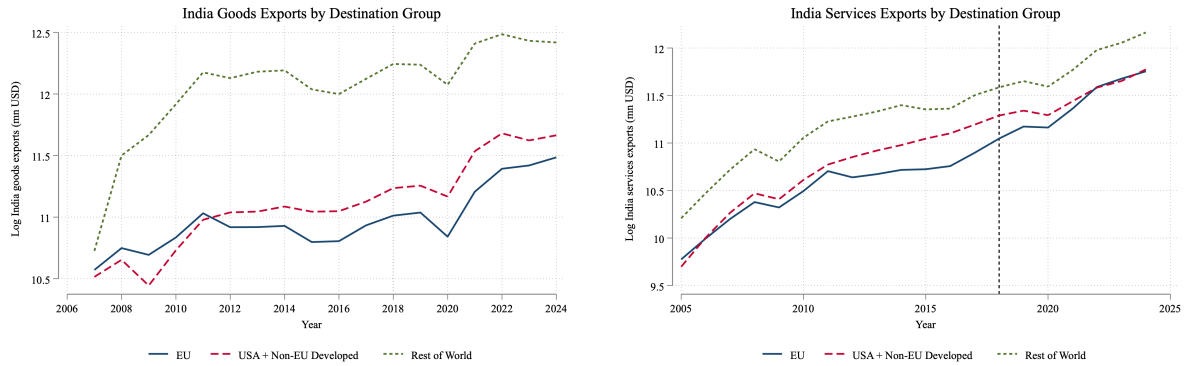
A.2 The GDPR and International Trade

An important feature of the GDPR is that it regulates the collection, processing, storage, and protection of personal data rather than international trade itself. The regulation does not impose tariffs, quotas, rules of origin, or other border measures that directly affect firms' ability to export goods or services to the European Union. Instead, it establishes legal obligations governing how organizations process personal data in the course of their activities.

The applicability of the GDPR depends on whether organizations process the personal data of individuals located in the European Union, rather than on the act of exporting itself. Firms engaged in international trade may therefore become subject to the regulation when their commercial activities involve the collection or processing of personal data relating to individuals located in the European Union. Consequently, although compliance with the GDPR may increase the cost of serving European customers, these costs arise from legal obligations governing the processing of personal data rather than from measures regulating cross-border trade itself.

One potential concern is that the additional compliance costs associated with the GDPR could induce firms to reduce exports to the European Union and redirect sales toward alternative foreign markets. Such a mechanism would resemble the export reallocation documented following tariffs and other trade policy shocks, where firms substitute exports across destinations. Figure 1 examines this possibility using aggregate Indian exports by destination group. We distinguish exports to the European Union, exports to the United States and other developed non-EU economies, and exports to the rest of the world. For both goods and services, exports to the European Union continue to follow trends similar to those observed for the other destination groups after the GDPR entered into force in 2018. There is no visible aggregate decline in exports to the European Union accompanied by a corresponding increase in exports to alternative destinations.

These aggregate patterns cannot rule out firm-level adjustments in export destinations. However, they do not indicate a substantial aggregate reallocation of Indian exports away



(a) Goods exports

(b) Services exports

Figure 1: **Aggregate Indian Exports by Destination Group**

Notes: The figures report the logarithm of India's aggregate goods and services exports by destination group. Destination groups are the European Union (EU), the United States and other developed non-EU economies, and the rest of the world. The vertical dashed line denotes the implementation of the GDPR in May 2018.

from the European Union following the implementation of the GDPR.

A.3 Compliance Costs

Compliance with the GDPR entails both one-time implementation costs and recurring operational costs associated with the processing of personal data. Initial compliance requires firms to establish data governance procedures, map data flows, prepare records of processing activities, update privacy notices, implement technical and organizational safeguards, train employees, and, where applicable, appoint a Data Protection Officer. These investments represent fixed or sunk costs associated with implementing the regulatory framework.

Beyond the initial implementation, firms incur ongoing compliance costs throughout their data processing activities. Organizations must maintain records of processing activities, respond to requests for data access, rectification, portability, restriction, and erasure, conduct Data Protection Impact Assessments for new processing activities when required, maintain appropriate technical and organizational security measures, and monitor continued compliance with the regulation.

Several compliance costs depend on the scale and complexity of firms' data processing activities. Organizations that process larger volumes of personal data are generally more likely to receive requests from data subjects and must devote additional resources to documenting processing activities, maintaining secure data storage, conducting assessments for new processing operations, and ensuring continued compliance.

B Data

B.1 Patent data

We augment the CMIE Prowess firm-level database with patent information from the Indian Patent Office (IPO). The patent data are compiled from web-scraped records of the IPO public patent search portal and cover applications filed between January 2000 and March 2025. The resulting dataset contains 286,594 patent applications with information on applicant names, filing and publication dates, International Patent Classification (IPC) codes, invention titles, and field-of-invention descriptions. Patent records are linked to CMIE firms using fuzzy matching on applicant names. To balance coverage and precision, we retain matches with a similarity score of at least 80%, yielding 48,221 patent–firm links. We additionally construct a more conservative firm-level patent indicator using a 90% similarity threshold as a robustness check.

Classification of AI and ICT Patents Patents are classified into AI and ICT categories based on their IPC codes. AI patents are identified following the methodology of Baruffaldi et al. (2020) and the OECD AI patent taxonomy, which map IPC subclasses to artificial intelligence technologies. ICT patents are identified using a broader collection of IPC classes covering information and communication technologies, including telecommunications, computing, software, data processing, and electronic systems. To improve classification accuracy, particularly for the relatively small AI sample, we manually review patent titles and field-of-invention descriptions to remove obvious false positives. Patents not classified as AI or ICT are grouped into an "Other" category.

Patenting activity is uncommon among firms in the CMIE sample, and AI innovation is especially rare. Table A1 reports the prevalence of patenting across different samples. In the full sample, only 1.91% of firms hold at least one patent, while 0.09% possess an AI patent and 0.44% possess an ICT patent. Similar patterns emerge when restricting the sample to manufacturing firms or to the pre-GDPR period used in the main analysis. The scarcity of AI patents is an important feature of the empirical setting and distinguishes our context from the patent-intensive environments often studied in the innovation and trade literature.

Table A1: Prevalence of Patenting Across Samples

Sample	Number of Firms	Any Patent	AI Patent	ICT Patent
All sectors	55,029	1,049 (1.91%)	48 (0.09%)	241 (0.44%)
Manufacturing (NIC 10–33)	16,683	688 (4.12%)	15 (0.09%)	140 (0.84%)
Pre-GDPR (2013–2017), all sectors	39,337	940 (2.39%)	46 (0.12%)	214 (0.54%)
Pre-GDPR (2013–2017), manufacturing	11,374	636 (5.59%)	15 (0.13%)	128 (1.13%)

Notes: The table reports the prevalence of patenting across different samples. Entries in the patent columns show the number of firms with at least one patent in the corresponding category, with the share of firms in parentheses. AI and ICT categories are not mutually exclusive at the firm level, as a firm may hold patents of multiple types. The pre-GDPR sample includes observations from 2013–2017. Sources: CMIE Prowess, Indian Patent Office, and authors' calculations.

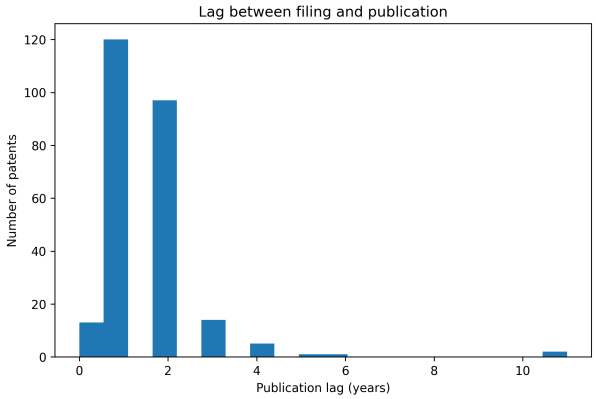


Figure 2: Distribution of time between filing and publication of Patent

Publication Lag Patent publication occurs with a delay relative to filing. In the matched sample, the median lag is approximately 23 months, while 75% of patents are published within two years of filing. Nevertheless, the distribution exhibits a long right tail, with some patents taking substantially longer to appear in the public records.

Table A2 summarizes the distribution of publication lags. Throughout the analysis, innovation timing is based on filing dates when available, thereby minimizing distortions due to publication delays.

Table A2: Publication Lag of Patents

Statistic	Lag (Years)	Lag (Months)
Observations	253.00	253.00
Mean	1.62	25.67
Standard deviation	1.19	13.53
Minimum	0.00	2.00
25th percentile	1.00	20.00
Median	1.00	23.00
75th percentile	2.00	29.00
Maximum	11.00	139.00

Notes: The table reports descriptive statistics for the lag between the filing year and the publication date of patents. The month-based measure is approximate because filing information is available only at the yearly level.

Qualitative Validation of Patent Categories To provide a qualitative assessment of our patent classification, Figures 3 and 4 present word clouds based on the field of invention and invention title variables, respectively. While both representations capture similar technological themes, the field descriptions provide broader contextual information, whereas invention titles emphasize the core technical contribution of each patent.

The AI patents are characterized by terms related to software, data processing, and intelligent systems. Frequent words include *computer*, *management*, *monitoring*, *graphical*, *representation*, *virtual*, *planning*, and *tabular*, suggesting applications involving information management, visualization, decision support, and computational methods. In contrast, ICT patents are dominated by vocabulary associated with communications and electronic technologies, in-

cluding *video*, *communication*, *network*, *electronic*, *control*, *signal*, *module*, and *cell*. The rest of the patents combine concepts from both domains. Prominent terms such as *method*, *apparatus*, *device*, *communication*, and *machine* reflect the broad technological scope of these innovations and the prevalence of standard patent language. Despite this overlap, the visualizations reveal distinct vocabularies for AI and ICT inventions, providing an intuitive validation of the IPC-based classification used throughout the paper. Overall, AI patents appear to be concentrated in computational and software-oriented innovations, whereas ICT patents span a wider range of communications, networking, electronic, and hardware-related technologies.

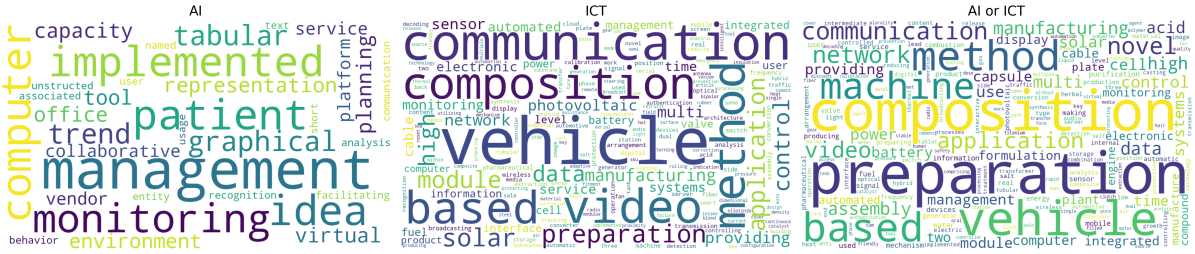


Figure 3: Word Clouds Based on the `FieldOfInvention` Description

Notes: Word clouds are constructed from the `FieldOfInvention` descriptions after removing common English stopwords and generic punctuation. The three panels correspond to patents classified as AI, ICT, and the union of AI or ICT patents. Word size is proportional to frequency within each category.

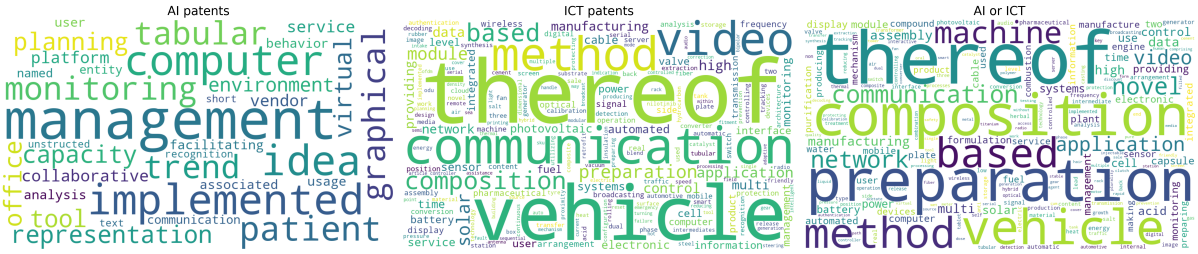


Figure 4: Word Clouds Based on Patent `InventionTitle`

Notes: Word clouds are constructed from patent titles after removing common English stopwords and generic patent terminology. AI patents are identified using IPC classifications following Baruffaldi et al. (2020) and subsequently manually validated to remove false positives. ICT patents are identified using IPC codes corresponding to information and communication technologies. Word size is proportional to frequency within each category.

Limitations Three limitations should be noted. First, the IPO public search portal may not provide complete historical coverage, implying that some patent applications are unobserved. Second, despite using conservative fuzzy-matching thresholds, linking applicant names to firms inevitably introduces some measurement error. Third, IPC classifications do not perfectly map into technological content, particularly for AI-related inventions, although manual validation substantially reduces this concern. Given the rarity of patenting in our sample, these issues are unlikely to affect the qualitative interpretation of the results but should be borne in mind when interpreting estimates based on narrow innovation categories.

B.2 LLM-Assisted CMIE-to-Trade-Code Crosswalk

This appendix describes the methodology used to construct a novel crosswalk between CMIE Prowess product descriptions and internationally recognized trade classifications: the Har-

monized System (HS6) for goods and the Extended Balance of Payments Services classification (EBOPS 2010) for services. Such a crosswalk is necessary because CMIE records firms' sales using free-text product descriptions that are often highly idiosyncratic, relying on firm-specific terminology, abbreviations, brand names, or accounting labels (e.g., `"MS PRODUCT"`, `"PUMP REPAIR SERVICES"`, or `"B-BLACK"`). Consequently, these descriptions cannot be directly mapped to standard international trade classifications using existing concordances.

The underlying data contain more than two million firm-year-product observations. After collapsing to unique combinations of product descriptions and industry classifications, the sample consists of 181,439 distinct product–industry triplets, corresponding to approximately 103,000 unique textual descriptions. Because identical descriptions may appear in different industries and refer to different underlying products, industry information is retained throughout the classification process to improve mapping accuracy and allow context-specific assignments.

To classify these observations, we develop a four-stage pipeline that combines semantic retrieval, cross-encoder reranking, deterministic preprocessing rules, and Large Language Model (LLM) adjudication. The resulting framework provides a scalable and reproducible approach for mapping firm-reported product descriptions to internationally comparable trade classifications and forms the basis for the product-level analyses in this paper.

Pre-LLM Deterministic Filtering Before applying semantic retrieval or LLM-based classification, we implement a sequence of deterministic preprocessing steps designed to identify observations whose category can be established without model inference. These filters reduce computational costs, improve the efficiency of the downstream pipeline, and minimize the risk of classifying observations that do not correspond to tradable goods or services.

We begin by standardizing all descriptions through basic text normalization, including lowercasing, removal of punctuation (except hyphens), whitespace normalization, and trimming. The resulting cleaned text serves as the input for all subsequent stages. Next, we exclude observations corresponding to accounting or profit-and-loss items rather than marketable outputs. Using a set of regular-expression rules, we identify descriptions referring to concepts such as interest income, dividends, royalties, rental income, foreign-exchange gains and losses, provisions, reimbursements, and related financial entries. These account for 22,802 product–industry triplets (12.6% of the corpus) and are removed before classification. We then separate observations corresponding to services from those representing goods. First, products belonging to industries classified as service sectors under the NIC 2008 taxonomy are routed directly to the services pipeline. Second, we apply a keyword-based screen to identify service activities embedded within goods-producing industries, such as installation, consulting, repair, maintenance, or warehousing. These observations are likewise redirected to the services classifier. Finally, we exclude descriptions that are too generic to support meaningful classification (e.g., "Others" or "Miscellaneous") together with observations whose descriptions are empty or effectively uninformative after text cleaning. Such cases are flagged separately and omitted from the goods-classification procedure.

The remaining observations enter the four-stage retrieval and classification cascade described below.

The Four-Stage Classification Cascade After deterministic filtering, the remaining observations are processed through a four-stage classification cascade designed to maximize accuracy while minimizing the need for LLM inference. The first three stages rely on retrieval-based methods and deterministic decision rules, reserving LLM adjudication for ambiguous cases.

Stage 1: Semantic retrieval. Each product description is embedded using a bi-encoder model and matched to the HS6 reference catalogue based on cosine similarity. For every observation, we retrieve the ten most similar candidate codes.

Stage 2: Cross-encoder reranking. The retrieved candidates are reranked using a cross-encoder model that jointly evaluates the product description and each candidate code. We retain the five highest-scoring candidates, with scores transformed to the unit interval to facilitate threshold-based decisions.

