UNIVERSITÀ
CATTOLICA
del Sacro Cuore

DIPARTIMENTO DI POLITICA ECONOMICA

A Deep Learning procedure for the identification of Artificial Intelligence technologies in patent data

Francesco D'Alessandro

Working Paper n. 50 - June 2025

VP VITA E PENSIERO

Università Cattolica del Sacro Cuore

DIPARTIMENTO DI POLITICA ECONOMICA

A Deep Learning procedure for the identification of Artificial Intelligence technologies in patent data

Francesco D'Alessandro

Working Paper n. 50 - June 2025

VP VITA E PENSIERO

Francesco D'Alessandro, Department of Economic Policy, Università Cattolica del Sacro Cuore, Milano, Italy

✉ francesco.dalessandro2@unicatt.it

Dipartimento di Politica Economica
Università Cattolica del Sacro Cuore – Largo A. Gemelli 1 – 20123 Milano
Tel. 02-7234.2921

✉ dip.politicaeconomica@unicatt.it

https://dipartimenti.unicatt.it/politica_economica

ISBN digital edition (PDF): 978-88-343-5994-5

www.vitaepensiero.it

Abstract

This paper introduces a deep learning methodology employing transformer-based models to systematically identify Artificial Intelligence (AI) patents. We develop two domain-specific classifiers tailored to two foundational AI fields: Learning and Symbolic Systems (LS-SYS) and Robotics and Autonomous Systems (RA-SYS). Building on the BERT (Bidirectional Encoder Representation from Transformers) for Patents foundational model (Srebrovic & Yonamine, 2020), we propose a fine-tuning pipeline that unfolds in three stages. First, we derive two domain-specific lists of 221 weighted n-grams by mining the AI scientific literature, which are used to assemble seed sets for both domains extracted from the Patent Universe. Second, we apply the patent landscaping procedure (Abood and Feltenberger, 2018) to expand these seed sets and generate anti-seed examples via negative sampling. Third, we fine-tune each BERT model on the resulting training corpora, yielding classifiers that achieve F1 scores above 0.90, which are publicly available on the Hugging Face Hub. Our models reveal a sharp post-2012 "Deep Learning Revolution" inflection in AI patenting activity, particularly for the LS-SYS domain, alongside pronounced geographic concentration in the United States and Asia. Sectoral analyses reveal that ICT industries lead LS-SYS inventions, while a heavy industry core drives the RA-SYS domain, with collaboration networks and CPC-based maps corroborating these domain distinctions. Firm-level rankings spotlight ICT and software incumbents (IBM, Microsoft, Google), robotics manufacturers (Fanuc, Yaskawa), automakers (Toyota, Honda, Ford), and agile newcomers (Waymo, Zoox, X Development).

JEL classification: C45; O31; O33; O34

Keywords: Artificial Intelligence; Learning Systems; Symbolic Systems; Robotics; Autonomous Systems; BERT; Patents.

1. Introduction

Artificial intelligence (AI) is reshaping industries, economies, and societies at an unprecedented pace. Its ubiquity and extensive versatility have led scholars to characterize AI as both a General-Purpose Technology (GPT) (Bresnahan & Trajtenberg, 1995; Trajtenberg, 2019) and a General-Purpose Invention in the Methods of Invention (GP-IMI) (Agrawal et al., 2019; Cockburn et al., 2019). Yet, its economic impact extends far beyond mere productivity enhancement effects: AI accelerates the production of scientific knowledge and spurs breakthrough technological innovations (Cockburn et al., 2019; Agrawal et al., 2023; Agrawal et al., 2024), trigger new waves of creative destruction (Schumpeter, 1939; Aghion & Howitt, 1992) that simultaneously disrupt existing industries and give rise to entirely new sectors (Carbonara & Santarelli, 2023), and fosters the creation of new innovative ventures in high-tech, knowledge intensive sectors (D'Alessandro et al., 2025) which are recognized as primary drivers of long-run economic growth (Audretsch et al., 2006; Acs et al., 2009; Braunerhjelm et al., 2010). Accordingly, the widespread adoption of AI technologies could catalyze a new technological revolution, ushering in a standalone AI-driven technological paradigm that comes to dominate the process of scientific discovery (Dosi, 1982, 1988; Damioli et al., 2025). All in all, by reshaping innovation processes, competitive dynamics, and entrepreneurial ecosystems, AI stands as a cornerstone of structural economic change and economic growth (Gonzales, 2023).

The analysis of the heterogeneous economic implications of AI demands a robust identification of related technologies at various levels of analysis (e.g., firm-level, regional-level, and sectoral-level). Yet, despite AI's transformative promise, mapping the diffusion of AI-related technologies is complex in nature, which, in turn, further hinders empirical research on its determinants and economic implications (Giczy et al., 2022). Definitional ambiguities cause a first hurdle (Baruffaldi et al., 2020), which is a specific by-product of AI's broad applicability. As highlighted by the High-level Expert Group on Artificial Intelligence (2019), AI technologies encompass a broad set of methods, tools, and systems that enable machines to perform tasks that would require human intelligence; in this respect, AI systems display a certain degree of autonomy. Despite the above, the boundaries of such a technological field remain rather elusive (Van Roy et al., 2020). Therefore, the breadth of AI defies simple taxonomy and often renders nearly impossible the development of classification schemes able to trace AI developments across space and time. AI's inherent dynamism and its rapid fusion across diverse disciplines give rise to continually evolving research frontiers. While robust classification schemes can capture the state of knowledge at a fixed point in time, static measurement frameworks will inevitably lag behind and overlook nascent subdomains unless they are regularly recalibrated.

In the face of these constraints, prior scholarly efforts have been devoted to developing a diverse toolkit of techniques to identify AI developments and track its diffusion. These techniques range from publication-based analyses (e.g., patent filings or scientific papers), labor market studies (job postings or employees' CVs), to web-scraping approaches (e.g., corporate websites). However, when focusing on publication-based analyses, and especially on studies leveraging patents to track AI innovations, the most popular methodologies tend to be rooted either in keyword-only (e.g., simple keyword matching leveraging the unstructured component of patents) and classification codes-only approaches (e.g., looking for patents having pre-determined classification codes, therefore using structured information), or a combination of the two. In addition to being

affected by several hurdles, these approaches might also be characterized by a lower precision and recall compared to modern Machine Learning (ML) methodologies that leverage Natural Language Processing (NLP). In this respect, both Supervised and Unsupervised[1] ML methods may be better suited to identify the technological content of patents, given the sheer volume and linguistic complexity of modern patent corpora. Machine Learning-based methods, as well as modern deep-learning architectures, unlock capabilities that far exceed manual or rule-based schemes. By ingesting millions of documents at once, ML algorithms can automatically discern subtle patterns in terminology, syntax, and citation linkages that no team of human coders could reliably detect at scale (Lamperti, 2024). In sum, by leveraging ML's ability to process massive datasets, extract complex semantic structures, and generate predictive insights, researchers gain a powerful toolkit for mapping AI's development and diffusion.

In this paper, we focus specifically on patent-based identification techniques to map AI innovation. Leveraging the richness of the unstructured information available in patent filings, we build a high-fidelity framework targeting two foundational AI domains: "Learning and Symbolic Systems" (LS-SYS) and "Robotics and Autonomous Systems" (RA-SYS). Our approach unfolds in three stages: we begin by crafting two domain-specific lists of n-grams drawn from the AI scientific literature, which are used to isolate a seed set for each domain. Each seed set is expanded using the patent landscaping procedure (Abood and Feltenberger, 2018), which is leveraged to identify anti-seed patents through negative sampling in our Patent Universe (consisting of patent families containing at least one application filed at the USPTO, EPO, or WIPO[2] from 1980 to 2021). Finally, using seed and anti-seed sets built in the previous stages, we fine-tune two BERT for Patents models (Srebrovic & Yonamine, 2020) to distinguish genuine AI disclosures from false positives initially included in the seed expansions.

We complement our methodological framework with an empirical, descriptive assessment of the classification results. In particular, we document a dramatic surge in AI patenting over time, particularly in the last decade following the Deep Learning Revolution, with the LS-SYS domain outpacing the RA-SYS field. We further highlight that the production of AI-related technologies is clustered in the United States and Asian countries, with Europe lagging behind. Our sectoral analysis [3] reveals that ICT-related industries overwhelmingly lead LS-SYS innovation, whereas RA-SYS output is more diffusely spread across core ICT, machinery, and transportation manufacturing. Furthermore, collaboration-network metrics confirm that a small number of sectors act as pivotal knowledge hubs, channeling critical AI expertise across industry boundaries, with this effect being increasingly dominant for the LS-SYS field. Finally, by ranking AI patent applicants in each domain, we demonstrate that our classification procedure reliably identifies principal actors driving the AI revolution, spanning major technology firms, established industrial incumbents, and specialized AI startups.

[1] As thoroughly described by Alloghani et al. (2020), the difference between Supervised and Unsupervised approaches is the use of labelled training data in the former as opposed to the latter. Supervised learning exploits a labelled dataset, where each input vector is paired with a ground-truth, predetermined label (e.g., AI vs. not AI). During training, a model is optimized to minimize a loss function, thereby reducing the rate of misclassification on held-out data. In contrast, unsupervised learning operates solely on an unlabelled dataset, seeking to discover intrinsic structures and hidden patterns in input data. The representations or clusters inferred by unsupervised algorithms can subsequently serve as feature transformations or initialization priors for downstream supervised tasks.
[2] United States Patent and Trademark Office, European Patent Office, and World Intellectual Property Organization.
[3] Using patents' applicants and their core NACE Rev. 2 sector as a reference.

Our work makes three key contributions. First, we provide an assessment of the most common methodologies and approaches implemented in prior studies to trace AI developments. Our synthesis emphasizes that each approach yields distinct yet equally vital insights into the patterns, timing, and economic impacts of AI diffusion. This comparison equips researchers with a clear roadmap for matching analytic tools to specific research questions, such as tracking nascent scientific breakthroughs, gauging real-time labor demand, or monitoring firm-level adoption.

Second, we develop and release two specialized BERT-for-Patents models, one tailored for LS-SYS and another for RA-SYS, that integrate n-gram seeding, patent landscaping, and targeted negative sampling to accurately identify patents within these two AI subdomains. Both models are freely accessible on the Hugging Face Hub, allowing researchers to extend and refine them. Alongside these models, we provide two curated keyword lists that can be used independently or in combination with machine-learning approaches. Together, these resources establish a solid foundation for future studies on AI technology diffusion and impact, going beyond traditional classification frameworks to capture the co-evolution of two interrelated AI fields (Cockburn et al., 2019).

Third, we present a descriptive analysis of global AI patenting activity, highlighting nuanced differences in how LS-SYS and RA-SYS innovations spread across sectors and over time. Our findings, besides being in line with prior research, further enrich the empirical literature on AI diffusion by unpacking domain-specific trajectories and offering fresh insights into the worldwide dynamics of AI innovation. In addition, through collaboration-network analysis, we identify a small set of sectors that function as critical AI knowledge hubs, channeling expertise across industry boundaries, an effect especially pronounced within LS-SYS. These findings point to two promising avenues for future inquiry: (1) investigating the co-evolution of LS-SYS and RA-SYS capabilities across firms, regions, and sectors; and (2) elucidating the mechanisms and determinants of AI-related knowledge flows.

The remainder of the paper is organized as follows. Section 2 delineates the three principal methodological paradigms for identifying AI technologies, with particular emphasis on patent-detection techniques. Section 3 describes the procedures employed to fine-tune the two deep-learning models. Section 4 walks through each step of the data preparation and training pipeline in detail. Section 5 reports the outcomes of the fine-tuning stage and the subsequent inference results. Section 6 offers an empirical validation of our classification results, benchmarking them against existing studies. Finally, Section 7 concludes by summarizing our key contributions, acknowledging the study's limitations, and outlining directions for future research.

2. Background literature on the identification of AI technologies

In recent years, scholarly work has produced an extensive repertoire of methods for detecting AI innovations and monitoring their dissemination. These approaches fall into three broad categories: (a) Publication-based analyses, which mine and leverage scientific articles or patents filings to map AI research and its technological developments (Cockburn et al. 2019; WIPO, 2019; Alderucci et al., 2020; Baruffaldi et al., 2020; Dunham et al., 2020; Van Roy et al, 2020; Bianchini et al., 2022; Giczy et al. 2022; Montobbio et

al., 2022; Savin et al., 2022; Miric et al., 2023; Mann & Püttmann, 2023; Pairolero et al. 2025) (b) Labor market studies, which examine job advertisements and professional résumés to infer demand for AI skills, track workforce evolution and proxy AI-related investments (Alekseeva et al., 2021; Squicciarini & Nachtigall, 2021; Acemoglu et al., 2022; Borgonovi et al., 2023; Babina et al., 2024); (c) Web-based inquiries, which scrape corporate websites for evidence of AI adoption, from product features to strategic initiatives (Colombelli et al., 2023; Dernis et al., 2023; Dahlke et al., 2024; Dahlke et al., 2025). Some researchers even combine multiple sources, integrating publication, labor-market, and web-derived signals, to gain a more holistic picture of AI diffusion (Calvino et al., 2022; Calvino et al., 2024). Despite their reliance on such varied data, ranging from highly structured records (e.g., patent databases) to unstructured texts (e.g., websites), all these approaches rest on a shared premise: the massive volume of data now available constitutes one of the few systematic archives of technological progress, offering a window into how AI is developed, disseminated, and implemented[4].

Crucially, each approach illuminates a different dimension of AI diffusion. Scientific publications reveal three aspects: firstly, the frontier of basic research, pinpointing where and how new AI concepts emerge; second, the adoption of AI-related tools in science, underscoring AI's role as a GP-IMI (Bianchini et al., 2022); third, the AI research strength of universities and research centers, serving as a proxy for the local supply of AI talent (Babina et al., 2024). Patent-based indicators map the trajectory of applied research and AI innovation, offering insights into which AI technologies are moving toward market deployment. These metrics are also used as a proxy of the AI endowments of firms, sectors, and regions (Cicerone et al., 2023; Grashof & Kopka, 2023). Labor-market signals flag the organizational uptake of AI by indirectly inferring the adoption and use of or investments in AI technologies by firms. Web-scraping techniques provide real-time visibility into firms' strategic use of AI, whether embedded in products, services, or internal processes. These methodologies are also suited to capture AI innovation, depending on the information available on companies' websites. Finally, the inherently relational structure of web data allows researchers to map collaboration networks and the diffusion pathways of AI across organizations (Dahlke et al., 2024; Dahlke et al., 2025).

2.1 Scientific literature, labor market insights, and web scraping

In this section, we briefly review works in the three complementary streams of AI measurement outlined above—mining the scientific literature, analyzing labor-market signals, and scraping corporate websites—to illustrate how each captures different facets of AI development and diffusion. We set aside patent-based approaches here and defer their detailed treatment to Subsection 2.2.

Among studies that have investigated the scientific literature, recent efforts have sought to delineate AI-related developments by leveraging both supervised/unsupervised approaches and keyword-driven retrieval strategies. Dunham et al. (2020) exploit the CoRR subject taxonomy on arXiv.org, comprising 39 categories,

[4] In contrast to the thorough analysis provided by Calvino et al. (2024), we deliberately excluded survey-based studies, leveraging sources such as the annual Eurostat Survey on ICT (Information and Communication Technologies) usage and e-commerce in enterprises, which only recently began incorporating questions on AI adoption. In this respect, access limitations and confidentiality constraints severely restrict researchers' ability to exploit these data sources.

to isolate six core AI domains (Artificial Intelligence; Computation and Language; Computer Vision and Pattern Recognition; Machine Learning; Multiagent Systems; Robotics). They assemble a corpus of 85,670 papers tagged in at least one of these domains and train six one-versus-all classifiers based on the SciBERT architecture, each dedicated to detecting publications in a single AI area[5]. Finally, for out-of-domain inference, the authors applied these models to the broader Web of Science corpus. Bianchini et al. (2022) apply NLP techniques to create a comprehensive "Neural Network-related" search list, mining publications on arXiv.org in the areas of Computer Science, Mathematics, and Statistics. The authors estimated vector representations of the words in the vocabulary (extracted from papers' abstracts), using Word2Vec. After estimating word embeddings, the authors performed a cluster analysis and retained the 30 most frequent n-grams belonging to the Neural Network cluster. This list of keywords was then used to retrieve scientific publications in the Web of Science Core Collection related to Neural Networks, whose title, keywords, or abstract contain at least one of the selected terms.

In parallel, labor market analyses are characterized by a similar array of methodologies. Alekseeva et al. (2021) proposed a list of 71 skills in the Burning Glass Technologies job vacancies dataset[6], which were then used to identify AI-related job vacancies, defined as job postings containing at least one AI skill. Squicciarini and Nachtigall (2021) leverage the Burning Glass Technologies dataset and employ the list of keywords compiled by Baruffaldi et al. (2020)[7]. Keywords are split into three groups: "generic", "AI approaches", and "AI applications". Furthermore, the authors augmented the list by including AI software and libraries. Finally, this list is used to trace AI-related jobs, namely those mentioning at least two different AI terms. Building on the foundation laid by Squicciarini and Nachtigall (2021), Borgonovi et al. (2023) distinguish between "generic" and "specific" AI skills: the authors define job vacancies as AI-related if their text contains at least two generic or one specific skill[8]. As noted by Alekseeva et al. (2021), vacancy data is affected by a critical limitation: while providing a detailed proxy of the demand for AI skills, it does not account for what happens next, namely, whether the position is filled. To this end, Babina et al. (2024) leverage employee résumés, in addition to Burning Glass data, to measure the actual stock of AI workers in each firm using Cognism, which also provides disambiguated information on employment records of individuals. Starting from the Burning Glass data, the authors compile a data-driven list of AI-related skills: they compute an AI-relatedness score for each unique skill based on its co-occurrence with core AI competencies[9]. This step led the authors to compile a list of 67 keywords, covering AI-related skills, which were then used to mine individuals' CVs: a subject is designated as an AI specialist if their résumé, including publication and patent sections, contains at least one keyword.

[5] Accordingly, for each CoRR AI area, the seed set comprises all papers tagged with the target subject, while the anti-seed set contains papers tagged with any of the remaining AI domains.
[6] As described in a white paper published in 2019, Burning Glass Technologies analyses the text of each job vacancy using Big Data and NLP techniques to identify relevant skills required to perform each job. Each skill is then organized by type (Baseline, Technical, and Software) and in a three-layer taxonomy (Skill, Skill Clusters, and Skill Cluster Families). In 2019, the list included more than 17,000 skills.
[7] The list of keywords is borrowed from Baruffaldi et al. (2020), which will be detailed in sub-section 2.2.
[8] The authors provide the following examples: "machine learning", "artificial intelligence", "computer vision", and "machine translation" for generic AI skills. Instead, specific AI skills are "gradient boosting", "natural language processing", "convolutional neural networks", and "deep learning".
[9] The core AI skills are "artificial intelligence", "machine learning", "natural language processing", and "computer vision".

Finally, web-based studies have combined keyword heuristics with language modeling to identify AI adopters among firms and startups. Colombelli et al. (2023) develop a top-down, bottom-up process to identify AI-related startups in Italy, leveraging their web pages. The authors begin with an initial lexicon of 72 AI terms to scrape 9,881 Italian startup websites, identifying 521 "unambiguous" AI ventures. These 521 websites were further mined to enrich the initial list of keywords, yielding a final list of 272 AI-related keywords. The authors repeated the scraping exercise using the extended list, enabling the discovery of an additional 11 AI startups. Dahlke et al. (2024) web-scraped over 1.1 million websites of firms in Germany, Austria, and Switzerland, gathering textual and hyperlink-based relational data. To identify firms adopting AI technologies, the authors fine-tuned a sentence-transformer language model to categorize relevant paragraphs on firms' websites as related to "Deep AI knowledge" or "Superficial AI knowledge"[10]. Relevant paragraphs were selected by implementing a keyword search, retrieving 247,846 passages with at least one keyword match. Subsequently, the authors selected 3,000 paragraphs for the fine-tuning procedure. The resulting language model was then used to predict whether each of the remaining paragraphs supplies Deep vis-à-vis Superficial AI knowledge. A firm is classified as an AI adopter if its website hosts at least one paragraph predicted to exhibit Deep AI knowledge.

2.2 AI technologies and Patent data

The accurate delineation of AI-related inventions in patents remains a formidable challenge, owing to both the multifaceted technical disclosures characteristic of patent filings and the dynamic nature of the AI realm. Over the past decade, scholars have therefore pursued a spectrum of methodologies to detect AI advances within patents, leveraging their metadata and full texts. Early efforts typically relied on rule-based heuristics (such as curated keyword searches or filtering by International Patent Classification (IPC) or Cooperative Patent Classification (CPC) codes) to flag potentially relevant documents. More recently, studies have also proposed mixed approaches as well as pure machine-learning models.

Cockburn et al. (2019) introduce a hybrid, two-pronged methodology for the identification of AI inventions in the U.S. patent corpus, whereby the AI domain is defined as comprising three different fields: Learning Systems, Symbolic Systems, and Robotics. First, they exploit the U.S. Patent Classification (USPC) system: all patents assigned to class 901 are designated as robotics-related, while those falling under specific subclasses of class 706 are further partitioned into "Symbolic Systems" and "Learning Systems." Second, they conduct a targeted title-based keyword search, employing curated term lists organized by AI subdomain (Learning Systems, Symbolic Systems, Robotics) to capture additional patents whose titles refer explicitly to AI methods or applications. After deduplication, the union of the classification-based and keyword-based retrievals constitutes the final AI-patent dataset.