Stage 3: Threshold-based routing. Observations are assigned to one of three groups. If the highest-ranked candidate is sufficiently confident and clearly separated from the runner-up, the observation is automatically assigned its HS6 code. If this criterion fails but all top candidates belong to the same HS2 chapter, the observation is assigned at the HS2 level. All remaining cases are forwarded to the LLM for adjudication. Thresholds were calibrated on a hand-labeled development sample to maximize automatic classification accuracy.

Stage 4: LLM adjudication. The final stage of the classification pipeline uses Anthropic's `claude-sonnet-4-5` model. For each observation, the model receives the original CMIE product description together with the top semantic retrieval candidates and is asked to assign the most appropriate HS classification. The retrieved candidates are intended to assist the model but are not treated as binding, allowing the LLM to override them when they are inconsistent with the underlying product description.

The production prompt was developed iteratively through multiple rounds of testing and refinement on held-out validation samples. It instructs the model to rely on expertise in the Harmonized System, international trade statistics, Indian manufacturing products, and CMIE Prowess descriptions while recognizing that the input data may contain abbreviations, OCR errors, trade names, service revenues, accounting items, or highly generic labels. In particular, the prompt emphasizes that the model should never invent HS codes, should prefer broader but correct classifications (e.g., HS2 or HS4) over uncertain HS6 assignments, and should distinguish physical goods from services, income items, accounting entries, and overly generic descriptions. The opening instructions of the production prompt are reproduced below:

> *You are an expert in international trade classification using the Harmonized System (HS) at the 6-digit subheading level. You classify product descriptions from Indian firm-level data into HS6 codes. When a description cannot be classified as a good, you must also categorize WHY – distinguishing too-vague product names (where the general category is recoverable) from services, income, and accounting items (where it is not). NEVER invent HS codes. PREFER broader-but-correct categories (HS2 or HS4) over narrower-but-wrong ones (HS6).*

All LLM outputs are validated against the official HS 2007 catalogue. If the returned six-

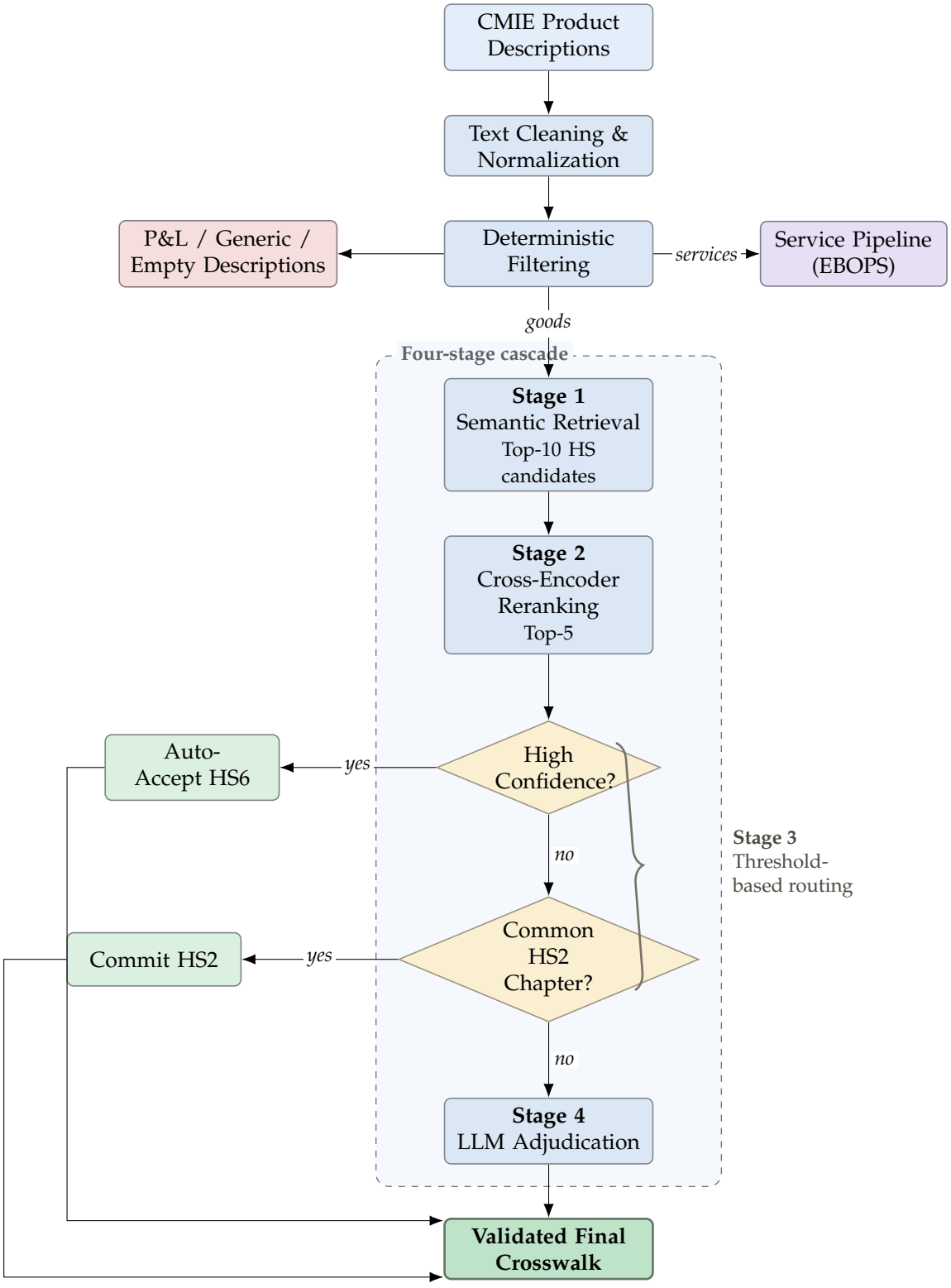


Figure 5: Overview of the CMIE-to-trade-code classification pipeline. Product descriptions are standardized and deterministically filtered to remove accounting items and generic or empty descriptions (red) and to divert service observations to the EBOPS classifier (purple). Remaining goods descriptions enter a four-stage cascade: semantic retrieval (Stage 1), cross-encoder reranking (Stage 2), threshold-based routing (Stage 3), and LLM adjudication (Stage 4). High-confidence matches are auto-accepted at HS6 and common-chapter agreements are committed at HS2; all accepted assignments feed the validated final crosswalk.

digit code is invalid but its four-digit heading exists, the observation is retained at the HS4 level. Otherwise, if only the corresponding two-digit chapter is valid, it is retained at the HS2 level. Observations without any valid HS prefix are excluded from the final goods crosswalk. This hierarchical fallback substantially reduces information loss while preventing the use of invalid HS codes.

For observations identified as services during the deterministic filtering stage are classified separately using the Extended Balance of Payments Services (EBOPS 2010) nomenclature. The services pipeline mirrors the goods-classification procedure, replacing the HS reference catalogue with the corresponding EBOPS taxonomy while using the same retrieval, reranking, and `claude-sonnet-4-5` adjudication framework. Travel-related EBOPS categories are excluded because they correspond to expenditures by individuals rather than firm-reported operating revenues.

Finally, we assign each observation a composite confidence tier (*high*, *medium*, or *low*) based on three factors: the precision of the assigned code (HS6, HS4, or HS2), whether the selected code appears among the top reranked candidates, and the degree of ambiguity reflected in the LLM output. This measure is independent of the model's self-reported confidence and is strongly correlated with manual validation accuracy. In the empirical analysis, we use the confidence tiers to distinguish highly reliable HS6 matches from coarser but still informative HS4- and HS2-level classifications.

Crosswalk Assembly The outputs from the deterministic routing stages and the LLM classifier are combined into a single master crosswalk keyed by the product description and its associated industry classification. For each observation, the crosswalk records the assigned HS code (when available), the level of classification precision (HS6, HS4, or HS2), the routing method, the composite confidence tier, and any applicable rejection category.

The resulting master file contains 181,439 unique product–industry triplets distributed across the status categories reported in Table A3.

Table A3: Crosswalk Status Distribution (Goods + Services)

Status	N triplets	% of total
Classified good	79,567	43.85
Classified service	43,185	23.8
PL	25,130	13.85
Rejected good	19,804	10.91
Not found	7,254	4.00
Rejected service	4,161	2.29
Too generic	2,022	1.11
Empty	316	0.17
Total	**181,439**	**100.00**

As a final consistency check, all assigned HS codes are validated against the official HS 2007 nomenclature and standardized to preserve leading zeros where required. This validation step ensures that all retained codes correspond to valid entries in the reference catalogue.

Coverage and Heterogeneity Diagnostics The final crosswalk combines broad coverage with substantial within-industry heterogeneity. Table A4 reports coverage by NIC 2-digit sector. Across manufacturing industries, sales-weighted coverage is consistently higher than unweighted coverage, indicating that unmatched observations are concentrated in relatively low-revenue product lines rather than firms' core activities. By construction, service industries exhibit lower HS coverage because they are routed to the EBOPS classification pipeline.

Table A4: Crosswalk Coverage by NIC 2-Digit Sector: Goods (HS) and Services (EBOPS)

			Unweighted Coverage (%)			Sales-Weighted Coverage (%)	
NIC	Industry	N	HS	EBOPS	Total	HS	Total
01	Crop and animal production, hu	3,668	59.2	4.9	64.1	89.0	91.0
02	Forestry and logging	150	61.3	3.3	64.7	96.2	96.5
05	Mining of coal and lignite	119	42.0	11.8	53.8	31.5	44.6
06	Extraction of crude petroleum	117	35.0	8.5	43.6	92.5	92.6
07	Mining of metal ores	429	49.4	7.7	57.1	86.8	88.5
08	Other mining and quarrying	1,149	62.1	5.0	67.0	90.4	91.8
10	**Food products**	8,121	64.7	4.3	69.0	83.3	84.4
11	**Beverages**	1,787	57.2	5.7	62.9	82.5	83.2
12	**Tobacco products**	264	56.8	5.7	62.5	87.4	87.7
13	**Textiles**	5,023	58.3	7.2	65.5	60.6	65.8
14	**Wearing apparel**	1,075	52.7	7.5	60.3	70.0	79.5
15	**Leather and related products**	693	61.9	4.0	65.9	76.0	77.1
16	**Wood and wood products**	681	66.5	4.4	70.9	92.8	94.2
17	**Paper and paper products**	1,607	56.9	6.1	63.0	85.1	86.6
18	**Printing and recorded media**	228	36.8	14.9	51.8	58.5	79.6
19	**Coke and refined petroleum**	958	59.6	6.8	66.4	49.1	49.3
20	**Chemicals and chemical products**	11,009	65.8	4.0	69.7	72.4	74.8
21	**Basic pharmaceuticals**	3,820	66.0	4.3	70.3	67.3	69.2
22	**Rubber and plastic**	4,974	66.5	4.9	71.5	65.4	67.2
23	**Other non-metallic mineral**	2,533	57.6	5.8	63.4	92.7	93.0
24	**Basic metals**	7,880	67.0	4.6	71.6	60.9	64.6
25	**Fabricated metal products**	3,312	59.1	8.2	67.3	75.2	82.8
26	**Computer and electronic products**	4,255	61.0	7.8	68.8	76.1	79.4
27	**Electrical equipment**	5,448	65.7	6.8	72.5	77.1	80.9
28	**Machinery and equipment n.e.c.**	8,339	61.9	7.8	69.8	78.1	81.4
29	**Motor vehicles**	4,457	66.5	4.7	71.3	78.7	80.5
30	**Other transport equipment**	1,079	55.5	8.1	63.6	56.9	58.2
31	**Furniture**	156	55.1	9.0	64.1	79.7	82.5
32	**Other manufacturing**	1,805	57.1	6.4	63.5	75.2	77.8
34	(unknown)	4,865	18.9	5.7	24.5	0.0	0.0
35	Electricity, gas, steam and air conditioning	1,411	27.4	11.6	39.0	25.1	28.2
36	Water collection, treatment and supply	29	17.2	17.2	34.5	43.6	45.4
37	Sewerage	36	25.0	19.4	44.4	45.9	81.0
38	Waste collection and treatment	14	7.1	14.3	21.4	1.0	23.0
41	Construction of buildings	3,560	25.4	14.0	39.5	7.9	25.8
42	Civil engineering	3,717	32.6	18.3	50.9	8.1	56.4
43	Specialized construction activities	174	17.8	16.7	34.5	19.1	55.4
45	Trade and repair of motor vehicles	1,399	43.7	9.9	53.6	79.5	80.8
46	Wholesale trade	30,289	54.3	9.2	63.6	54.2	58.1
47	Retail trade	3,326	43.1	8.5	51.7	42.7	45.2
49	Land transport and pipelines	940	4.9	20.6	25.5	0.8	50.3
50	Water transport	459	3.3	28.5	31.8	0.0	0.0
51	Air transport	338	2.1	20.7	22.8	11.1	17.9
52	Warehousing and transport support	2,604	7.7	19.5	27.2	30.7	41.8
53	Postal and courier activities	130	3.1	22.3	25.4	0.0	25.7
55	Accommodation	1,674	5.2	6.9	12.1	5.2	7.9
56	Food and beverage service activities	81	11.1	4.9	16.0	43.2	43.7
58	Publishing activities	603	5.8	9.3	15.1	15.5	21.5
59	Motion picture, video and TV activities	806	4.1	16.9	21.0	0.5	36.4
60	Broadcasting activities	423	5.9	14.4	20.3	0.0	45.2
61	Telecommunications	1,322	4.2	16.9	21.1	0.2	45.8
62	Computer programming and consultancy	2,401	4.6	25.9	30.5	0.7	74.7
63	Information service activities	951	3.0	21.7	24.7	0.8	12.1
64	Financial services, except insurance	19,379	11.9	8.9	20.8	4.6	16.1
65	Insurance, reinsurance and pension funding	87	1.1	24.1	25.3	0.0	24.9

Table A4: Crosswalk Coverage by NIC 2-Digit Sector: Goods (HS) and Services (EBOPS) (continued)

			Unweighted Coverage (%)			Sales-Weighted Coverage (%)	
NIC	Industry	N	HS	EBOPS	Total	HS	Total
66	Other financial activities	1,940	1.8	14.8	16.6	0.6	14.7
68	Real estate activities	38	0.0	18.4	18.4	0.0	91.5
69	Legal and accounting	53	1.9	26.4	28.3	0.0	32.4
70	Head offices and management co	1,701	9.3	17.3	26.6	1.4	29.9
71	Architecture and engineering	829	7.7	20.9	28.6	0.9	68.1
72	Scientific research and develo	300	8.7	15.0	23.7	0.0	14.2
73	Advertising and market researc	615	1.8	19.8	21.6	1.1	50.9
74	Other professional and technic	39	0.0	17.9	17.9	0.0	32.9
77	Rental and leasing activities	4,607	14.6	9.3	23.9	0.0	0.0
78	Employment activities	470	5.1	16.4	21.5	3.0	13.6
79	Travel agency and reservation	368	1.9	6.8	8.7	0.0	3.8
80	Security and investigation	142	4.9	14.8	19.7	0.0	0.0
82	Office administrative and supp	669	4.5	17.6	22.1	0.4	21.5
84	Public administration and defe	279	3.9	9.3	13.3	8.0	21.2
85	Education	760	2.4	23.4	25.8	0.2	22.6
86	Human health activities	1,211	5.0	14.5	19.4	1.1	34.8
90	Creative, arts and entertainme	107	2.8	8.4	11.2	0.0	0.0
93	Sports and recreation activiti	343	5.8	12.8	18.7	22.7	23.9
94	Membership organizations	348	1.7	6.3	8.0	0.0	2.9
95	Repair of computers and person	435	8.3	20.5	28.7	9.3	40.2
96	Other personal service activit	31	0.0	6.5	6.5	0.0	0.0

Notes: Coverage = share of (product, NIC) triplets classified. Unweighted: share of triplets. Sales-weighted: triplets weighted by total CMIE sales. Goods = HS classification (BACI compatible). Services = EBOPS (BaTIS compatible). **Manufacturing sectors (NIC 10-33) in bold.** Source: CMIE Prowess, authors' calculations.

Table A5 reports the distribution of service classifications across EBOPS categories. The classified service observations span a broad range of internationally traded activities, with the largest categories including manufacturing services on physical inputs owned by others (15.4%), technical and other business services (13.7%), professional and management consulting services (12.4%), construction (11.4%), computer services (9.2%), transport (8.3%), and financial services (6.1%). These categories account for the majority of classified service observations. The distribution indicates that the classification pipeline captures a diverse set of internationally traded services rather than concentrating observations in a small number of broad residual categories. In particular, the prominence of business, consulting, computer, and technical services is consistent with the composition of India's internationally integrated service sector and provides the level of detail required to construct service-specific foreign demand measures using the EBOPS classification.