[10] A company supplies Deep AI knowledge if: (a) provides products or services with integrated AI company know-how; (b) provides products or services with implemented AI from other companies' AI external know-how; (c) has or seeks personnel with AI expertise personnel know-how. Instead, a company supplies Superficial AI knowledge if it provides general information on the topic of AI, such as news, events or blog articles.

The World Intellectual Property Organization (WIPO, 2019) proposes a three-fold heuristic framework for delineating AI-related inventions within patent records by combining patent-classification filters with targeted keyword searches. Patents are deemed AI-relevant if they satisfy any of the following criteria: (1) Patents should be assigned to at least one CPC code (group or subgroup levels) included in the "Block 1" list, which enumerates the CPC entries most closely associated with core AI technologies; or (2) At least one keyword from the "Block 2 K1" [11] lexicon, comprising terms that denote fundamental AI concepts, must appear in patents title, abstract, or claims; or (3) Patents title, abstract or claims should contain at least one of the keywords related to general computing or mathematical concepts included in the "K2" list (frequently used in AI technologies, but not specific to them). In addition, patents should also be assigned to one of the (less detailed) classification codes identified in lists C1, C2, C3, and C4; the resulting intersection represents the "Block 3". AI-related patents are, therefore, those identified by the union of the three main search blocks.

Similarly to its predecessors, Baruffaldi et al. (2020) developed a methodology that combines both patent classification codes and keywords. This work distinguishes itself by proposing a meticulously curated lexicon of 193 AI-specific terms, extracted via comprehensive text analysis of leading AI journals and conference proceedings[12]. A patent is deemed AI-related if it satisfies any of the following criteria: (1) it is classified under at least one of the IPC codes listed in Table C.2.1; or (2) it carries an IPC or CPC codes listed in Tables C.2.2 and C.2.3, respectively, and contains at least one keyword from the "AI-193" list; or (3) its title, abstract, or claims include three or more distinct terms drawn from the AI-193 lexicon.

Leveraging modern machine learning methodologies and the automated patent landscaping procedure (Abood and Feltenberger, 2018), Giczy et al. (2022) employed long short-term memory (LSTM) neural networks to automatically classify patents in eight AI-related fields[13], using USPTO patent data. In more detail, for each AI component, the authors built a bespoke LSTM neural network. Seed sets are generated automatically by aggregating patents tagged with a curated list of classification codes specific to each AI domain[14], while "anti-seed" examples are randomly sampled from patents outside the initial landscaping-related expansion[15]. Furthermore, input features comprise the textual embeddings of abstracts and claims, augmented by one-hot encodings of backward and forward citations. Building upon their earlier LSTM-based pipeline, Pairolero et al. (2025) integrate a domain-specific transformer encoder into their classification architecture and enrich their training corpus. Firstly, the authors estimate word embeddings using BERT for Patents model (Srebrovic & Yonamine, 2020). Secondly, each training set is augmented by including, on the one hand, patent documents manually labeled by USPTO patent examiners during the evaluation of the original approach. On the other hand, they also selected and manually labelled "boundary" patents, namely, those

[11] The keywords list "K1" was built via a deep exploration of AI-related bibliographic records, and terms were selected based on their high degree of specificity (WIPO, 2019).
[12] The authors leveraged Scopus® "All Science Journals Classification" (ASJC), which classifies scientific publications helping readers find publications in specific areas, including Artificial Intelligence.
[13] Knowledge processing, Speech Recognition, AI Hardware, Evolutionary Computation, Natural Language Processing, Deep Learning, Computer Vision, Planning and Control.
[14] The CPC system, the IPC system, the USPC system, and Derwent's patent index.
[15] The authors employed a two-level expansion: firstly, each seed set was expanded by family members, using relevant CPC codes and citations; the resulting patents constitute Level 1 expansions. Furthermore, the L1 set was further expanded using family members and citations, creating Level 2 expansions. Therefore, anti-seed sets are selected from patents that should theoretically be technologically distant from those constituting seed sets.

documents that were originally assigned a predicted AI probability within a narrow band around the 0.50 decision threshold. By retraining the enhanced model on this augmented dataset, the authors demonstrate that these inclusions yield more discriminative representations and improve the separability of patents that exhibit subtle AI characteristics.

Miric et al. (2023) develop a supervised ML pipeline to detect AI-related patents employing any form of "statistical learning" methodology in the broader USPTO corpus. They begin by assembling a training corpus of 4,000 patent documents, each annotated according to Nilsson's (2010) foundational definition of AI and the taxonomy introduced by Cockburn et al. (2019)[16]. The authors mainly focus on the "Learning Systems" field identified by Cockburn et al. (2019), leveraging their list of keywords and general conceptualization. With the abovementioned training set[17], the authors train and compare a suite of classification algorithms using features derived from patent abstracts.

Mann and Püttmann (2023) focus on the broad domain of "automation," which they define as inventions enabling a device to carry out a task with little or no human intervention. Their scope spans both software innovations (such as financial-management applications and automated e-mail workflow systems) and hardware solutions like industrial assembly robots and self-checkout kiosks. Crucially, however, they restrict their analysis to patents with clearly identifiable end-use applications, excluding any filings lacking an immediately recognizable practical implementation (e.g., methods relating to deep learning). To classify these inventions, the authors train a Bernoulli Naïve Bayes model on a hand-labeled sample of 560 patent documents. They then apply the trained classifier to more than five million USPTO-granted patents from 1976 through 2014, systematically mapping the evolution of automation technologies over nearly four decades. Similarly, Santarelli et al. (2023) dissect the structural underpinnings of "automation technologies" by proposing a core–periphery analysis using USPTO patent data. Their retrieval strategy pairs Van Roy et al. (2020)'s curated AI keyword taxonomy with CPC classifications mapped from legacy USPC codes 901 (robotics) and 706 (narrow AI).

Finally, several studies have exclusively concentrated on robotics and autonomous systems. The United Kingdom Intellectual Property Office (UKIPO) (2014) combined IPC and CPC classification codes with targeted n-grams (e.g., "robot," "unmanned vehicle," "self-driving car") to identify 35,151 relevant patent families filed worldwide between 2004 and 2013. More recent research has augmented such mixed search strategies with unsupervised learning methods to uncover deeper thematic structures. In particular, Montobbio et al. (2022), first retrieve robotics patents from the USPTO using keywords and CPC codes, then isolate labor-saving inventions by detecting specific trigrams, namely, combinations of verbal predicates, direct objects, and attributes, which are indicative of labor-saving heuristics[18]. They further apply Latent Dirichlet Allocation (LDA) to estimate the relevance of 20 topics, which enabled a thorough differentiation between general robotics patents and those emphasizing labor-saving functionalities. Likewise, Savin et al. (2022) employ a

[16] In particular, Cockburn et al. (2019) recall a famous passage from Nilsson (2010), who defined AI as "that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment."
[17] 20% of documents are AI-related.
[18] Such as "lower labor cost" or "automate employees' tasks".

similar hybrid search and LDA modeling approach, supplementing topic estimates with textual descriptions from the International Federation of Robotics to differentiate "service-related" robotics patents (e.g., logistics and healthcare applications) from traditional industrial robots used in assembly-line contexts.

While this section does not aim to provide an exhaustive review of the existing literature, it acknowledges foundational contributions to patent-based AI identification methodologies. A clear trend emerging from the evolution of this field is the growing adoption of modern computational techniques, such as text mining, natural language processing (NLP), and machine learning, that are better equipped to capture AI's complexity and rapid evolution. A common limitation among earlier studies lies in their overreliance on patent classification codes. On the one hand, the dynamic and rapidly evolving nature of technological innovation renders classification systems susceptible to obsolescence, requiring constant revision and adaptation to effectively capture emerging fields. On the other hand, these codes may fail to fully represent the underlying technological content of patents, leading to a higher risk of Type II errors (i.e., false negatives). In response, many studies have adopted hybrid strategies that combine classification codes with keyword-based searches. However, this approach introduces a different challenge, namely, an increased risk of Type I errors (i.e., false positives), particularly when simple keyword searches are performed, given their lack of contextual or semantic sensitivity. As emphasized by Abood and Feltenberger (2018), modern machine learning approaches offer a promising alternative by addressing many of these limitations. Specifically, such models: (a) learn latent patterns and semantic regularities within a curated seed set of AI-related patents; (b) improve the ability to distinguish these from unrelated (anti-seed) patents; (c) generalize effectively to previously unseen patent documents during inference; and (d) substantially reduce the burden on human experts, as the model autonomously captures complex domain-specific signals without requiring exhaustive manual labeling or deep field-specific knowledge. Consequently, machine learning, and particularly deep learning, provides a scalable and context-aware framework for AI patent identification.

3. The proposed methodology

In alignment with prior research, we adopt a methodology that leverages recent advancements in deep learning to identify AI-related patents. Specifically, we fine-tune the BERT for Patents model (Srebrovic & Yonamine, 2020), an extension of the BERT architecture (Devlin et al., 2018) trained on over 100 million patent publications from the U.S. and other countries[19]. This approach allows for a more accurate classification of AI patents compared to relying solely on keyword searching or classification codes.

Following Cockburn et al. (2019), we define the AI field to be composed of two main macro-domains. First, the "Learning and Symbolic Systems" domain encompasses both symbolic AI, characterized by rule-based reasoning and knowledge representation, and machine learning techniques that allow systems to learn from data. Symbolic AI focuses on manipulating symbols and applying logical rules to simulate human

[19] As detailed in section A7, the authors train and release the first BERT model pre-trained exclusively on patent text, using a custom tokenizer that preserves long, technical patent terms. These adaptations enable more accurate, context-aware synonym suggestions for prior-art searching. The proposed architecture allows immediate reuse of the resulting encoder representations for downstream tasks–such as multiple or binary classifications–without the need to retrain the core Transformer architecture.

reasoning. Learning-based approaches, such as deep learning, enable the extraction of patterns from large datasets to make predictions, adapt over time, and progressively refine their performance, representing the brain-inspired domain (Sejnowski, 2018). Collectively, the LS-SYS is often referred to as "narrow AI", since only the software component is considered[20]. Secondly, the "Robotics and Autonomous Systems" domain pertains to AI applications that integrate perception and actuation to perform tasks in either fixed or dynamic environments. This domain, by contrast, focuses on the physical embodiment of intelligence, creating mechanical systems that interact with and navigate the physical world, enabling them to perform physical and human-like tasks. This includes the development of autonomous vehicles, drones, and robots that can navigate, perceive, and interact with their surroundings with or without human intervention. By focusing on the physical instantiation of intelligence, RA-SYS captures the "acting and planning" dimension of AI, complementing LS-SYS's emphasis on "thinking and learning" (Van Roy et al., 2020), and underscores the inseparable interplay of hardware and software in modern AI systems. The abovementioned discourses led us to define AI following a broader perspective (Damioli et al., 2024).