Table A6 reports the level of precision achieved by the classification pipeline. Among classified goods, approximately 40% of observations are assigned directly to the most detailed HS6 level, while the remaining goods are classified at the broader HS4 (21.4%) or HS2 (22.4%) levels when the available information does not support a more granular assignment. For services, classifications are assigned either at the detailed EBOPS Level 2 (7.9%) or, where appropriate, at the broader EBOPS Level 1 (8.8%).

This distribution reflects the conservative design of the classification pipeline. Rather than forcing highly detailed assignments when product descriptions are ambiguous, the algorithm assigns observations to broader but more reliable classifications. This approach prioritizes

Table A5: EBOPS Code Distribution (Classified Services Triplets)

Code	Description	N Triplets	%
SA	Manufacturing services on physical inputs own	2,460	15.35
SJ3	Technical, trade-related, and other business	2,192	13.68
SJ2	Professional and management consulting servic	1,980	12.36
SE	Construction	1,828	11.41
SI2	Computer services	1,479	9.23
SB	Maintenance and repair services n.i.e.	1,402	8.75
SC	Transport	1,325	8.27
SG	Financial services	980	6.12
SK2	Personal, cultural, and recreational services	781	4.87
SH	Charges for the use of intellectual property	337	2.10
SK1	Audiovisual and related services	281	1.75
SC3	Other modes of transport	211	1.32
SI1	Telecommunications services	183	1.14
SC1	Sea transport	160	1.00
SJ1	Research and development services	152	0.95
SF	Insurance and pension services	88	0.55
SC2	Air transport	65	0.41
SC4	Postal and courier services	54	0.34
SI3	Information services	33	0.21
SJ	Other business services	18	0.11
SI	Telecommunications, computer, and information	11	0.07
SK	Personal, cultural, and recreational services	2	0.01
Total		**16,022**	**100.00**

classification accuracy while preserving as much product-level information as possible for both goods and services.

Table A6: Match Precision: HS (Goods) and EBOPS (Services)

Type	Match Level	N Triplets	% of Classified
Goods (HS)	HS6	37,729	39.47
Goods (HS)	HS4	20,471	21.42
Goods (HS)	HS2	21,367	22.35
Services (EBOPS)	EBOPS L2	7,571	7.92
Services (EBOPS)	EBOPS L1	8,451	8.84
Total		**95,589**	**100.00**

The main advantage of the LLM-based approach over conventional concordances is its ability to preserve within-industry product heterogeneity. As shown in Table A7, a baseline assigning each NIC4D industry its modal HS6 code uses only 47 distinct HS6 codes across the sample, whereas the proposed approach identifies 3,391 distinct HS6 codes. Figures 6 and 7 illustrate the richness of the resulting classifications. The crosswalk recovers substantial within-industry heterogeneity across virtually all sectors, with particularly large numbers of distinct HS6 codes in wholesale trade, chemicals, machinery and equipment, and basic metals, and a broad distribution of EBOPS categories in financial services, wholesale and retail trade, machinery and equipment, and other business services. These patterns highlight the ability of the crosswalk to distinguish firms producing different goods and services within the same industry.

Figure 8 provides complementary evidence at the firm level. The distribution based on

Table A7: LLM Crosswalk vs. NIC4D Baseline (All Sectors)

	LLM	NIC4D Baseline
Coverage (%)	27.2	31.8
N triplets covered	33,827	39,567
HS6-precise (% of covered)	27.2	100.0
Distinct HS6 codes used	3,391	47

Notes: The baseline assigns the modal HS6 code to each NIC4D industry using average Indian exports over 2007–2009 from BACI. The key distinction is the number of distinct HS6 codes employed: the LLM-based crosswalk preserves substantial within-NIC4D heterogeneity, whereas the baseline collapses most products to a single HS6 code. Source: CMIE Prowess and BACI HS07 V202601.

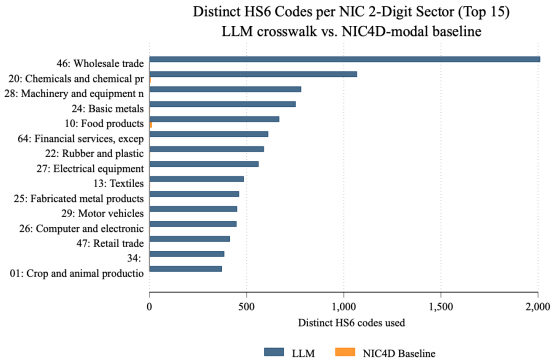


Figure 6: HS6 codes by NIC2D sector

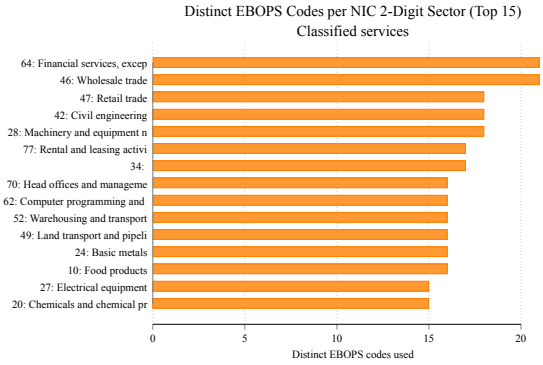


Figure 7: EBOPS by NIC2D sector

sales-weighted coverage lies well above the distribution based on the share of products classified, implying that products left unmatched account for only a small fraction of firms' revenues. In other words, although some niche or generic product descriptions remain unclassified, the crosswalk captures the products that are economically most important for the typical firm

Table A8 examines whether crosswalk coverage differs systematically across firms with different innovation profiles. Coverage is generally comparable across innovation groups within each broad sector, suggesting that the quality of the product classifications is not driven by firms' innovation status. As expected, manufacturing firms exhibit substantially higher HS coverage than service firms, reflecting the greater applicability of the Harmonized System to traded goods, while the inclusion of EBOPS classifications considerably improves overall coverage for service industries. Sales-weighted coverage consistently exceeds unweighted coverage across sectors and innovation groups, indicating that unmatched observations are concentrated among relatively low-revenue products. Overall, these patterns suggest that the crosswalk provides broad and balanced coverage across both innovators and non-innovators.

Figure 9 compares the distribution of firm-level crosswalk coverage across innovation groups. The left panel reports the share of products successfully classified, while the right panel weights products by firm sales. The distributions for AI innovators, ICT innovators, and other innovative firms are remarkably similar under both measures, indicating that classification perfor-

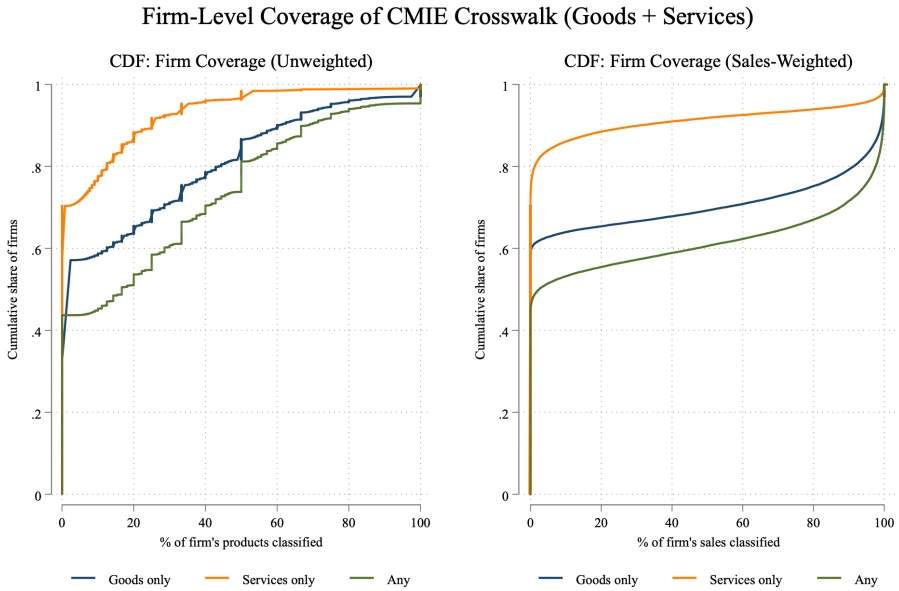


Figure 8: Coverage by sector

Table A8: Crosswalk Coverage by Innovation Status and Sector Type

			Goods Cov (%)		Any Cov (%)	
Sector	Innov Group	N	Unw	SW	Unw	SW
Manufacturing	AI	15	52.5	74.4	56.9	75.7
	ICT (not AI)	128	41.2	67.0	45.7	71.5
	Other innov	545	42.5	74.8	47.6	78.4
	Non-innovator	16018	40.5	63.4	45.4	67.2
Services	AI	27	2.1	0.5	19.6	54.5
	ICT (not AI)	54	3.6	0.4	16.1	34.4
	Other innov	154	3.2	4.9	16.3	26.8
	Non-innovator	22458	2.7	3.7	11.2	16.4
Trade	AI	4	22.3	29.6	30.8	52.7
	ICT (not AI)	16	29.1	58.6	36.6	62.2
	Other innov	78	34.2	53.2	41.3	57.8
	Non-innovator	9203	28.9	45.8	35.8	51.2
Other	AI	2	62.1	20.3	69.5	70.2
	ICT (not AI)	6	7.3	1.8	16.7	8.8
	Other innov	20	16.8	18.2	27.2	33.2
	Non-innovator	6376	6.6	7.5	13.9	20.8

Notes: Mean firm-level crosswalk coverage rates by sector type and innovation status. Sector type assigned from primary NIC: Manufacturing (NIC 10-33), Services (NIC 34, 49+), Trade (NIC 45-47). Goods coverage uses HS classification; Any coverage = HS + EBOPS. Unw = unweighted (products); SW = sales-weighted.

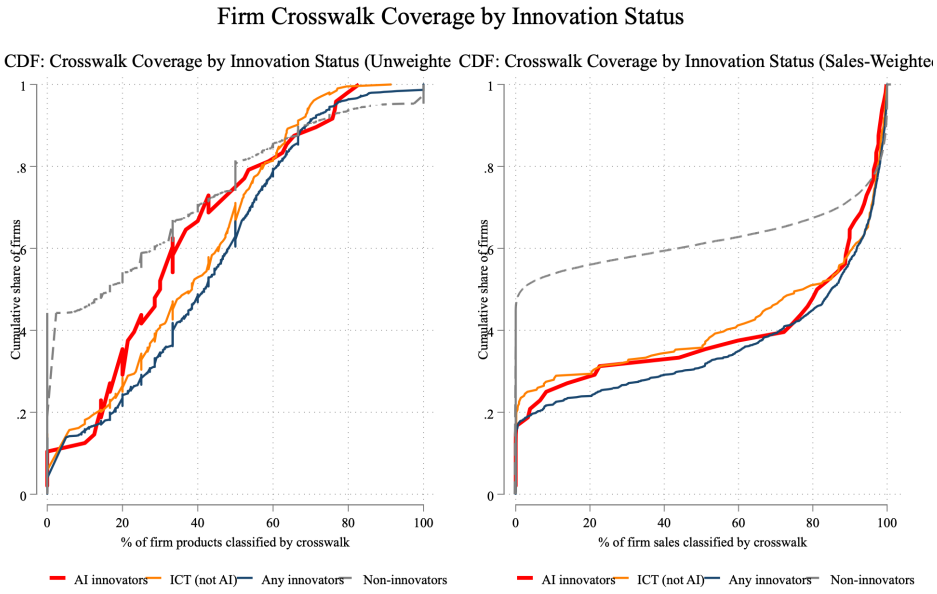


Figure 9: Coverage by innovation status

mance does not vary systematically with firms' innovation status. If anything, non-innovative firms exhibit lower coverage, particularly in the sales-weighted measure, where a larger share of firms have relatively low coverage. These patterns suggest that the crosswalk does not selectively perform better for AI innovators and that the economically most important products of innovative firms are successfully classified.

These diagnostics suggest that the proposed methodology achieves high effective coverage while preserving the product-level variation needed to construct informative measures of trade exposure and foreign demand.

Validation Diagnostics To assess the quality of the final crosswalk, we conduct a blind manual validation on a random sample of 450 product descriptions drawn from the full goods corpus. The sample is selected independently of the prompt development and threshold calibration process and labeled without access to the LLM's predictions. Overall, 89.55% of observations are classified correctly.

For each observation, we determine the appropriate HS classification and evaluate agreement with the automated assignment at the HS6, HS4, and HS2 levels. They also identify observations that should not be classified as tradable goods and, where relevant, assign them to one of the rejection categories (generic description, service, income item, or non-product).

The validation provides the primary measure of classification accuracy reported in this appendix and serves to calibrate the composite confidence tiers used in the empirical analysis

Limitations Despite the multiple validation and post-processing steps incorporated into the pipeline, the resulting crosswalk remains subject to several limitations. First, as with any LLM-based classification procedure, the model may occasionally generate invalid or overly specific HS codes. We mitigate this risk by validating every assignment against the official HS 2007

catalogue and automatically degrading invalid classifications to valid HS4 or HS2 prefixes whenever possible.

Second, the deterministic preprocessing rules used to identify accounting items, services, and generic descriptions rely on manually curated keyword and regular-expression lists. Although these rules are fully documented and reproducible, alternative specifications could lead to small differences in the set of observations entering the classification pipeline.

Finally, while the use of `temperature=0.0` substantially reduces stochasticity, LLM outputs are not guaranteed to be perfectly deterministic across independent runs. To ensure reproducibility, all production classifications are generated once, validated, and stored as a fixed crosswalk used throughout the empirical analysis.

B.3 Patent-Product Alignment

This appendix describes the construction of the patent-product alignment measure used in the paper. To characterize the relationship between firms' innovation activity and their commercial activity, we map both patents and products into a common product-classification space based on the Harmonized System (HS) and ask: *for each patent of a firm, how closely is the patent's underlying technology related to the products the firm actually sells* The measure is motivated by the idea of data accumulation: firms whose patents are concentrated in the product space of their commercial activity may be more likely to innovate using information, capabilities, and data generated through those activities, whereas firms whose patents lie farther from their observed products may instead be diversifying into technologically distinct domains. More generally, the measure provides a descriptive indicator of the proximity between firms' innovation activities and their observed product portfolios.

Mapping patents to products. We map each patent to a set of Harmonized System six-digit (HS6) product codes using the probabilistic IPC-to-HS concordance of Lybbert and Zolas (2014), which links International Patent Classification (IPC) codes to traded products. For each patent, we extract its IPC classifications from the Indian Patent Office data, truncate them to the four-character subclass level (e.g., `A61K` from `A61K 31/00`), and assign all HS6 codes with concordance weights above 0.10. This procedure yields a set of products plausibly associated with the patented technology.

Our baseline specification uses the primary (first-listed) IPC code, which typically represents the patent's main technological classification. As a robustness check, we also consider a "union" specification that pools HS6 assignments across all IPC codes reported for a patent.

Mapping products to HS codes. On the product side, we use the LLM-assisted CMIE-to-HS crosswalk described in Appendix B.2. Products are assigned at the HS6, HS4, or HS2 level depending on the precision of the match. Accordingly, we construct nested product sets for each firm and evaluate patent-product alignment at multiple levels of aggregation.

Alignment categories. For each patent, we compare its HS assignments with the firm's product portfolio and classify the relationship into one of four mutually exclusive categories:

- **Same HS6**: at least one patent HS6 code exactly matches a firm's HS6 product code.
- **Same HS4**: no HS6 match exists, but the patent and product share the same four-digit HS heading.
- **Same HS2**: no HS6 or HS4 match exists, but the patent and product belong to the same two-digit HS chapter.
- **Unrelated**: no overlap exists at any level of HS aggregation.

These categories capture progressively broader notions of technological proximity, ranging from exact product correspondence at the HS6 level to broader industry relationships at the

HS2 level. Evaluating alignment at multiple levels allows us to account for the varying precision of the CMIE-to-HS crosswalk while retaining information on whether firms' innovation activities are closely related to their observed product portfolios.

Results Table A9 reports patent-product alignment by innovation type. Alignment increases mechanically as the HS classification becomes more aggregated, with between 52% and 67% of patents matching firms' product portfolios at least at the HS2 level.

Table A9: Patent-Product Alignment by Innovation Type

Innovation Type	N firms	Same HS6	Same HS4	Same HS2	Unrelated
AI	48	0.010	0.278	0.667	0.333
ICT	75	0.048	0.144	0.612	0.388
Other	123	0.069	0.168	0.516	0.484

Notes: Sample restricted to firms with at least one patent (excluding 48 scraper-artifact patents) and at least one product at any HS precision. Baseline specification uses the first-listed IPC of each patent and a Lybbert-Zolas concordance weight threshold of 0.10. "Same HS6" is the share of the firm's patents that match a firm product at the full six-digit HS code; "Same HS4" is the cumulative share matching at the four-digit heading (subsumes same HS6); "Same HS2" is the cumulative share matching at the two-digit chapter (subsumes same HS4 and same HS6); "Unrelated" is the residual. Innovation categories are mutually exclusive: AI patents take precedence; ICT excludes AI; Other excludes both. *Preliminary*: the AI sample of 3 firms is expected to expand substantially once the EBOPS services classification is complete (currently 45 AI-flagged firms have all-services product portfolios pending EBOPS assignment).

AI innovators exhibit the highest degree of alignment, followed by ICT innovators and other firms. At the HS2 level, approximately 67% of AI patents align with firms' product portfolios, compared with 61% for ICT patents and 52% for other patents. However, exact HS6 matches are relatively uncommon across all innovation types. This reflects both the fine granularity of the HS6 classification and the fact that some products are only mapped at the HS4 or HS2 level, making exact six-digit matches impossible by construction. Consequently, alignment at the HS4 and HS2 levels provides a more informative characterization of technological proximity in our setting.

Robustness to specification. Table A10 presents results using the union specification, in which HS assignments are constructed from all IPC codes associated with a patent. Alignment rates increase modestly under this approach, particularly for ICT patents, reflecting the larger set of candidate HS codes. Importantly, the qualitative ranking across innovation types remains unchanged.

Robustness to patent count. Table A11 restricts the sample to firms with progressively larger patent portfolios. Alignment at the HS2 level increases from 55% among firms with at least one patent to 67% among firms with at least three patents, suggesting that firms with more extensive innovation activity tend to patent in technological domains more closely related to their observed product portfolios. Estimates for firms with five or more patents should be interpreted cautiously because of the limited sample size.

Table A10: Patent-Product Alignment by Innovation Type, Union Specification

Innovation Type	N firms	Same HS6	Same HS4	Same HS2	Unrelated
AI	48	0.000	0.278	0.667	0.333
ICT	77	0.096	0.238	0.759	0.241
Other	125	0.075	0.169	0.520	0.480

Notes: As in Table A9, but using the union specification in which all IPC codes assigned to a patent contribute to the patent's HS6 set. The ranking of AI > ICT > Other at the HS2 level is preserved. The ICT shares are notably higher under union than under primary, reflecting that ICT patents in our data carry more IPC codes on average than other categories. See Section X.Y for further discussion.

Table A11: Alignment by Minimum Patent Count

Min N patents	N firms	Same HS6	Same HS4	Same HS2
≥ 1	201	0.060	0.161	0.554
≥ 2	56	0.055	0.184	0.615
≥ 3	26	0.080	0.263	0.671
≥ 5	6	0.000	0.304	0.767
≥ 10	2	0.000	0.292	0.500

Notes: Each row restricts the firm sample to those with at least the indicated number of matched patents. Alignment shares increase modestly with the patent-count threshold up to $n_{\text{patents}} \geq 5$, then become noisy at very high thresholds due to small sample size. The pattern is consistent with the interpretation that more-innovative firms innovate more coherently within their commercial space.

End-Use Classification The HS-based alignment measure captures proximity in the product classification hierarchy but does not account for products' economic function. As a complementary robustness exercise, we classify products according to the Broad Economic Categories (BEC) nomenclature (United Nations Statistics Division, 2002), following the approach of Bergeaud et al. (2024). This allows us to distinguish patents that relate to a firm's exact products from those associated with products serving the same broad end use.

Using the UN HS2007-to-BEC concordance, we map each HS6 code to one of five broad end-use groups: Consumption, Intermediate Primary, Intermediate Processed, Capital, and Other. Patents inherit BEC groups via their mapped HS6 codes, whereas firms inherit them via their HS6-classified products.

For each patent, we define three mutually exclusive alignment categories:

- **Horizontal**: the patent shares an HS6 code with one of the firm's products.
- **Non-competing**: the patent does not share an HS6 code but belongs to the same BEC end-use category as one of the firm's products.
- **Vertical**: the patent belongs to a different BEC end-use category than the firm's products.

Results Table A12 reports the BEC-based alignment measures. The BEC-based classification broadly confirms the findings from the HS-hierarchy analysis. Most patents (60–67%) fall into the non-competing category, indicating that firms tend to innovate within the same broad end-

use space even when patents do not correspond to their exact products. Horizontal matches are relatively uncommon, while roughly one-third of patents are classified as vertical.

Table A12: Patent-Product Alignment by Innovation Type, BEC End-Use

Innovation Type	N firms	Horizontal	Non-competing	Vertical
AI	48	0.000	0.667	0.333
ICT	66	0.054	0.603	0.343
Other	108	0.079	0.630	0.292

Notes: Sample restricted to firms with at least one patent and at least one HS6-precise product with a BEC end-use mapping. Following **?**, patents are classified by alignment with firm products: "Horizontal" indicates patent HS6 matches a firm product HS6 (same product); "Non-competing" indicates patent and product share a BLW end-use group but differ in HS6 (same end-use, different product); "Vertical" indicates patent and product differ in BLW end-use group entirely. BLW groups are: Consumption (BEC 112, 122, 522, 61, 62, 63), Intermediate primary (BEC 111, 21, 31), Intermediate processed (BEC 121, 22, 321, 322), Capital (BEC 41, 42, 51, 521, 53), and Other (BEC 7). Source: CMIE Prowess + IPO + Lybbert and Zolas (2014) + UN HS-BEC concordance.

The ranking across innovation types is also consistent with the baseline results: AI patents exhibit the strongest within-end-use alignment, followed by ICT and other patents. Thus, using end-use categories rather than the HS hierarchy yields the same qualitative conclusion: AI-related innovation is more closely associated with firms' observed commercial activities.

Table A13 reports the cross-tabulation between patent and firm-product BEC groups. Most observations lie on the diagonal, particularly for Intermediate Processed and Capital goods, indicating that innovation predominantly occurs within the same broad end-use category. The largest off-diagonal cells involve firms producing intermediate goods while patenting capital goods or downstream processed products, suggesting innovation in complementary technologies along the production chain.

Table A13: Cross-Tabulation: Patent BLW $\times$ Product BLW

Patent BLW \ Product BLW	Consumption	Intermediate primary	Intermediate processed	Capital	Other
Consumption	10.5	23.4	19.0	14.2	0.8
Intermediate primary	2.6	4.8	1.4	2.8	0.0
Intermediate processed	14.2	24.5	30.8	15.3	0.1
Capital	20.2	36.5	23.9	36.2	2.0
Other	0.0	0.0	0.0	0.0	0.0

Notes: Each cell reports the weighted frequency of patent-product BLW end-use pairs. A patent with multiple BLW mappings is fractionally distributed across rows. Diagonal cells correspond to alignment within the same end-use category.

Overall, the BEC-based analysis provides a robustness check that complements the HS-hierarchy approach and reaches the same qualitative conclusion: firms' innovation activities are concentrated in technological domains closely related to their commercial activities.

C Additional Descriptive Evidence

This appendix provides supplementary descriptive evidence on the distribution of AI innovation across firms, sectors, locations, and time. The results complement the summary statistics presented in the main text and further illustrate the concentration of AI activity within the Indian economy.

Table A14: Descriptive Statistics by AI Innovator Status

	Non-AI Firms			AI Firms		
Variable	Mean	SD	N	Mean	SD	N
Net sales (INR mn)	867.21	4930.62	42525	5657.40	13721.71	47
Export intensity	0.39	0.64	42525	0.05	0.13	47
Labor productivity	3.24	20.72	668	0.47	0.57	11
Log data expenditure	0.21	0.53	42525	0.72	1.35	47
R&D intensity	0.06	0.40	42525	0.75	1.85	47
Firm age	29.80	16.81	42490	36.89	23.73	47
Exporter (=1)	0.40	0.49	42525	0.83	0.38	47
Observations		42,525			47	

Notes: This table reports firm-level descriptive statistics by AI innovator status. AI firms are defined as firms with at least one AI-related patent. Export intensity and R&D intensity are expressed as ratios. Labor productivity is measured as output per worker. The number of observations varies across variables due to missing data.

Table A15: Descriptive Statistics

	Mean	SD	p25	Median	p75	N
Net sales (INR mn)	872.50	4951.05	10.50	83.00	398.30	42572
Export intensity (exports/sales)	0.39	63.56	0.00	0.00	0.00	42572
Labor productivity (def. sales/employees)	3.20	20.56	0.06	0.24	0.73	679
Log data expenditure	0.21	0.53	0.00	0.02	0.16	42572
R&D intensity (log)	0.06	0.40	0.00	0.00	0.00	42572
Firm age (years)	29.81	16.82	18.00	28.00	37.00	42537
AI innovator (=1)	0.00	0.03	0.00	0.00	0.00	42572
ICT innovator (=1)	0.01	0.07	0.00	0.00	0.00	42572
Exporter (=1)	0.40	0.49	0.00	0.00	1.00	42572

Notes: Firm-level statistics (one observation per firm). Sample: 42572 unique firms, 30 years, 333738 firm-year observations. 47 AI-patenting firms (83.0% are exporters). 240 ICT-patenting firms, 4 any-patent firms. 17139 exporting firms (40.3% of sample). 57 2-digit and 270 4-digit NIC industries covered.

Table A16: Exports and Innovation in the Manufacturing Sector (Full Panel)

NIC	Description	Firms	Mean per Firm		Trade		% Innovator				Mean Patents	
			Wage Bill (Cr)	Sales (Cr)	% Exporter	Exports/Exp. (Cr)	Innov.	AI	ICT	Other	AI	ICT
10	Other semi-processed, processed or insta	1598	0.6	6.0	25.0	7.5	1.3	0.0	0.2	1.2	0.00	0.34
11	Manufacture of distilled, potable, alcoh	355	0.5	6.2	22.6	6.5	0.8	0.0	0.0	0.8	0.00	0.00
12	Manufacture of cigarettes, cigarette tob	33	0.6	7.4	31.9	13.1	14.8	0.0	0.0	14.8	0.00	0.00
13	Weaving, Manufacture of cotton and cotto	1312	0.6	6.0	40.9	6.3	0.9	0.0	0.1	0.9	0.00	0.36
14	Manufacture of wearing apparel, except f	327	0.5	6.0	47.0	9.0	1.8	0.0	0.2	1.8	0.00	0.10
15	Manufacture of other footwear n.e.c.	169	0.5	5.4	47.0	7.7	1.0	0.0	0.1	1.0	0.00	0.10
16	Manufacture of products of wood, cork, s	103	0.5	5.6	30.1	7.5	2.1	0.0	0.0	2.1	0.00	0.00
17	Manufacture of paper and paper rolls not	436	0.5	5.8	23.4	7.1	1.2	0.0	0.1	1.2	0.00	0.11
18	Other printing activities like screen pr	55	0.6	4.8	19.8	10.1	2.9	0.0	2.1	2.9	0.00	3.28
19	Production of liquid and gaseous fuels,	153	0.7	7.0	30.2	9.4	6.9	1.5	3.1	6.5	0.79	2.05
20	Manufacture of pesticides and other agro	1754	0.6	6.2	45.4	7.2	5.9	0.1	0.8	5.9	0.01	0.23
21	Manufacture of other pharmaceutical and	918	0.6	6.0	49.1	10.0	18.3	0.0	2.4	18.1	0.00	0.31
22	Manufacture of other plastics products n	993	0.6	5.7	41.7	7.4	4.1	0.1	0.5	3.6	0.02	0.35
23	Manufacture of clinkers and cement	562	0.6	6.2	31.4	7.1	3.0	0.0	0.5	3.0	0.00	0.16
24	Manufacture of other iron and steel cast	1735	0.5	6.5	29.6	7.5	2.3	0.1	0.3	2.3	0.09	0.39
25	Manufacture of steam or other vapour gen	595	0.5	5.9	37.1	8.2	1.7	0.3	0.5	1.7	0.39	6.85
26	Manufacture of industrial process contro	726	0.9	5.6	43.5	7.3	7.5	0.4	3.3	6.6	0.10	3.71
27	Manufacture of electric power distributi	858	0.8	6.0	40.3	7.1	5.9	0.0	1.8	5.3	0.00	0.70
28	Manufacture of other special-purpose mac	1153	0.7	6.0	52.6	7.3	8.0	0.0	1.3	7.9	0.01	0.23
29	Manufacture of diverse parts and accesso	913	0.7	6.7	44.0	8.5	8.4	0.6	3.0	7.7	0.18	2.48
30	Building of ships and boats	131	0.7	6.9	42.5	8.4	11.1	0.6	5.9	10.1	0.05	2.01
31	Other manufacturing	1804	0.4	5.9	16.6	7.3	1.2	0.2	0.4	1.2	0.13	0.52
	All Manufacturing	**16683**	**0.6**	**6.1**	**37.3**	**7.7**	**4.7**	**0.1**	**1.0**	**4.5**	**0.05**	**0.90**

Notes: Yearly firm averages over 1988–2024. Wage Bill, Sales, and Exports are measured in INR crores (1 Cr = 10 million INR) and are deflated. Sector codes correspond to the NIC 2008 2-digit classification. % Innov. denotes the share of firms with at least one patent. % AI, % ICT, and % Other denote the share of firms with at least one patent in each category; these categories are not mutually exclusive. Mean AI and Mean ICT report the average cumulative patent stock per firm-year. Source: CMIE Prowess and authors' calculations.

Table A17: Exports and Innovation in the Manufacturing Sector (Pre-GDPR)

NIC	Description	Firms	Mean per Firm		Trade		% Innovator				Mean Patents	
			Wage Bill (Cr)	Sales (Cr)	% Exporter	Exports/Exp. (Cr)	Innov.	AI	ICT	Other	AI	ICT
10	Other semi-processed, processed or insta	1059	0.6	6.7	25.7	10.6	1.0	0.0	0.2	0.9	0.00	0.32
11	Manufacture of distilled, potable, alcoh	227	0.5	6.7	23.4	9.0	1.1	0.0	0.0	1.1	0.00	0.00
12	Manufacture of cigarettes, cigarette tob	26	0.6	7.5	31.8	14.9	10.9	0.0	0.0	10.9	0.00	0.00
13	Weaving, manufacturing of man-made fiber	875	0.6	6.6	40.7	7.8	0.8	0.0	0.1	0.8	0.00	0.21
14	Manufacture of wearing apparel, except f	211	0.5	6.7	49.2	9.4	1.7	0.0	0.1	1.7	0.00	0.07
15	Manufacture of other footwear n.e.c.	103	0.5	6.2	51.1	8.1	1.4	0.0	0.0	1.4	0.00	0.00
16	Manufacture of products of wood, cork, s	66	0.5	5.9	28.8	.	1.8	0.0	0.0	1.8	0.00	0.00
17	Manufacture of paper and paper rolls not	288	0.5	6.4	25.2	11.9	1.6	0.0	0.1	1.6	0.00	0.05
18	Printing of magazines and other periodic	42	0.6	5.4	26.7	11.1	3.1	0.0	2.5	3.1	0.00	3.40
19	Production of liquid and gaseous fuels,	102	0.7	7.6	35.2	11.1	7.6	1.1	2.8	6.9	0.48	1.09
20	Manufacture of synthetic or artificial f	1178	0.6	6.8	50.3	9.2	6.1	0.1	0.6	6.1	0.01	0.17
21	Manufacture of pharmaceuticals, medicina	625	0.6	6.7	53.8	12.1	19.4	0.0	2.1	19.2	0.00	0.24
22	Manufacture of other plastics products n	728	0.6	6.3	46.3	9.0	3.7	0.2	0.2	3.4	0.05	0.08
23	Manufacture of clinkers and cement	378	0.7	6.8	30.7	9.1	3.4	0.0	0.6	3.4	0.00	0.18
24	Manufacture of other basic iron and stee	1242	0.5	7.0	30.2	9.4	2.1	0.1	0.2	2.1	0.06	0.26
25	Forging, pressing, stamping and roll-for	469	0.5	6.2	37.8	9.6	1.4	0.2	0.3	1.4	0.36	5.57
26	Manufacture of industrial process contro	483	1.0	5.9	45.6	8.8	8.5	0.6	3.2	7.7	0.12	2.94
27	Manufacture of other domestic appliances	643	0.8	6.4	43.7	8.7	6.0	0.0	2.0	5.5	0.00	0.81
28	Manufacture of other special-purpose mac	910	0.7	6.2	55.3	8.7	7.8	0.1	1.3	7.8	0.01	0.26
29	Manufacture of parts and accessories for	744	0.7	7.0	48.7	9.8	7.3	0.4	2.4	6.6	0.11	1.72
30	Building of ships and boats	100	0.7	7.1	45.8	9.6	11.2	0.2	6.2	10.2	0.02	1.49
31	Manufacture of stationary articles such	875	0.5	6.6	18.6	9.2	1.1	0.1	0.2	1.1	0.11	0.40
	All Manufacturing	**11374**	**0.6**	**6.6**	**39.8**	**9.5**	**4.7**	**0.1**	**0.9**	**4.6**	**0.04**	**0.73**

Notes: Yearly firm averages over 2013–2017. Wage Bill, Sales, and Exports are measured in INR crores (1 Cr = 10 million INR) and are deflated. Sector codes correspond to the NIC 2008 2-digit classification. % Innov. denotes the share of firms with at least one patent. % AI, % ICT, and % Other denote the share of firms with at least one patent in each category; these categories are not mutually exclusive. Mean AI and Mean ICT report the average cumulative patent stock per firm-year. Source: CMIE Prowess and authors' calculations.