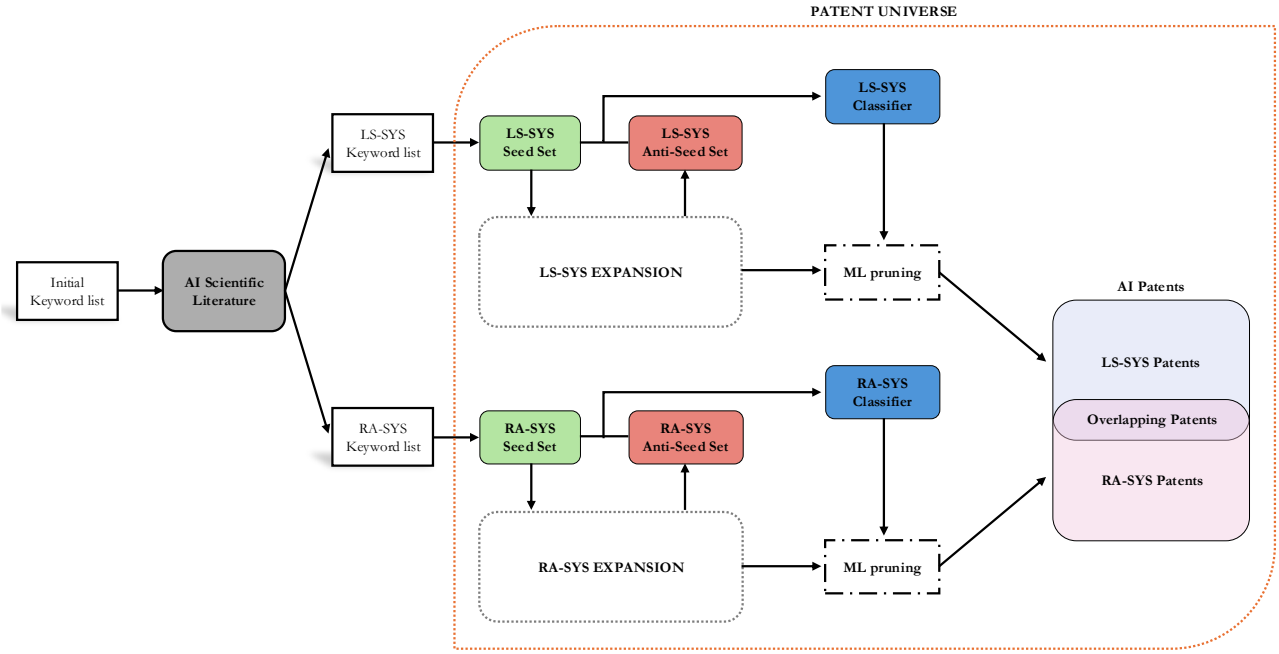
Our fine-tuning proceeds in several stages. Firstly, we employ a hybrid top-down and bottom-up strategy (Colombelli et al., 2023), exploring the AI scientific literature through text mining techniques to compile a list of 221 weighted n-grams. Each term is assigned to one of two principal AI domains: Learning and Symbolic Systems and Robotics and Autonomous Systems. The resulting keyword lists are then used to query our Patent Universe, composed of patents filed to various Patent Offices over the last 40 years, to derive accurate and representative seed sets for training two different machine learning models (one for each AI domain). By spanning this four-decade horizon, the models are thus able to capture and adapt to the diachronic evolution of the AI terminology. As in Giczy et al. (2022) and Pairolero et al. (2025), we adopt the Patent Landscaping Procedure (Abood and Feltenberger, 2018) to expand the seed sets and collect anti-seed sets. In particular, we randomly select patents that were not included in the expansions to construct the negative (anti-seed) training samples. Subsequently, leveraging the pre-trained BERT for Patents backbone and the training samples described above, we fine-tune two separate classifiers, one for each AI-related domain. After fine-tuning, we employ the two domain-specific classifiers to systematically prune patents erroneously retained during the seed-set expansion phase, thereby ensuring that only documents truly pertinent to each AI subdomain are preserved.

Our analysis diverges from that of Giczy et al. (2022) in several key respects. First, we build upon the extensive training of an existing neural network with a transformer-based architecture, which enables a more refined interpretation of patent content by capturing contextual relationships within the text via self-attention (Vaswani et al., 2017; Choi et al., 2022). Second, leveraging the PATSTAT database, we apply our model to patents filed across multiple patent offices. Given data availability, this cross-jurisdictional scope necessitates reliance solely on English-language patent titles and abstracts. Third, our methodology constructs the initial seed sets using a combination of weighted keywords and targeted heuristics, which in turn supports the

[20] Typically, most studies define AI from a narrow perspective, whereas our work provides the tools to expand the analysis. By fine-tuning two distinct models, it is possible to flexibly target either the LS-SYS or RA-SYS domain, or both, depending on specific research questions.

development of a taxonomy that differentiates between two fundamental AI domains: LS-SYS and RA-SYS. Although this curated vocabulary may bias the models toward conservative classifications, its breadth and semantic diversity ensure robust generalization across evolving AI subfields. Moreover, it allows scalability and full reproducibility. Figure 1 summarizes the proposed procedure.

Figure 1. Methodological diagram



4. The AI identification procedure

In this section, we provide a detailed assessment of the procedure used to create two fine-tuned BERT for Patents models. We start by explaining the procedure that resulted in the creation of the two lists of keywords. We then move to the seed sets construction and landscaping procedure. Finally, we detail the models' architecture.

4.1. A new keyword list

The purpose of this subsection is to outline the procedure that culminated in a new, comprehensive list of keywords aimed at capturing not only narrow AI tools (i.e., those reasoning-inspired and software-based) but also a broader set of technologies that have progressively contributed to the development of the AI technological paradigm and overall "computational intelligence field" (van Eck & Waltman, 2007; Baruffaldi et al., 2020; Vannuccini & Prytkova, 2023). The construction of this list follows a methodology similar to that of Colombelli et al. (2023), combining a top-down approach with a complementary bottom-up procedure to ensure both theoretical grounding and empirical robustness. The process begins with a simplified version of the keyword list proposed by Damioli et al. (2024), which serves as the initial query input in Elsevier's Scopus database. Using this list, we retrieved scientific publications that contain at least one of the predefined n-grams

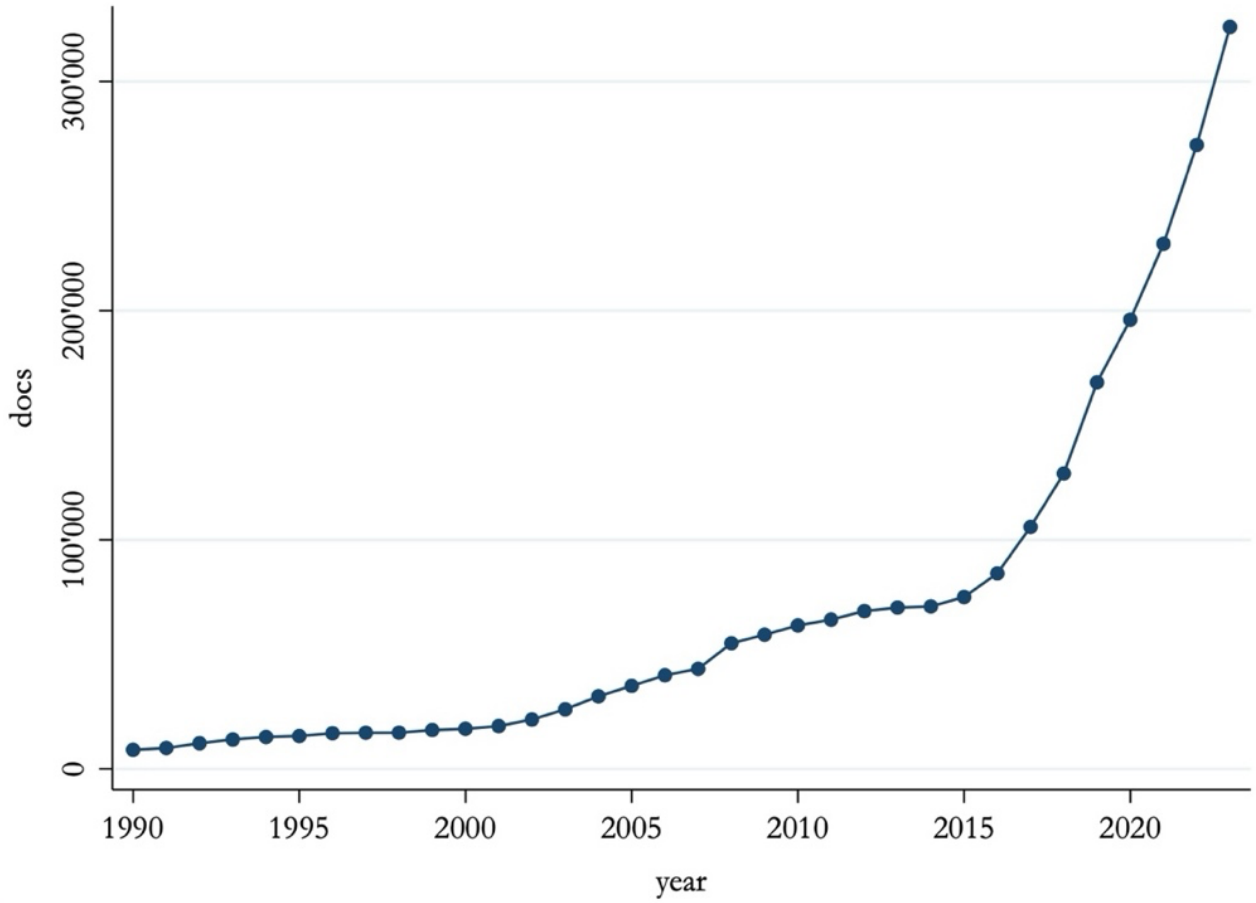
in their abstract, title, or authors' keywords, as detailed in Table 1. To minimize noise and ensure thematic focus, the search was restricted to core disciplinary areas that are foundational to AI research: namely, Computer Science, Engineering, and Mathematics (Bianchini et al., 2022). This filtering ensures that the resulting keyword list is both representative of the domain and sufficiently refined to support downstream analytical tasks.