C.1 Evidence on Stylized Facts

Figure 10 compares the distribution of total factor productivity among AI-patenting exporters and AI-patenting firms serving only the domestic market. Exporters exhibit systematically higher productivity, suggesting positive selection into foreign markets even within the subset of AI innovators.

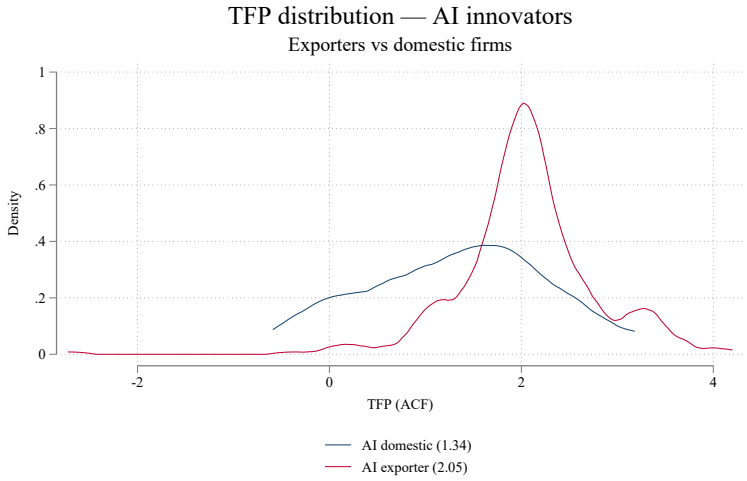
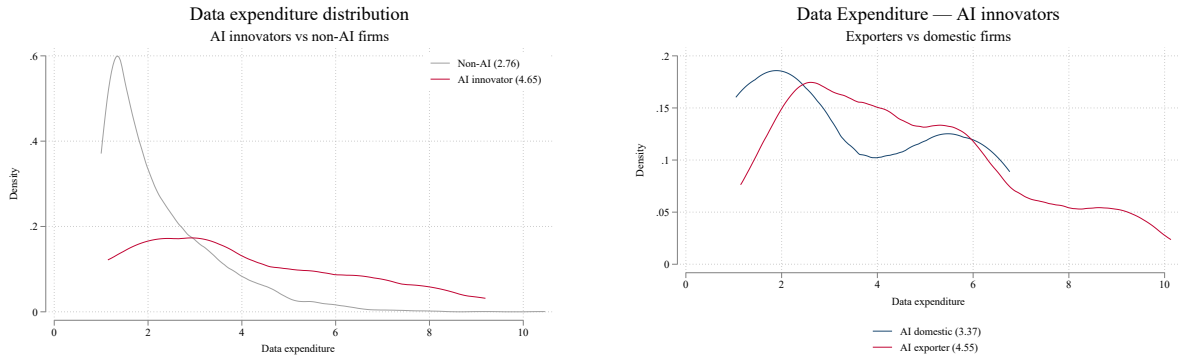


Figure 10: TFP distribution among AI innovators: exporters versus domestic firms



(a) AI vs. non-AI firms

(b) AI exporters vs. AI domestic firms

Figure 11: Distribution of data-related expenditure

Notes: The figure reports kernel density estimates of log data expenditure. The left panel compares AI innovators with non-AI firms, while the right panel compares exporting and non-exporting firms within the set of AI innovators. Data expenditure is constructed as the sum of communications services, web hosting, software development, internet services, satellite links, and IT-enabled services.

C.1.1 Concentration of AI innovation

Figure 12 plots the distribution of value added across five firm groups: domestic non-exporters, exporters, ICT innovators, ICT-and-AI innovators, and AI innovators. The distributions shift monotonically to the right as we move up the innovation ladder, indicating that more innovative firms tend to be larger and more productive. This pattern is consistent with a setting

in which the ability to collect, store, and process data is unevenly distributed across firms. In particular, firms with greater scale, organisational capacity, and technological infrastructure are better positioned to transform data into innovation outputs. [8]

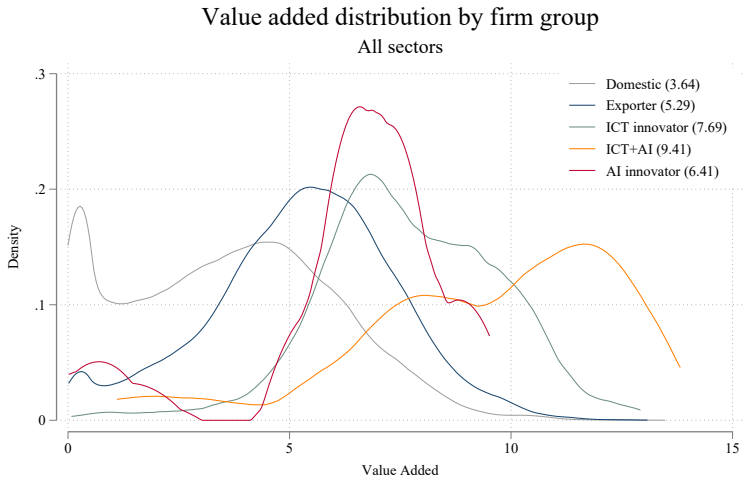


Figure 12: Value-added distribution by firm group

We next examine how innovation is distributed across sectors, locations, and time.

Table A18: Sectoral Distribution of Firms by AI Innovation Status

	AI status		
Sector	Non-AI firms	AI firms	Total
Agriculture	806	0	806
Manufacturing	16,668	15	16,683
Mining	317	0	317
Services	37,190	32	37,222
Total	54,981	47	55,028

Notes: This table reports the number of unique firms by sector and AI innovation status. AI firms are defined as firms with at least one AI patent. Non-AI firms are those with no AI patents.

Sectoral concentration. AI innovation is overwhelmingly concentrated in service sectors, which account for the majority of AI firms, while manufacturing plays a more limited role and other sectors are largely absent (Table A18). This pattern is consistent with the idea that AI innovation is closely tied to data-intensive activities, which are more prevalent in services than in traditional goods-producing sectors.

Geographic concentration. AI activity is also highly concentrated geographically. A small number of states—particularly Maharashtra, Delhi, Karnataka, and Tamil Nadu—account for the majority of AI firms (Table A19). Moreover, some regions, such as Karnataka, host a disproportionately large share of AI firms relative to their overall firm base, suggesting the presence

[8] At the same time, the sorting is not perfect. There is substantial overlap across distributions, suggesting that while these characteristics increase the likelihood of engaging in AI innovation, they are neither necessary nor sufficient conditions

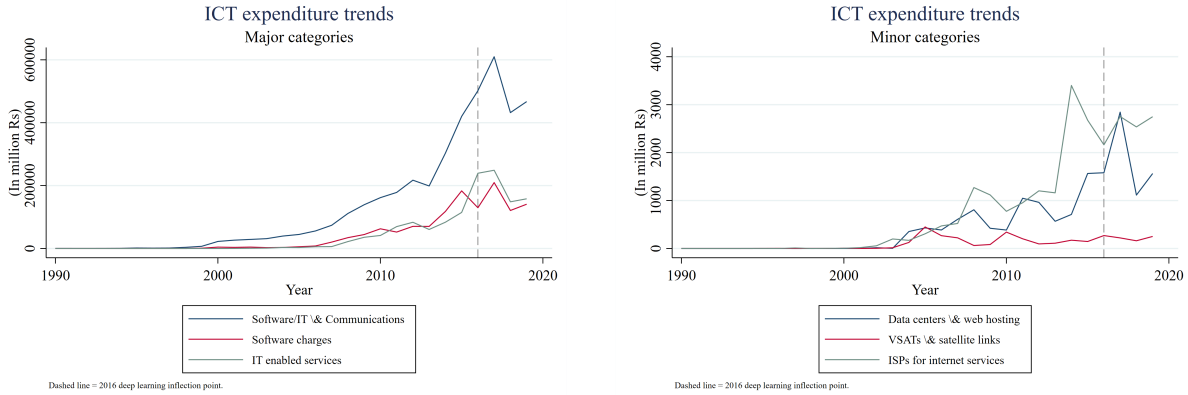
of localized technology clusters. This spatial concentration is consistent with agglomeration forces and knowledge spillovers playing an important role in AI innovation.

Table A19: Geographic Distribution of Firms by AI Status (Top 10 States)

	Firm Type		
Location	Non-AI Firms	AI Firms	Total
Maharashtra	12446	21	12467
NCT of Delhi	6853	6	6859
Tamil Nadu	3831	5	3836
Gujarat	3738	2	3740
West Bengal	3591	1	3592
Telangana	2243	1	2244
Karnataka	2228	9	2237
Haryana	1162	1	1163
Uttar Pradesh	1088	0	1088
Rajasthan	752	0	752
Other States	6472	1	6473
Total	**42404**	**47**	**42451**

Notes: This table reports the number of unique firms by location and AI status. States are ranked by total number of firms, and only the top 10 are shown individually. The remaining states are grouped into "Other States". AI firms are defined as firms with at least one AI-related patent.

Time trends. Finally, we examine the evolution of data-related investment over time. Figure 13a shows that key ICT expenditure components—particularly software development and IT-enabled services—increase sharply after 2016. In contrast, Figure 13b shows more gradual growth in other ICT components. This divergence suggests that certain types of digital investment, particularly those directly related to data processing and software capabilities, have become increasingly important in recent years.



(a) Main ICT components

(b) Secondary ICT components

Figure 13: Evolution of ICT expenditure components

Patent-Product Alignment To examine whether firms innovate in technological domains related to their commercial activities, we construct a patent-product alignment measure that maps both patents and products into the Harmonized System (HS) classification. We link

patents to HS6 products using the IPC-to-HS concordance of Lybbert and Zolas (2014), while mapping firms' products to HS codes using our LLM-assisted CMIE-to-HS crosswalk. We then compare each patent's associated HS codes with the firm's observed product portfolio and classify the relationship into four mutually exclusive categories: (i) an exact HS6 match, (ii) a match at the HS4 level, (iii) a match at the HS2 level, or (iv) no match. Because products are not always classified at the six-digit level, evaluating alignment at broader HS aggregations provides a complementary measure of technological proximity while preserving coverage. Additional details on the construction of the measure are provided in Appendix B.3.

Table A9 reports the results by innovation type. AI innovators exhibit the highest degree of patent-product alignment, followed by ICT innovators and other firms. Approximately 67% of AI patents align with firms' product portfolios at least at the HS2 level, compared with 61% for ICT patents and 52% for other patents. A similar ordering emerges at the HS4 level, suggesting that AI-related innovation is particularly concentrated in technological domains adjacent to firms' existing commercial activities.

Exact HS6 matches are relatively uncommon across all groups, reflecting both the granularity of the HS classification and the varying precision of the product crosswalk. More broadly, the evidence indicates that firms' patenting activity is not randomly distributed across technologies but is systematically concentrated in areas related to their observed product portfolios. This pattern is consistent with the idea that firms innovate in domains where they possess product-specific expertise and accumulate commercially relevant information, providing suggestive evidence for a data-accumulation channel linking production and innovation.

C.1.2 Consequences of AI innovation

We descriptively document how firm behaviour evolves around the first instance of AI innovation. The following event-study specification traces conditional correlations between firm outcomes and proximity to first AI patenting; we return to causal identification in section 4.

$$Y_{it} = \alpha_i + \lambda_{st} + \sum_{\tau \neq -1} \beta_\tau \, 1\{t - T_i = \tau\} + \varepsilon_{it}, \tag{8}$$

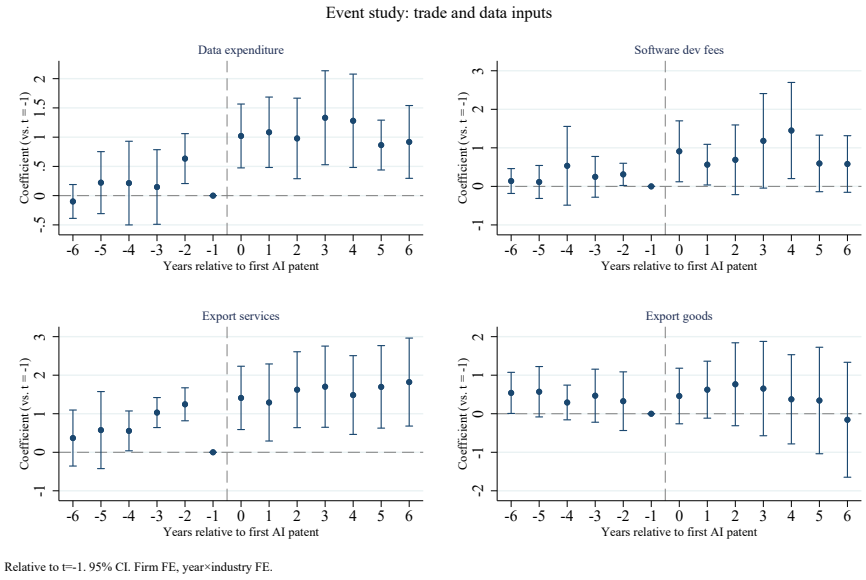
where Y_{it} is the outcome of interest for firm i in year t, α_i are firm fixed effects, and λ_{st} are year-by-4-digit-industry fixed effects absorbing shocks common to all firms in the same narrowly defined industry in a given year. The variable T_i denotes the year of firm i's first AI patent, and $\tau = -1$ is the omitted reference period. Standard errors are clustered at the 2-digit industry-by-location level. The pre-event coefficients ($\tau < 0$) capture differential trends prior to AI innovation; the post-event coefficients ($\tau \geq 0$) describe how outcomes evolve around the patent event.

Two features of the institutional context bear on interpretation. First, the timing of first AI patenting is endogenous to the firm's investment trajectory: firms that eventually innovate in AI have been accumulating relevant capabilities for years before filing — the patent event reflects the culmination of a gradual process, not its onset. Second, in India, patent applica-

tions are granted only after several years of examination, so the observed event time lags the underlying innovation activity by a substantial margin. Together, these features imply that the estimated coefficients should be interpreted as descriptive correlations rather than causal effects of AI innovation.

Trade and data investment. Figure 14 provides core evidence for the data accumulation mechanism. Data expenditure, R&D investment, and service exports begin to rise several periods before the patent event and continue increasing afterwards, tracing a gradual build-up of the inputs required for AI innovation. This pre-event expansion suggests that firms progressively scale up their data infrastructure and deepen foreign service relationships in the years leading up to innovation, with the patent filing marking the point at which accumulated inputs become sufficient for model development. This dynamic pattern is consistent with the data accumulation channel: sustained export activity contributes to the gradual growth of the data stock D_{jt}, which in turn shapes firms' investment trajectories around the innovation event. By contrast, we find no comparable pre- or post-event dynamics for goods exports. This asymmetry indicates that the observed patterns are not driven by general firm expansion, but rather by service-based interactions that generate and leverage data.

Figure 14: Event-study: Trade and Data Expenditures Around AI Innovation



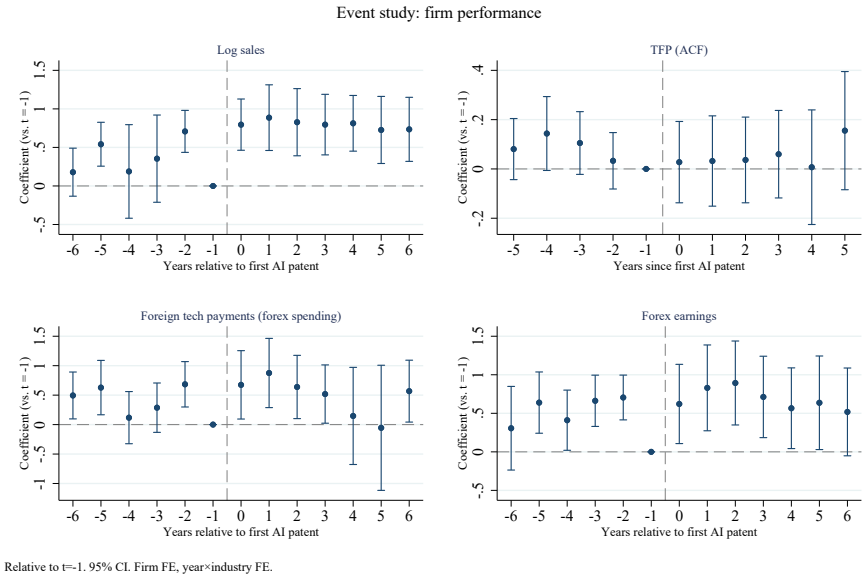
Notes: This figure plots event-study coefficients around the first AI patent at the firm level ($\tau = 0$). The dependent variables capture data-related investments and exports on goods and services. All specifications include firm fixed effects and year×4-digit industry fixed effects. The omitted category is $\tau = -1$. Standard errors are clustered at the 2-digit industry-by-location level, and 95% confidence intervals are shown. The sample includes firms observed in the event window around their first AI patent.