Table 1. Initial list of keywords

artificial intelligence	facial recognition	robot
automatic classification	gesture recognition	self-driving
automatic control	knowledge representation	sentiment analysis
autonomous car	machine intelligence	speech recognition
autonomous vehicle	machine learning	statistical learning
bayesian model	natural language processing	supervised learning
computer vision	neural network	transfer learning
data mining	object detection	unmanned aerial vehicle
decision tree	predictive model	unmanned aircraft system
deep learning	probabilistic model	unsupervised learning
evolutionary computation	random forest	voice recognition
face recognition	reinforcement learning	

Approximately 2.4 million scientific publications were retrieved, published between 1990 and 2023. As illustrated in Figure 2, the annual volume of publications exhibits a pronounced inflationary trend in recent years. For instance, the cumulative number from the entire final decade of the 20th century is slightly less than one-quarter of that generated during the last three years of the observation period (2021–2023). This significant temporal imbalance necessitated dividing the textual analysis into distinct historical periods. Segmenting the analysis in this way serves three main purposes: (a) it facilitates a clearer understanding of the evolution of the AI-related technological-scientific paradigm by identifying the dominant terminologies in each period; (b) it allows for the detection of emerging and evolving technologies across distinct historical phases; and (c) it prevents the underrepresentation of foundational technologies and associated keywords from earlier periods, which might otherwise be overshadowed due to their comparatively lower publication volumes.

Figure 2. Retrieved scientific publications over time



The textual analysis was therefore divided into four distinct periods, or "waves": 1990–1999, 2000–2015, 2016–2020, and 2021–2023. For each wave, similar to Baruffaldi et al. (2020), we constructed a co-occurrence network of relevant keywords extracted from titles and abstracts using text mining techniques implemented through VOSviewer (van Eck & Waltman, 2023). This software enables co-occurrence analysis and the calculation of various network metrics that facilitate the identification of meaningful n-grams. To ensure relevance and comparability across periods, keywords were initially retained if their frequency exceeded a threshold defined for each wave, thereby controlling for differing publication volumes[21]. This filtering was subsequently refined using additional criteria, including total co-occurrences, the relevance score[22], and the co-occurrence strength between new candidate terms and those in the preliminary list. These steps ensured that the final set of keywords was not only relevant but also thematically coherent and representative of the underlying knowledge dynamics.

The final list was compiled by merging the results across all four waves and encompasses 221 unique n-grams. To ensure interpretability and avoid noise, we retained only non-ambiguous terms[23].

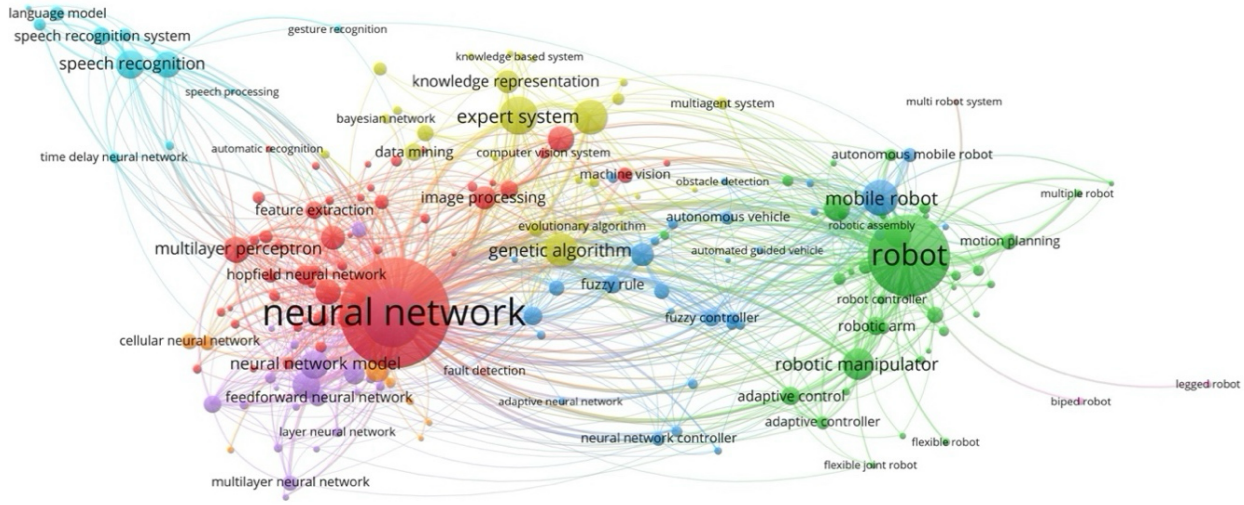
[21] As in Baruffaldi et al. (2020), we set the threshold to 100 occurrences in 1990-1999, 400 for the waves 2000-2015 and 2016-2020, and 500 for the wave 2021-2023. In this way, while accounting for the overall scientific production over the period, we analyze the evolution of the AI field over the last 30 years.

[22] It is computed by VOSviewer based on the centrality of the lemma within individual sentences and its overall prominence in the broader text. For more details regarding the metrics, see van Eck & Waltman (2009), van Eck et al. (2010), Waltman et al. (2010), and van Eck & Waltman (2014).

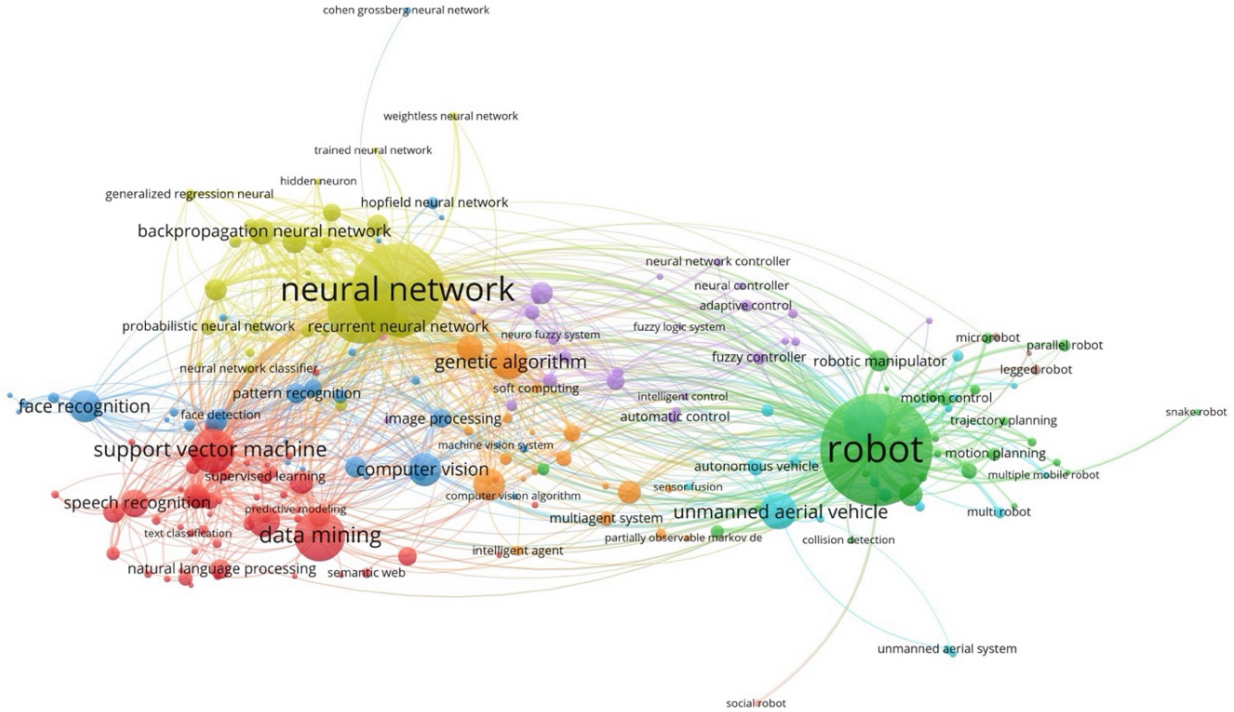
[23] In particular, most of the unigrams were not retained.

Figure 3. Keywords network by wave

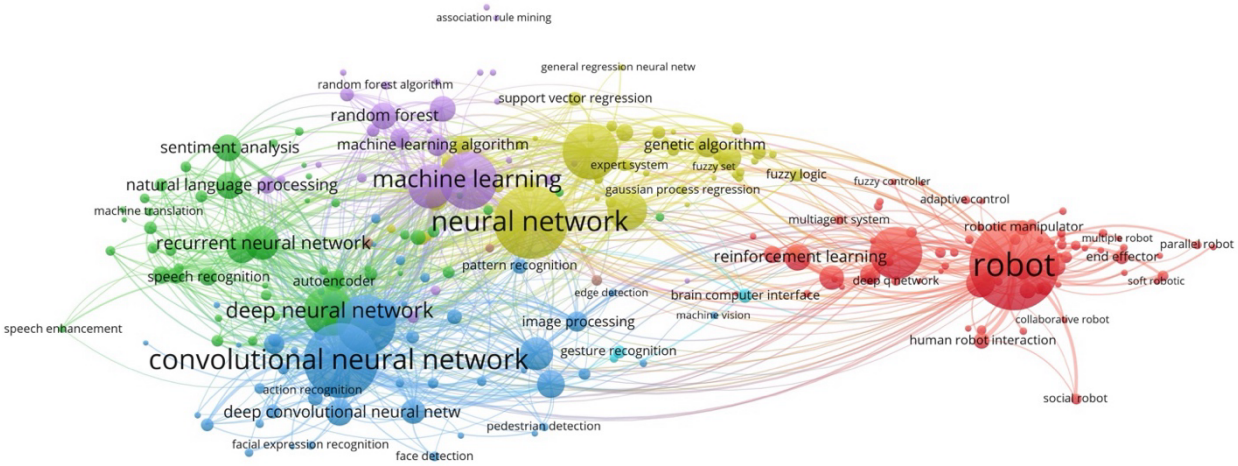
(a) Wave 1990–1999:



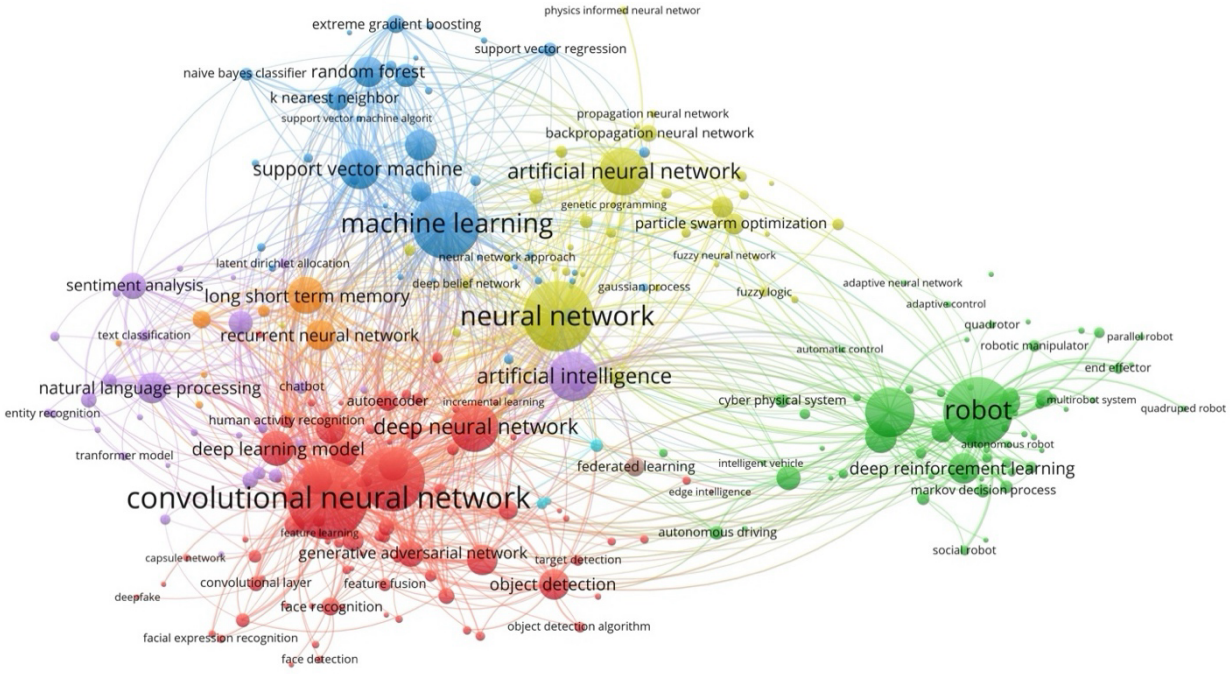
(b) Wave 2000–2015:



(c) Wave 2016–2020:



(d) Wave 2021–2023:



The network visualizations presented in Figure 3 delineate the scientific knowledge space for each temporal wave. In these graphs, each node represents a distinct n-gram extracted via VOSviewer, while edges denote co-occurrence relationships quantified by the association strength metric between keywords (van Eck & Waltman, 2009). Link thickness is proportional to the magnitude of this co-occurrence. Node placement is governed by the VOS (Visualization of Similarities) mapping algorithm, which spatially arranges items to preserve relational proximities and minimize layout distortion, thereby avoiding artificial circular structures (van Eck et al., 2010). Clusters of closely positioned nodes, therefore, reflect semantically coherent groupings within the AI knowledge space. Leveraging these network structures, we systematically assigned each n-gram to one of the two AI subdomains, LS-SYS or RA-SYS, based on its clustering context.

Each n-gram was subsequently attributed a weighting coefficient of either 0.5, 0.75, or 1.0. N-grams having a score of 1.0 denote higher domain specificity, and keywords scoring 0.5 indicate broader, more

general applicability. This scoring system helps distinguish core AI concepts from more peripheral or broadly applicable terms, enabling more precise classification and downstream analysis of patent content. These scores were assigned based on (a) an in-depth examination of the two AI domains; (b) an assessment of each term's structural role within the keyword co-occurrence networks; and (c) supplementary validation using GenAI tools, such as ChatGPT, to support semantic interpretation and contextual relevance[24]. The results for the two domains are shown in Tables 2 and 3, respectively.

Table 2. Final list of keywords - Learning and Symbolic Systems

LEARNING AND SYMBOLIC SYSTEMS					
Keywords	**Score**	**Keyword**	**Score**	**Keyword**	**Score**
action recognition	0,75	emotion detection	0,75	machine intelligence	0,75
active learning	0,75	emotion recognition	0,75	machine learning	1
activity recognition	0,75	ensemble classifier	0,75	markov random field	0,75
adaptive boosting	1	ensemble learning	1	meta learning	1
adaptive learning	0,75	entity recognition	0,75	model quantisation	1
adversarial attack	0,75	evolutionary algorithm	1	multi label classification	0,75
adversarial learning	1	evolutionary computation	1	multi layer perceptron	1
adversarial train*	1	evolutionary programming	1	multi task learning	1
apriori algorithm	0,75	expert system	1	natural language generation	1
artificial immune system	0,75	expression recognition	0,75	natural language processing	1
artificial intelligen*	1	extreme learning machine	1	neural architecture search	1
association rule learning	1	face recognition	0,75	neural classifier	1
association rule mining	0,75	feature engineering	1	neural controller	1
attention mechanism	1	feature extraction	0,75	neural machine translation	1
autoencoder	1	feature learning	1	neural net*	1
automatic classification	0,5	feature pyramid network	1	object classification	0,75
automatic detection	0,5	feature representation	0,75	object detection	0,75
automatic generation	0,5	federated learning	1	object recognition	0,75
automatic identification	0,5	feedforward network	1	opinion mining	1
automatic recognition	0,5	fuzzy inference system	0,75	pattern classification	0,75
automatic segmentation	0,5	fuzzy logic	0,75	pattern recognition	0,75
backpropagation	1	fuzzy system	0,5	predictive model	0,75
base classifier	0,75	gated recurrent unit	1	probabilistic model	0,75
base learner	0,75	gaussian mixture model	0,75	proximal policy optimization	1
batch normalisation	1	generative adversarial network	1	q learning	1
bayes classifier	0,75	generative model	0,75	radial basis function network	1
bayesian learning	0,75	generative pretrained transformer	1	random forest	0,75
bayesian model	0,75	genetic algorithm	1	regression tree	0,75

[24] The adoption of GenAI tools, such as ChatGPT, is increasingly gaining traction in economic research, as evidenced by recent studies (e.g., Jha et al., 2024; Eloundou et al., 2023; Davidsson & Sufyan, 2023; Korinek, 2023).

bayesian network	0,75	genetic network programming	1	reinforcement learning	1
beam search	0,5	genetic programming	1	representation learning	1
bidirectional associative memory	1	gesture recognition	0,75	restricted boltzmann machine	1
bidirectional encoder representations from transformers	1	gradient boosting	1	self organising map	1
bootstrap aggregation	1	gradient descent	1	semantic web	0,5
brain computer interface	0,5	graph attention network	1	sentiment analysis	1
capsule network	1	haar cascade	1	sentiment classification	1
cerebellar model arithmetic computer	1	haar classifier	1	shapley additive explanation	0,75
character recognition	0,5	hidden markov model	0,75	soft computing	0,75
chatbot	0,75	histogram of oriented gradient	0,75	speaker recognition	0,75
cluster analysis	0,5	image captioning	0,75	speech recognition	0,75
competitive learning	1	image classification	0,75	statistical learning	1
computational intelligence	0,75	image recognition	0,75	supervised learning	1
computational neuroscience	0,75	image segmentation	0,75	support vector machine	0,75
computer vision	1	imitation learning	1	support vector regression	0,75
conditional random field	0,75	inductive logic programming	1	swarm intelligence	1
connectionist temporal classification	1	intelligent agent	0,5	swin transformer	1
continual learning	0,75	k means	0,75	text classification	0,75
contrastive learning	1	k nearest neighbor	0,75	text mining	0,75
convolutional layer	1	knowledge based system	0,75	topic model	1
convolutional network	1	knowledge distillation	1	transfer learning	1
data mining	0,5	knowledge graph	0,75	transformer model	1
decision tree	0,75	knowledge process automation	0,75	unsupervised domain adaptation	1
deep belief network	1	knowledge representation	0,75	unsupervised learning	1
deep learning	1	large language model	1	vision transformer	1
deep q network	1	latent dirichlet allocation	1	voice recognition	0,75
dilated convolution	1	learning vector quantisation	1	web mining	0,75
edge detection	0,75	long short term memory	1	word embedding	1

Table 3. Final list of keywords - Robotics and Autonomous Systems

ROBOTICS AND AUTONOMOUS SYSTEMS					
Keywords	**Score**	**Keyword**	**Score**	**Keyword**	**Score**
active vision system	0,75	extended kalman filter	0,5	self driv*	0.75
adaptive control	0,5	fall detection	0,75	sensor data fusion	0,75
automated optical inspection	0,75	fault classification	0,75	sensor fusion	0,75
automatic control	0,5	fault detection	0,75	simultaneous localization mapping	0,75
automatic target detection	0,75	forward kinematic	1	trajectory planning	0,75
automatic target recognition	0,75	intelligent vehicle	0,75	trajectory prediction	0,75
autonomous aerial vehicle	1	inverse kinematic	1	unmanned aerial system	1
autonomous car	1	machine vision	0,75	unmanned aerial vehicle	1
autonomous driving	1	manipulator	0,75	unmanned aircraft system	1
autonomous ground vehicle	1	motion planning	0,75	unmanned aircraft vehicle	1
autonomous system	0,5	multiagent system	0,5	unmanned ground vehicle	1
autonomous underwater vehicle	1	obstacle avoidance	0,75	unmanned surface vehicle	1
autonomous vehicle	1	pedestrian detection	0,75	unmanned underwater vehicle	1
central pattern generator	0,5	pid controller	0,75	unmanned vehicle	1
closed loop control system	0,5	quadcopter	1	vehicle detection	0,75
collision avoidance	0,75	quadrotor	1	vehicular ad hoc network	0,75
collision detection	0,75	rapidly exploring random tree	0,75	visual servoing	1
end effector	0,75	robot	1		

4.2. The patent universe

The Patent Universe comprises all patent families with at least one application filed at either the European Patent Office, World Intellectual Property Organization, or United States Patent and Trademark Office, and published between 1980 and 2021[25]. The dataset was constructed using the PATSTAT database and includes approximately 27 million patent applications, corresponding to nearly 14 million DOCDB[26] patent families. Given data availability constraints and the objective of ensuring cross-office comparability, our machine learning–based analysis relies exclusively on the information contained in patent titles and abstracts. The patent universe was used to (a) retrieve LS-SYS and RA-SYS seed sets for training the two machine learning models, (b) perform the patent landscaping procedure, leading to identifying anti-seed sets for the two neural networks, and (c) conduct the inference on patents included in the expansion.

[25] This scope is chosen to mitigate data-availability limitations for inventors and applicants.
[26] The DOCDB ("Documentation Database") patent family groups all documents that share exactly the same priority or combination of priorities (i.e., they originate from the same original filing). Since applications within the same DOCDB families claim the same "active" priority, they are considered to have the same technological content (Martinez, 2010).