Firm responses to AI innovation. Firms systematically adjust their production structure following AI innovation. In particular, they reallocate toward more data- and skill-intensive activities: skill intensity rises, capital intensity declines, and data expenditure increases substantially, consistent with a shift away from traditional capital-intensive production toward data-driven operations. These patterns suggest that AI complements high-skill labour rather than substituting for it. At the same time, firms expand in scale and market engagement, expe-

riencing higher sales and export earnings alongside increased spending on foreign technology. By contrast, we find limited evidence of immediate gains in measured productivity, consistent with innovation-to-commercialisation lags: productivity improvements from AI may materialise over longer horizons or operate primarily through scale expansion and organisational change rather than short-run efficiency gains captured in aggregate TFP measures. Additional evidence on input reallocation, innovation activity, and firm performance is presented in Appendix C.1.2.

We document two additional dimensions: broader innovation activity (Figure 16) and detailed input substitution (Figure 17). All specifications follow equation (8).

Figure 15: Event-study: Firm Performance Around AI Innovation



Notes: This figure plots event-study coefficients around the first AI patent at the firm level ($\tau = 0$). The dependent variables capture firm performance outcomes, including sales, export earnings, foreign technology payments, and productivity. All specifications include firm fixed effects and year×4-digit industry fixed effects. The omitted category is $\tau = -1$. Standard errors are clustered at the 2-digit industry-by-location level, and 95% confidence intervals are shown.

C.2 Descriptive Panel Regressions

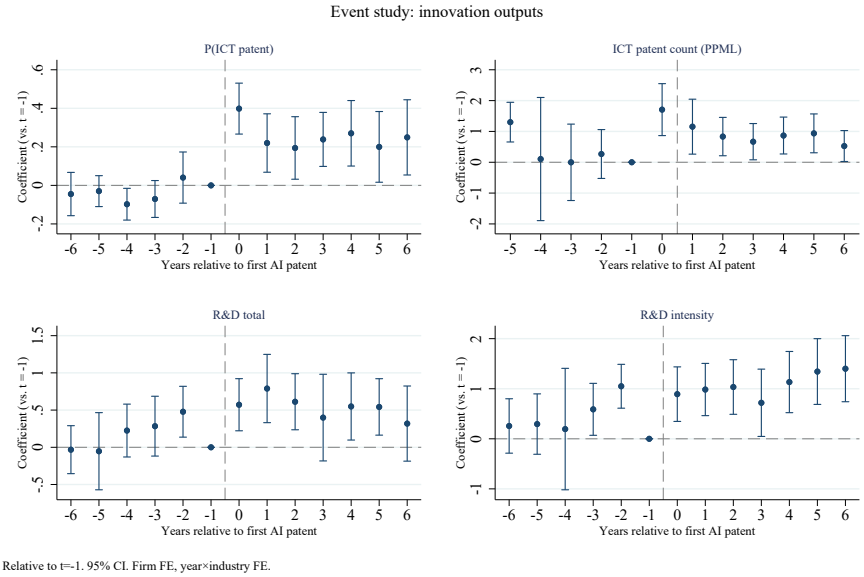
As a complement to the descriptive statistics presented in the main text, we estimate panel regressions relating firms' trade activity to AI innovation. These specifications are intended to document conditional correlations rather than identify causal effects, as firms self-select into export markets and time-varying unobservables may jointly influence exporting and innovation (Melitz, 2003).

The baseline specification is

$$\mathrm{AI}_{it} = \alpha_i + \alpha_{kt} + \beta\,\mathrm{Trade}_{it} + \mathbf{X}_{i0}\delta_t + \varepsilon_{it}, \tag{9}$$

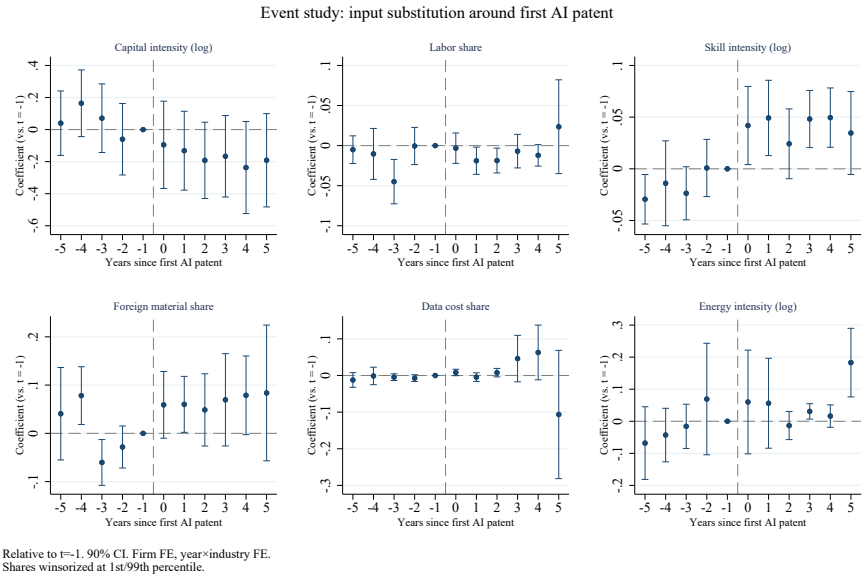
where α_i denotes firm fixed effects, α_{kt} are four-digit industry-by-year-by-location fixed

Figure 16: Event-study: Innovation Activity Around AI Innovation



Notes: This figure plots event-study coefficients for innovation outcomes around the first AI patent. The dependent variables include ICT patenting and R&D expenditure (levels and intensity). All specifications include firm fixed effects and year×4-digit industry fixed effects. The omitted category is $\tau = -1$. Standard errors are clustered at the 2-digit industry-by-location level, and 95% confidence intervals are shown.

Figure 17: Event-study: Inputs Around AI Innovation



Notes: This figure plots event-study coefficients for inputs around the first AI patent. The dependent variables include capital intensity, labor share, skill intensity, foreign material share, and energy intensity. All specifications include firm fixed effects and year×4-digit industry fixed effects. The omitted category is $\tau = -1$. Standard errors are clustered at the 2-digit industry-by-location level, and 95% confidence intervals are shown.

effects, and $\mathbf{X}_{i0}\delta_t$ represents baseline firm characteristics interacted with year indicators. Standard errors are clustered at the two-digit industry-by-location level.

Table A20 reports estimates of the relationship between trade exposure and AI innovation.

Table A20: Effect of Trade on AI Innovation

	AI innovation			AI innovation (patents)		
	(1)	(2)	(3)	(4)	(5)	(6)
Export (services)	0.004*	0.005*	0.006**	0.010*	0.018**	0.019**
	(0.002)	(0.003)	(0.003)	(0.006)	(0.008)	(0.009)
Export (goods)	-0.002	-0.004*	-0.004	-0.013	-0.039	-0.043
	(0.001)	(0.002)	(0.003)	(0.012)	(0.034)	(0.038)
Imports	0.003*	0.005	0.005	0.008*	0.019*	0.019*
	(0.002)	(0.003)	(0.004)	(0.005)	(0.011)	(0.011)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry×Year FE	No	No	Yes	No	No	Yes
Pre-char.×Year	No	Yes	Yes	No	Yes	Yes
Observations	9082	2657	2447	9082	2657	2447
R^2	0.008	0.079	0.086	0.010	0.068	0.079

Notes: Dependent variable: AI patent dummy (cols 1–3), AI total patents (cols 4–6). Col (1)/(4): firm + year FE, clustered by firm. Col (2)/(5): adds pre-period characteristics×year. Col (3)/(6): firm + year×4d industry FE, clustered by industry×location. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

We find that service exports are positively and significantly associated with AI innovation across all specifications. The estimates are stable to the inclusion of different controls and fixed effects. The magnitude is larger when using AI patent counts (cols. 4–6), indicating that service exports are associated not only with a higher likelihood of innovating, but also with greater innovation intensity. In contrast, goods exports show no consistent positive relationship with AI innovation, with coefficients that are small, negative, and often insignificant. This contrast suggests that the trade-innovation link is not driven solely by market size, but is stronger in activities that involve richer client interactions and data generation. Importantly, this pattern is specific to AI innovation. We find no corresponding effect on ICT or other patenting activity, ruling out a general innovation or scale-based explanation.

We next examine the temporal structure of this relationship. Table A21 shows that the effect of service exports on AI innovation is not purely contemporaneous, but emerges with a lag. Both one- and two-period lags of service exports are positively associated with AI innovation, with magnitudes that remain stable or increase across specifications. This lagged pattern is consistent with a mechanism in which exporting gradually generates usable data that feeds into AI development, rather than affecting innovation instantaneously through market-size effects alone. By contrast, we find no comparable lag structure for ICT innovation, reinforcing the idea that the dynamics are specific to AI.

These findings motivate the causal analysis based on firm-level foreign demand shocks presented in the main text.

Table A21: Lag Structure: Exports and Innovation

	AI innovation		ICT innovation	
	(1)	(2)	(3)	(4)
Export (services) t-1	0.006**	0.010**	0.004	0.016*
	(0.003)	(0.005)	(0.003)	(0.009)
Export (services) t-2	0.002	0.011*	0.003	0.018
	(0.003)	(0.006)	(0.004)	(0.017)
Imports t-1	0.002	0.010	0.007	0.034
	(0.002)	(0.010)	(0.005)	(0.022)
Imports t-2	0.004	0.002	0.001	0.015
	(0.003)	(0.009)	(0.004)	(0.020)
Firm FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Industry×Year FE	Yes	Yes	Yes	Yes
Pre-char.×Year	Yes	Yes	Yes	Yes
Observations	1929	1918	1929	1918
R^2	0.104	0.056	0.145	0.101

Notes: Cols (1)–(2): AI patent dummy and count. Cols (3)–(4): ICT patent dummy and count. All specifications include firm FE, year×4d industry FE, and pre-period characteristics×year. Clustered by industry×location. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

D Identification Strategy

This appendix documents the validity of the shift-share design used to construct our foreign demand instrument. As discussed in Section 4.1, the Borusyak et al. (2022) framework rests on two pillars: shock exogeneity (quasi-random assignment of product-level shocks conditional on the controls and fixed effects we include) and a sufficiently rich shock structure. The first is a substantive identification assumption that we discuss in the main text; the second is a finite-sample concern, namely whether the design draws on enough independent shock variation for inference to be reliable. The remainder of this appendix addresses the second concern through four diagnostics.

D.1 Number of Underlying Shocks

A second concern in shift-share designs is whether the data contain enough shock-level variation for asymptotic inference to be meaningful. Borusyak et al. (2022) formalize this as a condition on the number of effective shocks underlying the instrument. Our foreign demand series exhibits substantial variation along this dimension: 91,151 HS6-by-year growth rates ΔFD_{pt} are computed from 18 years of BACI bilateral trade data across 5,479 distinct six-digit product codes, following the construction in Barrows and Ollivier (2021). The number of underlying shocks is therefore an order of magnitude larger than the firm-year observation count in our estimation sample.

Beyond raw counts, the more relevant diagnostic is the effective sample size implied by the

Table A22: Shift-Share Instrument Diagnostics

Panel A: Shock-level statistics (HS6×year)	
Number of distinct HS6 products	5,479
Number of years	18
Number of HS6-by-year shocks	91,151
Mean shock	0.0342
Standard deviation of shocks	0.3332
P10 – P90 range of shocks	[-0.299, 0.365]
Interdecile range (P90−P10)	0.663
Panel B: Effective sample size	
Distinct HS6 products contributing	2,068
HHI ($\sum_p \bar{s}_p^2$)	0.00635
Effective number of shocks (1/HHI)	157.6
Panel C: Concentration of exposure	
Top-5 HS6 share of total exposure	0.145
Top-10 HS6 share of total exposure	0.175
Top-20 HS6 share of total exposure	0.222
Panel D: Firm-level exposure variation (Bartik HS6)	
Firms with non-missing exposure	3,745
Cross-firm SD of average exposure	0.045700
Cross-firm IQR of average exposure	0.054527

Notes: Panel A reports the number and dispersion of product-level foreign demand shocks ΔZ_{pt} used in the construction of the Bartik instrument. Shocks are computed at the HS6-by-year level using bilateral trade flows from BACI (excluding India) and the Davis-Haltiwanger growth rate following Barrows and Ollivier (2021). Panel B reports the effective sample size implied by the inverse Herfindahl of average exposure weights, $N_{\text{eff}} = 1/\sum_p \bar{s}_p^2$, where $\bar{s}_p = N^{-1}\sum_i w_{ip}$ is the average exposure weight across all firms, following **?**. Values below 100 would indicate that a small set of products drives the shift-share variation. Panel C reports the cumulative share of average exposure accounted for by the most heavily weighted HS6 products. Panel D documents heterogeneity in average Bartik exposure across firms.

inverse Herfindahl of average firm exposure weights:

$$N_{\text{eff}} = \frac{1}{\sum_p \bar{s}_p^2}, \qquad \text{where} \qquad \bar{s}_p = \frac{1}{N}\sum_{i=1}^{N} w_{ip}, \tag{10}$$

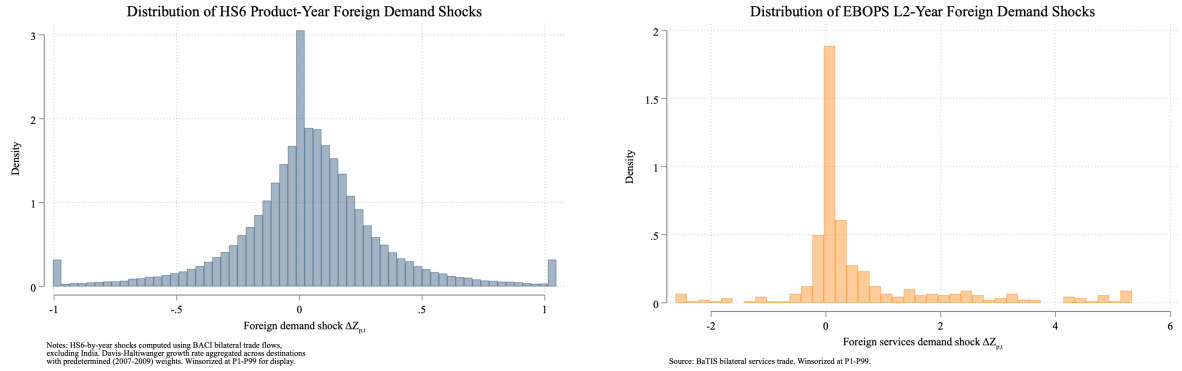
and the sum in the denominator runs over all HS6 products p for which $\bar{s}_p > 0$, with $\bar{s}_p$ averaging the firm-level exposure weight w_{ip} over the full sample of N firms (including firms with $w_{ip} = 0$). The denominator $\sum_p \bar{s}_p^2$ is the Herfindahl index of the distribution of average exposure across products; the inverse measures the number of equally-weighted shocks that would generate the same dispersion.

In our setting, $N_{\text{eff}} = 157.6$ (or 435, depending on the shock), comfortably above the threshold of 100 that Borusyak et al. (2022) suggest for reliable inference. The cross-product distribution of average exposure is similarly diffuse: the top 5 HS6 products jointly account for only 14.5% of total exposure, the top 10 for 17.5%, and the top 20 for 22.2%. These patterns suggest that no narrow set of products underlies the identifying variation. Table A22 reports the full set of diagnostics.

D.2 Dispersion of Foreign Demand Shocks

The shift-share design requires that the shocks themselves carry meaningful variation. The distribution of ΔFD_{pt} across HS6-by-year cells is shown in Figure 18: shocks are centered near zero with substantial dispersion. Figure 18b reports the distribution of the EBOPS-level foreign demand shocks used to construct the firm-level service demand measure. The distribution is concentrated around zero, indicating that most service categories experience relatively modest year-to-year changes in foreign demand. At the same time, the distribution exhibits a pronounced right tail, reflecting episodes of exceptionally strong demand growth in a limited number of internationally traded service categories. This pattern is consistent with the concentrated nature of global services trade, where rapidly expanding sectors—such as business and digital services—can experience substantially larger growth rates than those typically observed for merchandise trade.