4.3. Seed sets construction

We extracted LS-SYS and RA-SYS patent applications from the Patent Universe by mining titles and abstracts, employing the two domain-specific keyword lists. We define the seed-set using a semi-automatic approach akin to Giczy et al. (2022), replacing their classification codes-based filtering with our weighted keyword lists and bespoke heuristics. As emphasized by Abood and Feltenberger (2018), the construction of seed sets must ensure both representativeness and accuracy, as their quality directly impacts the performance of machine learning models in patent landscaping tasks. First, to ensure representativeness, seed sets must capture the full diversity of subdomains within each technological area. If specific subfields are not included, the model is unlikely to learn their characteristics during training, thereby reducing its ability to recognize them during inference. The keyword list expansion described in Subsection 4.1 (resulting in 221 n-grams) was deliberately performed to mitigate this risk by furnishing a lexicon of sufficient breadth to encompass the entire spectrum of AI-related topics that have emerged over the past forty years. This expansion strategy helps reduce the risk of under-representativeness and improves the inclusiveness of the seed sets. Second, accuracy is essential to avoid error propagation throughout the machine learning pipeline. Training a model on patent documents retrieved solely through basic keyword matching can lead to Type I errors. To mitigate false positives, we implemented a fully reproducible, two-pronged filtering strategy. First, we develop a weighted scoring system in which each candidate patent receives a composite score based on the keywords' specificity and frequency in its title or abstract. In parallel, we also apply human-inspired heuristics to exclude ambiguous or marginally related documents, enabling us to compile high-quality seed sets for both the LS-SYS and RA-SYS domains that balance breadth with precision. Accordingly, the seed sets used to train the neural network consist of patents that satisfy heuristic (a) and one criterion between (b) or (c):

(a) The presence of at least one "tier-one" keyword, namely terms assigned the highest relevance score of 1.

(b) Patents having a final score of at least 2, combined with a keyword occurrence count exceeding the number of distinct matched keywords.

(c) Patents having a final score of less than 2, combined with a keyword occurrence count exceeding the number of distinct matched keywords by at least three.

These selection criteria were designed to balance precision and inclusivity in constructing training seed sets. The inclusion of patents containing at least one tier-one keyword ensures that the most central and unambiguous AI concepts anchor the dataset, providing a strong semantic foundation for the model. Moreover, these tier-one terms serve as markers of AI relevance—terms to which human annotators would instinctively attach strong semantic meaning when assessing a patent's AI content. We combined these tier-one keywords with a broader but consistent presence of AI-related terminology, measured through a combined score of at least 2 and a higher-than-expected number of keyword occurrences relative to unique terms. This criterion prioritizes conceptual depth and reduces noise from superficial mentions. Finally, to avoid overlooking patents that may be narrowly focused but still meaningful, we also include patents with lower overall scores if they demonstrate concentrated use of key terms. This ensures the inclusion of technically focused documents that could otherwise be filtered out by more rigid thresholds. In particular, for lower-scoring patents, we applied a

stricter threshold, requiring a higher degree of keyword density to ensure only those with a clear technical signal were included. This two-tiered programmatic labeling function allowed us to preserve high-confidence AI patents while minimizing noise and improving the overall representativeness of the seed sets used for model training[27]. Indeed, by focusing our labeling strategy on patent titles and abstracts, we capitalize on the sections where inventors distill and define their core inventive contributions and claimed subject matter.

Applying the heuristics outlined above in conjunction with the keyword lists reported in Tables 2 and 3, we identified 43,974 and 67,214 patent documents, which represent potential candidate entries for the LS-SYS and RA-SYS seed sets, respectively. To ensure the integrity of both seed sets, we retained unique, non-duplicate patent documents, which might introduce the risk of data leakage during training[28]. To systematically detect and remove duplicates or near-duplicates, we concatenated each patent's title and abstract into a single text string and applied standard natural language preprocessing steps, including stop-word removal and stemming. We then vectorized the documents using Term Frequency–Inverse Document Frequency (TF-IDF) weighting, following the approach in Lamperti (2024). To quantify similarity across documents, we computed pairwise cosine similarity scores, treating the dataset as a fully connected undirected graph, where nodes represent patents and edges encode their textual similarity. We imposed a similarity threshold of 0.5, such that if any two patents exceeded this cosine similarity, one was randomly removed from the seed set. This deduplication procedure ensured a diverse and representative sample of patent texts, minimized the risk of information leakage across data splits during the training and testing phases, and enhanced the robustness and generalizability of the downstream machine learning model. Finally, by leveraging CPC codes and keyword co-occurrence patterns, we checked for the presence of ambiguous or false-positive patents in both seed sets, which were then discarded. This procedure resulted in the construction of the final seed sets. The LS-SYS seed set ultimately comprises 10,392 unique patent filings, corresponding to 10,251 DOCDB families; the RA-SYS seed set comprises 11,387 filings across 10,924 families[29]. Table 4 reports some examples.

As a result, we are able to construct extensive and scalable seed sets that comprehensively encompass AI's multifaceted applications and developmental pathways (i.e., a typical byproduct of general-purpose technologies), which further mitigates the downward bias that would otherwise stem from fine-tuning on a limited number of positive instances.

[27] The criterion (a) is, obviously, the most penalizing. In this respect, we focus on patents where inventors explicitly disclose core AI technologies or techniques (i.e., tier one keywords). We might overlook inventions that describe AI-related technologies without the use of these keywords. The impact of this aspect, however, should be minimized during training and inference, for two key points: the number of "positive examples" is large enough to ensure a wide vocabulary heterogeneity, which, in turn, should allow the models to learn the characteristics of AI-related patents and generalize to other patents not containing "tier-one" keywords. This aspect is further investigated in Tables A3 and A4, where we show that the models are also able to classify as AI-related patents that do not explicitly contain tier-one keywords.

[28] A situation where metrics are inflated due to the presence of semantically equivalent samples in the training and test sets.

[29] This explains why, in most cases, we retain only one patent per DOCDB family in the seed set. Patent documents within the same family often share highly similar, if not identical, titles and abstracts, which increases the risk of redundancy and data leakage. As a result, we preserve a single representative application per family, except in rare instances where two or more documents within the same DOCDB family exhibit sufficiently distinct semantic structures to warrant separate inclusion.

Table 4. LS-SYS and RA-SYS seed sets examples

Application ID	DOCDB family ID	Title	Abstract	Keywords	Count	Score	Set
449665221	55229745	Sensor noise profile	The invention relates to **feature extraction** technique based on edge extraction. It can be used in **computer vision** systems, including image/facial/**object recognition** systems, scene interpretation and classification and captioning systems. It uses a model or profile of the noise in the sensor to improve **feature extraction** or **object detection** on an image from a sensor which may be linear data in the RAW domain. The model may be used to normalise **feature extraction** response. The extraction/detection may be based on **edge detection** possibly involving convolution of the edge response with a filter kernel e.g. a Gabor or CNN filter, a linear classifier such as a **support vector machine** (SVM) or the classification layer of a convolutional **neural network** (CNN).	7	9	5,75	LS
517889032	67541812	System and method for interfacing with biological tissue	There is provided a system and method for interfacing with biological tissue. The system includes: a **feature extraction** module to implement an extraction approach to extract one or more features from the one or more physiological recording signals; a **machine learning** module to apply a **machine learning** model based on input data to detect a physiological event or condition for classification, the input data including the extracted features, the **machine learning** model trained using a training set including feature vectors of time-series data labelled with known occurrences of the physiological event or condition; and an output module to output the classification of the **machine learning** module.	2	5	1,75	LS
331630746	43535428	Method and apparatus to plan motion path of **robot**	If a **manipulator** of a **robot** falls in local minima when expanding a node to generate a path, the **manipulator** may efficiently escape from local minima by any one of a random escaping method and a goal function changing method or a combination thereof to generate the path. When the solution of **inverse kinematics** is not obtained due to local minima or when the solution of **inverse kinematics** is not obtained due to an inaccurate goal function, an optimal motion path to avoid an obstacle may be efficiently searched for. The speed to obtain the solution may be increased and thus the time consumed to search for the optimal motion path may be shortened.	3	6	2,75	RA
456772040	588813834	Flow-based **motion planning** blueprint for **autonomous vehicles**	A system includes a memory device, and a processing device, operatively coupled to the memory device, to receive a set of input data including a representation of a drivable space for an **autonomous vehicle** (AV), generate, based on the representation of the drivable space, a **motion planning** blueprint from a flow field modeling the drivable space, and identify, using the **motion planning** blueprint, a driving path of the AV within the drivable space.	2	5	1,75	RA

4.4. Patent landscaping and Anti-Seed sets

The two seed sets presented above will serve as the foundation for the LS-SYS and RA-SYS landscapes. As thoroughly described by Abood & Feltenberger (2018), the Automated Patent Landscaping represents the process of finding patents related to particular topics, leveraging modern machine learning methodologies as well as patent metadata. More deeply, the human-curated seed sets are expanded using their CPC codes and family citations to identify a group of "probably related" patents, which are afterwards pruned by a machine learning model trained on the seed (positive examples) and anti-seed sets (negative examples, composed of patents that are not included in the expansion).

Consistent with established practices in the literature (Pairolero et al., 2025; Giczy et al., 2022; Abood and Feltenberger, 2018), we conducted two rounds of expansion for each technological landscape. The first, referred to as Level 1 Expansion, encompasses the following categories: (1) all patents, as well as their respective DOCDB family members, retrieved using the two keyword lists but excluded from the seed sets due to not meeting the threshold criterion; (2) all patents belonging to a DOCDB patent family that either cites or is cited by patent families included in the seed sets (i.e., forward and backward family citations); (3) all patents pertaining to DOCDB families assigned to relevant CPC codes[30]. The relevance of CPC codes is evaluated within each seed set. A CPC code is considered relevant if: (i) it appears in at least 0.5% of the patent families within the corresponding seed set (as in Choi et al., 2022); and (ii) the share of patent families containing the CPC code within the seed set relative to its share in the overall patent universe is greater than 50. This criterion ensures that only technology classes disproportionately represented in the seed sets are included in the expansion[31]. We report in the appendix the lists of CPC codes used (Tables A1 and A2), which in turn emphasize the main building blocks of both seed sets as well as the quality of our training data. Level 2 Expansion is designed to capture all backward and forward family citations of patent families included in the Level 1 Expansion. As noted by Abood and Feltenberger (2018), it is expected that the machine learning models will retain a higher proportion of Level 1 patents while pruning a larger share of Level 2 candidates. This is because patents included in Level 1 are, by construction, more proximally related to the core seed sets, whereas those in Level 2 represent more peripheral connections.