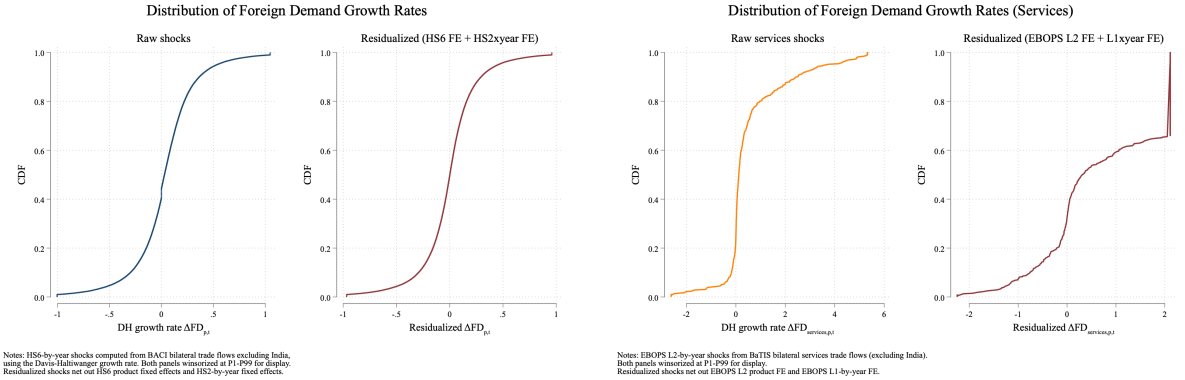
To verify that this variation is not driven entirely by stable cross-product differences or broad sectoral trends, we follow Barrows and Ollivier (2021) and residualize the shocks against HS6 product fixed effects (absorbing time-invariant product-level means) and HS2-by-year fixed effects (absorbing two-digit-sector trends). Figures 19a and 19b report the distributions of the product-level foreign demand shocks before and after residualizing them with respect to product and broad sector-by-year fixed effects. In both goods and services, the shocks exhibit substantial variation around zero and retain considerable dispersion after residualization, indicating that the identifying variation is not driven solely by common product or sectoral trends. While service-sector shocks display greater skewness and a heavier upper tail than goods shocks, reflecting the more concentrated nature of international services trade, both measures provide meaningful within-product variation for identification.



(a) Goods (HS6) (b) Services (EBOPS)

Figure 18: Distribution of Product-Level Foreign Demand Shocks

Notes: The left panel reports the distribution of annual foreign demand shocks for HS6 goods, constructed from CEPII BACI bilateral trade flows excluding India using the Davis–Haltiwanger growth rate and predetermined destination weights. The right panel reports the corresponding distribution for EBOPS service categories, constructed from OECD–WTO BaTIS bilateral trade flows using the same methodology. For visualization, both distributions are winsorized at the 1st and 99th percentiles; all empirical results use the non-winsorized shocks.



(a) Goods (HS6) (b) Services (EBOPS L2)

Figure 19: Distribution of Product-Level Foreign Demand Shocks

Notes: Each panel reports the cumulative distribution function (CDF) of the product-level foreign demand shocks before (left) and after (right) residualizing the shocks with respect to product fixed effects and broad sector-by-year fixed effects. Goods shocks are constructed from CEPII BACI bilateral trade flows at the HS6 level, while service shocks are constructed from OECD–WTO BaTIS bilateral trade flows at the EBOPS Level 2 classification. In both cases, India is excluded from the calculation of destination-specific demand growth, which is measured using the Davis–Haltiwanger growth rate. For visualization, the shocks are winsorized at the 1st and 99th percentiles. Residualized shocks retain substantial variation, indicating that the identifying variation is not driven solely by common product or broad sector-year trends.

D.3 Cross-Firm Exposure Variation

Finally, the identifying variation in our shift-share regressor comes not only from the underlying shocks but from differences in firm exposure to those shocks. We document the cross-firm dispersion of average Bartik exposure in Panel D of Table A22. Across 3,745 firms with non-missing exposure in the HS6 specification, the cross-firm standard deviation of average exposure equals 0.046 (IQR 0.055). Combined with the within-product temporal variation documented above, this cross-firm heterogeneity in product portfolios provides the variation that identifies β in Equation (5).

E The Combined Foreign Demand Measure

This appendix provides additional details on the construction of the combined foreign demand measure used throughout the paper. We describe the underlying data sources, the construction of firm-level exposure, and a series of diagnostic exercises evaluating the properties of the resulting shift–share measure. We also report robustness checks using alternative service classifications and separate goods and services measures.

Our preferred specification combines foreign demand shocks for both goods and services using firms' pre-period product mix. This construction allows firms' overall exposure to foreign demand to reflect the composition of their export portfolio rather than treating goods and services separately. This accounts for the fact that many Indian firms simultaneously export manufactured goods and services, particularly in technology-intensive industries such as software, information technology, engineering, and professional services. Restricting the analysis to either margin alone would therefore provide an incomplete measure of firms' exposure to international demand.

To do so, we combine two complementary measures of destination-specific demand growth, the Davis–Haltiwanger growth rates, and we aggregate them using predetermined destination shares. The firm-level product weights are constructed from CMIE Prowess using average product revenue shares over the pre-period (2002–2004). Products are classified into HS6 or EBOPS codes using the crosswalk described in Appendix B.2. We restrict our attention to classified products; weights are normalized as

$$w_{ik}^{combined} = \frac{\tilde{w}_{ik}^{pre}}{\sum_{k' \in K_i^{class}} \tilde{w}_{ik'}^{pre}},$$

where K_i^{class} denotes all classified goods and service products produced by firm i.

The combined foreign demand measure is

$$Z_{it}^{combined} = \sum_{k \in K_i^{class}} w_{ik}^{combined} \Delta Z_{kt},$$

where ΔZ_{kt} denotes the product-level foreign demand shock.

The measure can be decomposed into

$$Z_{it}^{combined} = z_{it}^{goods} + z_{it}^{services},$$

allowing us to quantify the contribution of each export margin.

Following Borusyak et al. (2022); Barrows and Ollivier (2021), Table A23 reports several diagnostic statistics for the combined shift–share measure. First, the effective number of shocks equals 435.4, well above commonly used thresholds in the shift–share literature. Second, exposure is dispersed across a large number of products. The twenty most heavily weighted products account for only 13.7% of aggregate exposure, indicating that identification is not driven by a small number of industries. Third, Figure 20 plots the distribution of raw and residualized product-level shocks after removing broad sector-by-year effects. Substantial within-sector variation remains, indicating that the identifying variation is not driven solely by aggregate sectoral trends.

Table A23: Shift-Share Instrument Diagnostics: Combined (Goods + Services)

Panel A: Pooled shock statistics (HS6 + EBOPS L2)	
Mean shock	0.0368
SD of shocks	0.3576
P10 – P90 range	[-0.298, 0.367]
Panel B: Effective sample size (combined)	
Total products contributing	2,081
HHI	0.00230
Effective N (1/HHI)	435.4
Panel C: Concentration	
Top-5 share	0.083
Top-10 share	0.105
Top-20 share	0.137

Notes: Combined Bartik treats goods (HS6) and services (EBOPS L2) as one product universe. Weights jk^{pre} are normalized to sum to 1 across all classified products (goods + services) per firm. Effective sample size = $1/where$
bars$_p = N^{-1}$
$sum_i w_{ip}$ (BHJ 2022).

F Appendix to Foreign Demand Results

F.1 Goods and Services Components of the Export Demand Shock

Our preferred specification uses a single firm-level export demand shock that combines changes in foreign demand for both goods and services according to firms' predetermined product portfolios. This reflects the fact that many firms, particularly in technology-intensive industries, export both goods and services and are therefore simultaneously exposed to both sources

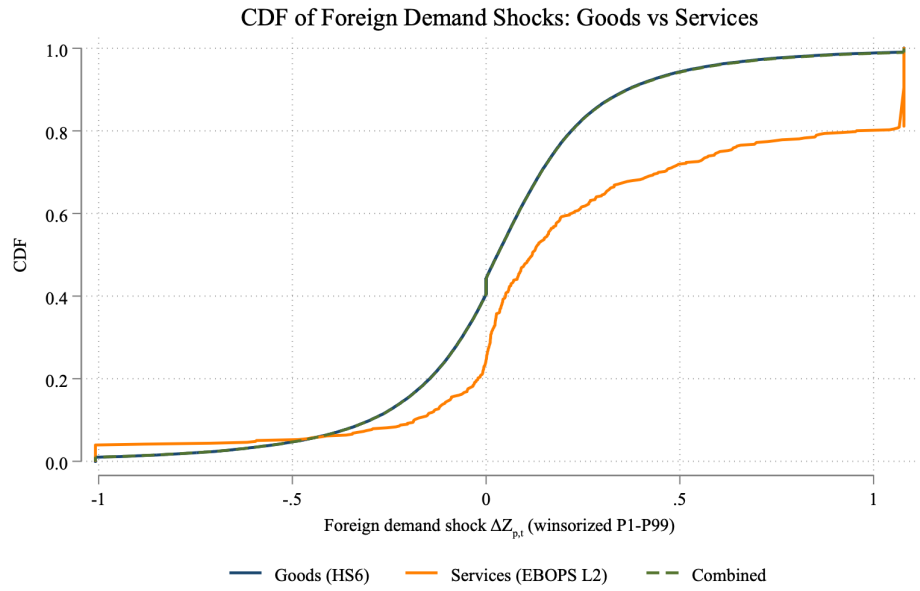


Figure 20: Distribution of Product Level Foreign Demand Shock Combined

of foreign demand.

Table A24 reports estimates in which the goods (HS6) and services (EBOPS Level 2) components of the export demand shock are entered separately. This decomposition allows us to assess whether the main results are driven by one margin of trade.

Table A24: Foreign Demand and Innovation: Side-by-Side (sales-weighted)

	Patent indicator (dummy)			Patent stock		
	(1) AI	(2) ICT	(3) Any	(4) AI	(5) ICT	(6) Any
Bartik Goods (HS6)	-2.668	-0.220	-0.111	-0.2080***	-0.140	0.0489
	(-0.15)	(-0.32)	(-0.70)	(-5.84)	(-1.00)	(0.50)
Bartik Services (EBOPS L2)	0.169***	0.0492	0.0257	0.0518***	0.0784**	0.0579*
	(4.49)	(1.16)	(0.57)	(4.22)	(2.89)	(2.01)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
NIC4D×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls×Year	Yes	Yes	Yes	Yes	Yes	Yes
N	1306	2417	14093	1597	6451	20435

Notes: Both goods (HS6) and services (EBOPS L2) Bartik instruments entered jointly. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The results indicate that the positive relationship between export demand and AI innovation is primarily driven by the services component of the export demand shock. The service-demand shock is positively associated with both the probability of filing AI patents and the accumulation of AI patent stocks, while the goods-demand shock is generally small and statistically insignificant for most outcomes. Together, these findings suggest that the main results reported in the paper are not driven by goods exports alone but instead reflect the importance of foreign demand for services, which is more closely related to the data-intensive activities underlying AI innovation.

F.2 Robustness

Table A25: Placebo Tests Using Future Foreign Demand Shocks

	AI patenting		AI patent stock	
	(1)	(2)	(3)	(4)
Foreign demand shock, $t+1$	-0.000		-0.535	
	(0.001)		(0.636)	
Foreign demand shock, $t+2$		-0.000		0.082
		(0.000)		(0.355)
Observations	51,180	51,180	185	258

Notes: This table reports placebo tests using future values of the foreign demand shock. Columns (1) and (2) use an indicator for AI patenting as the dependent variable, while columns (3) and (4) use the AI patent stock. Standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

G Appendix to GDPR Results

This appendix provides additional details on the construction of the treatment, the matching procedures, and the estimation strategy used in Section 6.

G.1 Constructing the GDPR Exposure Measure

CMIE Prowess reports firms' total exports but does not identify their export destinations. Consequently, we cannot directly observe each firm's exposure to the General Data Protection Regulation (GDPR) through its trade with the European Union. Therefore, we construct a predetermined firm-level measure of potential GDPR exposure by combining firms' pre-GDPR product portfolios with product-level information on the importance of the European Union as an export destination. The resulting measure summarizes the extent to which a firm's products were oriented toward EU markets prior to the GDPR's implementation.

Restricting the comparison to developed markets. A key design choice is to compute product-level destination shares within developed economies rather than relative to total world trade. This restriction serves two purposes. First, Indian firms exporting to developed markets differ systematically from firms serving developing economies in terms of size, innovation, and participation in global value chains.[9] Second, the natural counterfactual for an EU-oriented exporter is a firm exporting primarily to other developed economies that did not experience a GDPR-equivalent regulatory shock. Our measure, therefore, captures a firm's relative orientation toward the European Union within developed markets, rather than its orientation toward Europe relative to the entire world.

We define the set of developed economies (excluding India) as the United States, Canada, Japan, Australia, Korea, New Zealand, and the EU27 plus the United Kingdom, which remains subject to the GDPR during the post-Brexit transition period.

Step 1: Product-level EU orientation. For each product k (defined at the HS6 level for goods and EBOPS Level 2 for services), we compute the share of developed-country exports destined for the European Union:

$$\text{EU share}_k = \frac{\sum_{d \in \text{EU28}} X_{kd}^{\text{world}\setminus\text{IN}}}{\sum_{d \in \text{DEV}} X_{kd}^{\text{world}\setminus\text{IN}}}, \tag{11}$$

where $X_{kd}^{\text{world}\setminus\text{IN}}$ denotes exports of product k from all exporting countries except India to destination country d. The numerator sums exports destined for the European Union, while the denominator sums exports destined for all developed-market economies in DEV. We compute the product-level destination shares using goods trade flows from BACI and services trade flows from the WTO–OECD BaTIS. To ensure that exposure is predetermined, all shares are constructed using average trade flows over the pre-GDPR period 2013–2015.

[9] Restricting the comparison to developed-market trade reduces heterogeneity unrelated to GDPR exposure.

The measure EU share$_k \in [0,1]$ captures the share of exports of product k to developed-market destinations that is destined for the European Union. Products with values close to one are primarily exported to EU markets, whereas products with values close to zero are exported mainly to other developed economies.

Step 2: Firm-level product portfolios. For each firm f, we compute the share of each product in the firm's average sales during the pre-GDPR period:

$$w_{fk} = \frac{\overline{\text{Sales}}_{fk,2013\text{–}2015}}{\sum_{k'} \overline{\text{Sales}}_{fk',2013\text{–}2015}}, \tag{12}$$

where $\overline{\text{Sales}}_{fk,2013\text{–}2015}$ denotes average annual sales of product k between 2013 and 2015. Product sales are observed at the CMIE product-code level and mapped to HS6 (goods) and EBOPS Level 2 (services) using the crosswalk described in Appendix B.2. Portfolio weights are normalized so that $\sum_k w_{fk} = 1$ for every firm with at least one classified product.

Step 3: Firm-level GDPR exposure. We compute firm-level exposure as the weighted average of product-level EU orientation:

$$\text{EU exposure}_f = \sum_k w_{fk} \, \text{EU share}_k, \tag{13}$$

which ranges between zero and one.

We interpret this measure as follows. A firm with a high EU exposure$_f$ produces goods and services whose exports are disproportionately directed toward the European Union within developed markets, whereas a firm with a low EU exposure$_f$ specializes in products whose developed-market exports are primarily destined for other developed economies.

Step 4: Binary treatment definition. Firms with above-median values of EU exposure$_f$ are classified as treated ($D_f = 1$), while firms below the median constitute the control group. Firms with no classified exposure to developed markets are excluded from the main analysis because the exposure measure is undefined for these firms. Appendix G.4 reports robustness to alternative treatment thresholds and to continuous measures of exposure.

Step 5: Post-GDPR interaction. Treatment is indicated by an indicator equal to one from 2018 onward:

$$D_f \times \mathbf{1}[t \geq 2018], \tag{14}$$

which constitutes the treatment variable in the difference-in-differences specifications presented in Section 6.

Interpretation and identifying assumption. Identification exploits predetermined differences in firms' orientation toward European Union markets within developed-country trade, to-

gether with the common timing of GDPR implementation in 2018. The identifying assumption is that, absent the GDPR, firms with relatively high and low EU exposure would have followed parallel trends in the outcomes of interest, including AI patenting, R&D expenditure, and data-related investments. We assess this assumption using event-study specifications (Section 6).

The construction has three desirable features. First, product-level destination shares are computed using world exports excluding India, ensuring that the measure is not mechanically driven by Indian firms' own exports. Second, both product portfolio weights and destination shares are measured entirely before GDPR implementation, preventing endogenous adjustments in firms' product mix or export destinations from affecting exposure. Third, the measure captures firms' *potential* exposure to the GDPR rather than their realized exports to Europe after 2018, allowing identification to rely on predetermined differences in exposure rather than endogenous post-treatment destination reallocation.

G.2 Propensity Score Matching

Firms with high exposure to European markets differ systematically from less exposed firms in terms of size, innovation activity, and export behavior. Directly comparing these groups could therefore confound the effects of the GDPR with pre-existing differences in firm characteristics. To improve comparability, we combine propensity score matching with a difference-in-differences design. Following Bessen et al. (2020); Demirer et al. (2024), we estimate a logit model for the probability of treatment using pre-GDPR firm characteristics measured in 2016. The covariates include log assets, log sales, log employment, firm age, R&D expenditure, exporter status, log goods exports, log services exports, and log data expenditure.

Using the estimated propensity scores, we implement kernel matching with an Epanechnikov kernel and impose common support. Each treated firm is matched to a kernel-weighted average of control firms with similar propensity scores, allowing all comparable control observations to contribute to the counterfactual while assigning greater weight to firms with more similar observable characteristics.

The matched sample is used to estimate the following difference-in-differences specification,

$$Y_{it} = \beta(D_i \times Post_t) + \alpha_i + \alpha_t + \varepsilon_{it}, \tag{15}$$

where D_i denotes the treatment indicator, $Post_t$ equals one from 2018 onward, and observations are weighted using the kernel matching weights. All specifications include firm and year fixed effects, and standard errors are clustered at the NIC two-digit industry $\times$ location level.

Table A26 reports covariate balance before and after matching. Before matching, treated firms are significantly larger, more R&D intensive, and more export oriented than control firms. After kernel weighting, these differences become economically small and statistically insignificant, indicating that the matching procedure successfully balances the observed characteristics

of the two groups. Figure 21 shows substantial overlap in the propensity score

Table A26: Balance Table: Before vs After PSM (Year 2016)

	Unweighted			PSM-weighted		
Variable	Treated	Control	Diff	Treated	Control	Diff
Log assets	6.98	6.97	0.02	6.96	6.88	0.08
Log sales	9.20	9.38	-0.18	9.14	9.17	-0.03
Log employees	7.23	7.43	-0.19	7.19	7.19	0.00
Firm age	54.72	49.75	4.97	50.86	52.37	-1.51
R&D intensity	3.29	3.30	-0.00	3.24	3.20	0.03
Exporter	1.00	1.00	0.00	1.00	1.00	0.00
Log goods exports	6.79	6.73	0.06	6.73	6.68	0.04
Log services exports	4.21	3.96	0.25	3.99	4.09	-0.10
Log data exp.	1.80	1.26	0.54*	1.61	1.59	0.02

Notes: Balance comparison before and after kernel propensity score matching. Treated = above median EU share within developed exposure. Control = below median. PSM uses Epanechnikov kernel with common support. Baseline year 2016 (2 years pre-GDPR). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

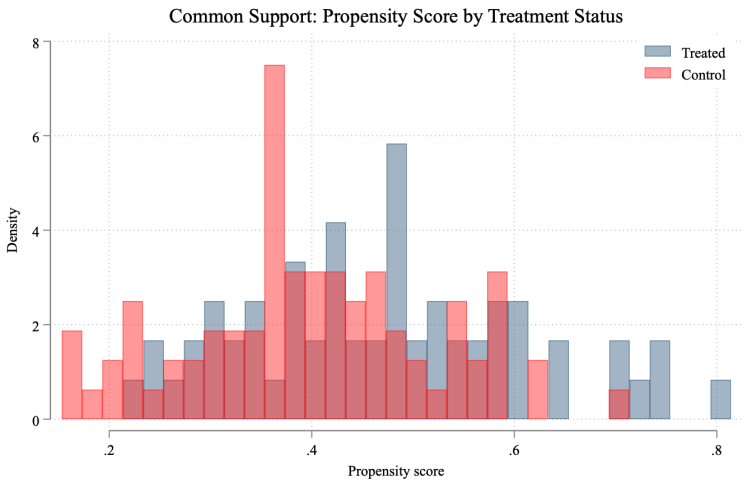


Figure 21: Propensity Score Distributions Before Matching

Notes: The figure shows the distribution of estimated propensity scores for treated and control firms before matching. Propensity scores are obtained from a logit model estimated using pre-GDPR firm characteristics, including firm size, sales, employment, firm age, R&D expenditure, exporter status, goods exports, services exports, and data-related expenditure. The substantial overlap between the two distributions supports the common-support assumption underlying the propensity score matching procedure.

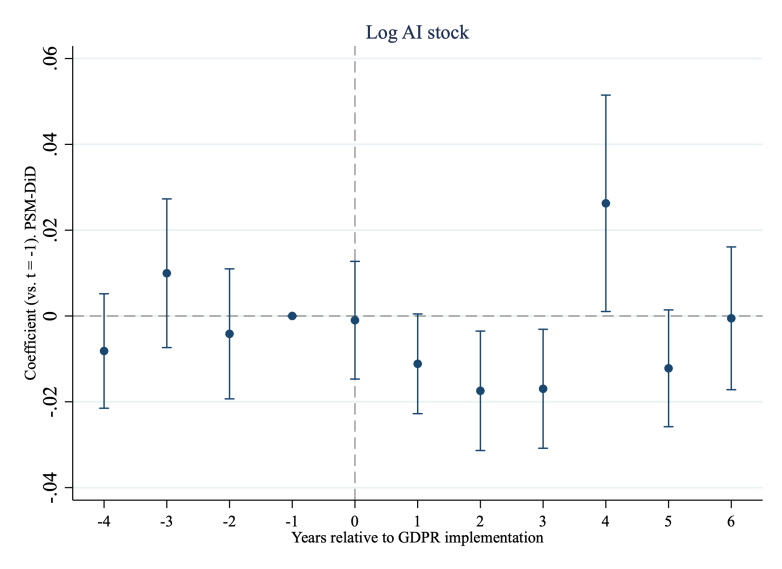


Figure 22: Propensity Score Matched Event-Study Estimates: GDPR Effects on AI Innovation

Notes: The figure reports event-study estimates for the log AI patent stock. Coefficients are estimated relative to the year immediately preceding the implementation of the GDPR (2017), which is omitted from the regression. The sample is weighted using Epanechnikov kernel propensity-score matching weights with common support. The specification includes firm and year fixed effects. Standard errors are clustered at the two-digit industry × location level. Vertical bars denote 95% confidence intervals.

Table A27: GDPR Effects on Innovation

	Patent Entry			Patent Intensity (Log Stock)		
	(1) AI	(2) ICT	(3) Any	(4) AI	(5) ICT	(6) Any
High EU Exposure × Post	0.002	0.024**	-0.009	-0.009*	-0.100	-2.518
	(0.003)	(0.009)	(0.018)	(0.005)	(0.092)	(2.194)
Estimator	Poisson	Poisson	Poisson	OLS	OLS	OLS
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Kernel PSM Weights	Yes	Yes	Yes	Yes	Yes	Yes
N	4,133	4,133	4,133	4,133	4,133	4,133

Notes: Difference-in-differences estimates using kernel propensity-score matching weights. High EU Exposure equals one for firms with above-median pre-GDPR exposure to the European Union, measured as the EU share within developed-country export exposure. Columns (1)–(3) report Poisson estimates of patent entry, while columns (4)–(6) report OLS estimates for the logarithm of patent stocks. All specifications include firm and year fixed effects. Standard errors are clustered at the two-digit industry × location level. Matching is performed using an Epanechnikov kernel with common support based on firms' 2016 characteristics, including assets, sales, employment, age, R&D intensity, exporter status, goods exports, services exports, and data-related expenditure.

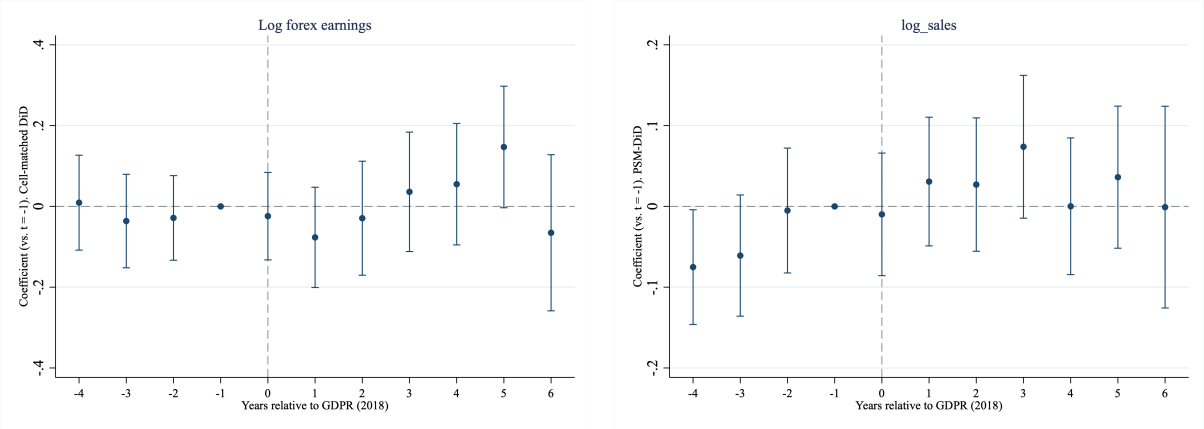
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A28: GDPR Effects on Trade and Firm Performance

	(1) Sales	(2) Goods Exports	(3) Services Exports	(4) Total exports	(5) TFP	(6) Exporter
High EU Exposure × Post	0.048	-0.141	0.915	-0.039	0.021	0.080
	(0.233)	(0.444)	(0.608)	(0.248)	(0.159)	(0.076)
Estimator	OLS	OLS	OLS	OLS	OLS	Poisson
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Kernel PSM Weights	Yes	Yes	Yes	Yes	Yes	Yes
N	3,373	2,893	1,817	3,221	3,059	4,133

Notes: Difference-in-differences estimates for firm performance and trade outcomes using kernel propensity-score matching weights. High EU Exposure equals one for firms with above-median pre-GDPR exposure to the European Union, measured as the EU share within developed-country export exposure. Columns (1)–(5) are estimated by OLS and column (6) by Poisson. All specifications include firm and year fixed effects. Standard errors are clustered at the two-digit industry × location level. Matching is performed using an Epanechnikov kernel with common support based on firms' 2016 characteristics, including log assets, log sales, log employment, firm age, R&D intensity, exporter status, log goods exports, log services exports, and log data expenditure.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

G.3 Event-study results



(a) Total exports

(b) Sales

Figure 23: Event-Study Estimates: GDPR Effects on Trade and Firm Performance

Notes: The figure reports event-study estimates for log total exports (Panel A) and log sales (Panel B). Coefficients are estimated relative to the year immediately preceding the implementation of the GDPR (2017), which is omitted from the regressions. The sample is weighted using Epanechnikov kernel propensity-score matching weights with common support. All specifications include firm and year fixed effects. Standard errors are clustered at the two-digit industry × location level. Vertical bars denote 95% confidence intervals.

G.4 Additional Robustness Checks

This appendix presents a series of robustness checks and placebo exercises, including alternative treatment definitions, tests of covariate balance and common support, evidence on firms' trade performance, and additional analyses evaluating the trade-cost channel. These exercises support the validity of the empirical design and the interpretation of the GDPR as a shock to firms' ability to collect and use commercially generated data rather than a reduction in access to European markets.

Aggregate Evidence on the Trade-Cost Channel An alternative interpretation of our results is that the GDPR acted primarily as an EU-specific trade cost, reducing Indian firms' access to European markets rather than limiting their ability to collect and process commercially generated data. If this were the case, one would expect Indian exports to the European Union to decline relative to exports to comparable non-EU developed destinations following the implementation of the GDPR. To examine this possibility, we conduct three complementary exercises using aggregate and product-level trade data.

First, we construct an aggregate panel of India's exports to major destination groups between 2013 and 2023 using BACI for goods and WTO–OECD BaTIS for services. We estimate a difference-in-differences specification comparing exports to the European Union with exports to other developed destinations before and after the implementation of the GDPR. The comparison group consists of the United States, Canada, Japan, Australia, Korea, and New Zealand. Table A29 reports the corresponding estimates. Contrary to the trade-cost interpretation, Indian exports to the European Union did not decline following the implementation of the GDPR. Instead, exports to the European Union grew faster than exports to other developed destinations after 2018. Although these aggregate patterns cannot identify firm-level responses, they provide complementary evidence that the GDPR did not materially reduce India's exports to the European Union.

Table A29: Aggregate India Trade by Destination

	Goods (BACI)		Services (BaTIS)	
	(1) Main	(2) + Placebo	(3) Main	(4) + Placebo
EU × Post	0.099*	0.102*	0.159***	0.161***
	(0.055)	(0.056)	(0.035)	(0.036)
USA-Placebo × Post		0.094		0.051
		(0.084)		(0.044)
Country FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	2,377	2,377	2,189	2,189

Notes: Country-year level regressions of log India exports on EU indicator interacted with Post-GDPR dummy. Panel comprises 2013–2023. Goods trade from BACI; services trade from BaTIS. If GDPR were primarily an EU-specific trade cost shock, the coefficient on EU × Post should be significantly negative. The joint specifications include the USA-placebo interaction to test whether any decline is EU-specific or a general developed-country effect. Standard errors clustered at destination country level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Second, we test whether products that were more exposed to the European market before the GDPR experienced a disproportionate decline in exports to EU destinations after the regulation came into force. We estimate a triple-difference specification using product-by-destination and year fixed effects, where the coefficient of interest interacts with a product's pre-GDPR EU orientation, an indicator for EU destinations, and the post-GDPR period. This coefficient captures whether exports to the European Union declined differentially for products that were initially more dependent on EU markets, relative to exports of the same products to other developed destinations. Under the trade-cost interpretation, the triple interaction is expected to be negative. The estimated coefficient is small and statistically indistinguishable from zero, providing no evidence of destination reallocation away from the European Union.

Table A30: Product-Level Test of the Trade-Cost Channel

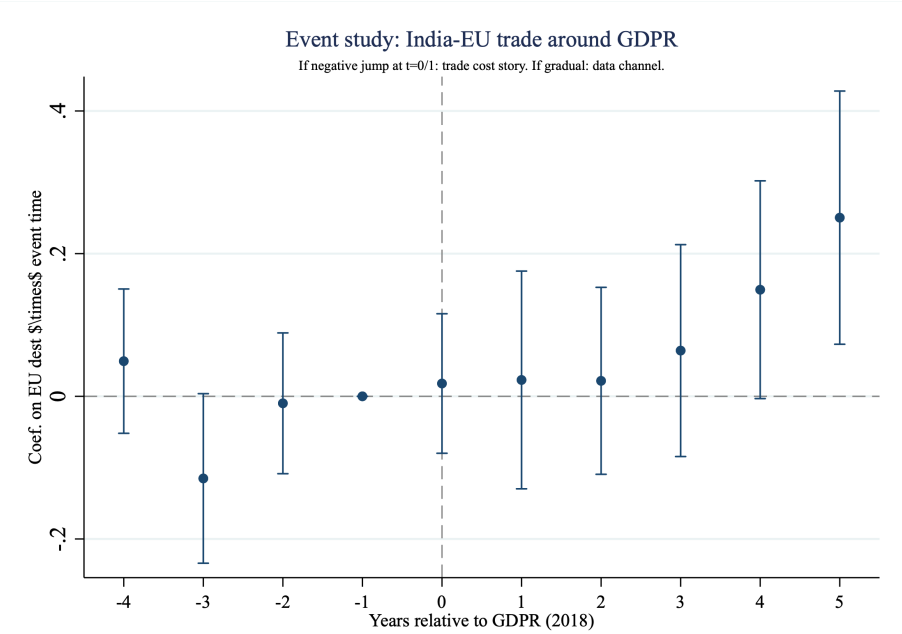
	(1) Log Exports
EU destination × Post	-0.218^{***}
	(0.055)
EU share × Post	0.216
	(0.133)
EU share × Post × EU destination	0.048
	(0.153)
Product × Destination FE	Yes
Year FE	Yes
Observations	97,660

Notes: Difference-in-differences estimates at the HS6 product × destination group × year level. The sample includes Indian goods exports to the EU28 and to the placebo group of developed destinations (United States, Canada, Japan, Australia, Korea, and New Zealand) over 2013–2023. EU share$_k$ denotes the pre-GDPR share of world (excluding India) exports of product k destined for the European Union. The coefficient of interest is the triple interaction, which tests whether products with greater pre-GDPR exposure to the European Union experienced a differential change in exports to EU destinations after the implementation of the GDPR. Under the trade-cost interpretation, this coefficient is expected to be negative. Standard errors are clustered at the HS6 product level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Additionally, Figure 24 presents an event study of India–EU trade around the implementation of the GDPR. We find no evidence of a discrete decline in exports immediately following the introduction of the regulation. Instead, trade evolves smoothly over time, providing little support for the view that the GDPR generated an immediate EU-specific trade-cost shock.

Overall, these results are difficult to reconcile with the view that the estimated effects on AI innovation were driven primarily by reduced access to European markets.

Figure 24: Event Study: India–EU Trade Around the GDPR



Notes: Event-study estimates of the differential evolution of India's exports to EU vs. non-EU destinations around GDPR implementation. Each point plots the estimated coefficient on the interaction of an EU destination indicator with a year dummy, using 2017 as the reference year. Coefficients are estimated from a country-year panel of log Indian exports (BACI, 2013–2023), controlling for country and year fixed effects. Vertical bars are 90% confidence intervals based on standard errors clustered at the destination-country level. The dotted vertical line marks the implementation of GDPR in May 2018. Under a trade-cost interpretation, a discontinuous negative drop should be visible at $t = 0$ or $t = 1$.