All patents not captured through the two-level expansion process are classified as part of the anti-seed group, serving as negative examples in the training process. From this group, we randomly sample 35,000 unique, non-duplicate patent documents to construct a representative set of negative training instances. This sampling strategy ensures a large volume of reasonably accurate negative examples, positioned at a clear semantic distance from the core seed sets. By design, the positive (seed) and negative (anti-seed) examples are intentionally drawn from opposite ends of the relevance spectrum. As such, patents with intermediate similarity, those falling in the "gray area" between clearly relevant and irrelevant, are underrepresented. While this approach leads to training a machine learning model to filter out documents that clearly differ from the seed

[30] At the subgroup level.
[31] This computation is performed at the level of DOCDB patent families, leveraging the fact that all applications within the same family share identical CPC code in PATSTAT, as in TLS225_DOCDB_FAM_CPC. This approach prevents the inflation of CPC frequency counts that could arise from counting multiple applications of the same invention (Martinez, 2010).

sets, it also has a key limitation: the binary sampling method can introduce bias and make the model less sensitive to borderline or ambiguous cases[32].

Table 5 shows some descriptive statistics regarding the two landscapes. We note that the two lists of keywords produced two seed sets and expansions that are quite similar in terms of numerosity. The two machine learning models presented in the next section will be used to prune irrelevant patents from Level 1 and Level 2 expansions.

Table 5. LS-SYS and RA-SYS landscapes

	LS-SYS	**RA-SYS**
Patent applications – seed set	10,392	11,387
Patent families – seed set	10,251	10,924
Patent applications – L1 expansion	477,293	443,117
Patent families – L1 expansion	267,156	214,397
Patent applications – L2 expansion	3,586,480	4,189,234
Patent families – L2 expansion	1,967,637	2,200,632
Patent applications – Anti-seed set	35,000	35,000
Patent families – Anti-seed set	35,000	35,000

4.5. BERT for Patents and fine-tuning

As previously noted, we leverage the two landscapes to train two separate machine learning models aimed at predicting whether patents included in the Level 1 and Level 2 expansions are related to the LS-SYS or RA-SYS domains. In line with prior research (Pairolero et al., 2025; Choi et al., 2022; Giczy et al., 2022; Alderucci et al., 2020; Abood and Feltenberger, 2018), a wide array of classification algorithms has been proposed to identify AI-related patents. These range from traditional machine learning models—such as Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machines (SVM)—to more advanced deep learning architectures, including Long Short-Term Memory networks (LSTM), Hierarchical Attention Networks (HAN), and Transformer-based models.

Transformer-based models have become the preferred choice for classification tasks due to their ability to capture complex linguistic patterns and contextual relationships within texts (Vaswani et al., 2017). Among these, BERT (Bidirectional Encoder Representations from Transformers) stands out for its bidirectional training approach, making it especially effective at handling the meaning of words in different situations (Devlin et al., 2018). As highlighted by Devlin et al. (2018), the massive pre-training provided to BERT[33]

[32] We acknowledge that such a limitation is addressed by Pairolero et al. (2025).
[33] The authors used BooksCorpus (800M words) and English Wikipedia (2,500M words) and trained the model on two unsupervised tasks: Masked Language Modelling (aimed at identifying a masked token in a string, given a specific context) and Next Sentence Prediction (which constitute a binary classification task where the model receives two sentences and predicts whether the second one

enables the model to leverage general language understanding, streamlining its fine-tuning for specific tasks, such as patent classification, with minimal architectural adjustments.

This is particularly beneficial in the patent domain, where documents often contain intricate and technical language (Chung & Sohn, 2020). Empirical studies have demonstrated BERT's effectiveness in patent classification tasks. For instance, Lee & Hsiang (2020) fine-tuned a pre-trained BERT model on a corpus of over two million USPTO patents to predict classification codes of each filing. Without altering the core architecture and by feeding the model patent-claim sections, they surpassed previous state-of-the-art approaches, such as CNNs with static word embeddings, in both precision and recall. In particular, the authors were able to achieve accurate classification results by implementing minimal changes to the original model[34]. This minimal-adjustment strategy highlights BERT's capacity to internalize fine-grained semantic and syntactic patterns within patents, enabling highly accurate downstream classification with relatively little task-specific data.

In their 2020 white paper, Srebrovic & Yonamine present a pioneering application of the BERT model tailored specifically for patent-related domains. Recognizing the unique linguistic characteristics and complexity of patent documents, they trained a BERT model exclusively on a vast corpus of patent texts, encompassing over 100 million publications from the U.S. and other countries[35]. This domain-tailored pre-training provides the model with a nuanced understanding of the highly technical, often idiosyncratic language employed in patents, terminology that diverges substantially from general-purpose corpora like Wikipedia, used in the pre-training phase of BERT (Devlin et al., 2018). Indeed, the pre-training on patent texts significantly boosts downstream patent classification performance compared to generic BERT, given its exposure to the full range of patent terminology and phrasing. Furthermore, Srebrovic & Yonamine (2020) also developed a custom tokenizer optimized for patent text. Traditional tokenizers, trained on general corpora, often struggle with the specialized terminology and lengthy compound words found in patent filings[36]; by creating a tokenizer attuned to patent-specific language patterns, the authors improved the model's ability to process and understand complex documents effectively.

To perform our binary classification task, we leveraged the extensive training of the BERT for Patents model[37] by creating two fine-tuned models, one per domain. Our fine-tuning strategy is minimal and leverages the Hugging Face Trainer API. Table 6 provides some descriptive statistics on the training samples used for the fine-tuning, along with the train-validation-test splits[38].

---

is the actual next sequence in a text or a random sentence from the corpus). The training did not involve humans' supervision and labelling.

[34] Lee & Hsiang (2020) perform a multi-label classification task, intending to assign patents to one or more CPC codes. Therefore, the authors leveraged the pre-trained BERT model and only changed the output layer by substituting the original softmax function (suitable for one-hot classification only) with a sigmoid cross-entropy with logits function.

[35] The Bert for Patents model is based on BERT large.

[36] To explain this aspect, the authors refer to an example: while BERT's tokenizer would split the word "prosthesis" into <pro><thes><is> tokens, BERT for Patents tokenizer will keep <prosthesis> as a single token given its optimization. Such an improvement significantly boosts predictive accuracy.

[37] The model, along with its checkpoints, is available on the Hugging Face Hub at the following link: https://huggingface.co/anferico/bert-for-patents.

[38] Following a 70:15:15 ratio.

Table 6. LS-SYS and RA-SYS training samples

	Positive examples	Negative examples
	LS-SYS	
Train	7,274	24,500
Validation	1,559	5,250
Test	1,559	5,250
	RA-SYS	
Train	7,971	24,999
Validation	1,708	5,250
Test	1,708	5,251

In preparing patent documents for the fine-tuning phase, we deliberately eschewed conventional text preprocessing techniques, such as stopword removal, punctuation stripping, and stemming or lemmatization. These traditional methods can disrupt the syntactic and semantic integrity of the text, potentially impairing the model's ability to capture subtle language patterns. Instead, we utilized the pre-trained BERT for Patents tokenizer, ensuring that the unique terminology and structure inherent in patent documents are effectively tokenized. Specifically, we concatenated the title and abstract of each patent, inserting the special token [SEP] between them. Such a configuration enables the model to process the title and abstract as separate yet contextually linked sequences, facilitating a more comprehensive understanding of the patent's content.

Following the approach of Devlin et al. (2018), we implemented a binary classification head atop the pre-trained BERT for Patents architecture to address our domain-specific classification task. Given the relatively limited size of our labeled training sets, we adopted a feature-based transfer learning strategy by freezing all layers of the base BERT model and fine-tuning only the classification head. This design choice mitigates common risks associated with fine-tuning large language models on small datasets, including overfitting, where the model captures noise or idiosyncrasies in the training data, and catastrophic forgetting, whereby updates to the full model may overwrite valuable representations learned during pre-training. By keeping the base model fixed, we retain its rich contextual language representations, enhance training efficiency by reducing the number of trainable parameters, and promote generalization to unseen data. This strategy also contributes to greater training stability, particularly in low-data regimes. We trained the classification head for a maximum of 10 epochs, implementing early stopping to prevent overfitting and ensure convergence once validation performance ceased to improve.

Each fine-tuned model produces a pair of logits representing unnormalized scores for two classes: class 0 (not related to either LS-SYS or RA-SYS) and class 1 (related to either LS-SYS or RA-SYS). These logits are passed through a softmax function, which transforms them into a probability distribution over the two classes. As a result, for each input example, the model outputs two probabilities that sum to one, indicating the model's confidence that the input belongs to either the negative or positive class. The predicted label is then determined by selecting the class with the highest probability. Consequently, this prediction process is equivalent to

applying a decision threshold of 50% on the probability of the positive class: if the probability of the positive class exceeds 0.5, it will be assigned class 1 (positive), and if it is less than or equal to 0.5, it will be assigned class 0 (negative).

5. Training and Inference Phase

In this section, we first present the key training and validation metrics for each fine-tuned model, evaluating their classification performance on held-out test sets. We then outline the inference procedures applied to the Level 1 and Level 2 expansion corpora and summarize the resulting classification outcomes.

5.1. Training and Testing Statistics

Table 7 presents the training and testing performance metrics for both domain-specific classifiers. Across all key indicators, each model demonstrates a robust ability to distinguish LS-SYS and RA-SYS patents from non-AI filings. Notably, both models achieve F1 scores exceeding 0.90 on their respective test sets.

Table 7. Training and Testing statistics for LS-SYS and RA-SYS fine-tuned models

	T. Loss	V. Loss	Accuracy	F1 score	Precision	Recall
			TRAINING			
LS-SYS	0.0647	0.0528	0.981	0.958	0.962	0.955
RA-SYS	0.0794	0.0668	0.975	0.947	0.96	0.935
			TESTING			
LS-SYS	–	–	0.981	0.958	0.965	0.95
RA-SYS	–	–	0.976	0.951	0.964	0.937

These high-performance scores are consistent with expectations and should not be considered surprising. As noted by Abood and Feltenberger (2018), elevated classification metrics are a common feature in automated patent landscaping tasks largely because the training data contains few borderline positive or negative examples. This design choice, favoring clearly relevant and irrelevant examples, enhances model precision and helps the classification system effectively remove patents that do not closely match the core seed topic. On the other hand, this approach may also introduce a limitation: during inference, the model may underestimate the full scope of the target domain by overlooking patents that are only moderately related.

Notably, the F1 score remains consistently high across both the training and testing phases for both models, indicating that the classifiers do not suffer from overfitting. Unlike other performance metrics, the F1 score is computed for the minority class (namely, the positive examples) and is particularly informative in the context of class imbalance. The strong F1 performance suggests that the models do not default to predicting the majority class, thereby avoiding a common pitfall in imbalanced classification tasks. Interestingly, the F1

scores obtained in our analysis are broadly consistent with those reported by Abood and Feltenberger (2018), who developed four distinct patent landscapes and trained separate classifiers for each.[39]

Tables 8 and 9 depict the confusion matrices for LS-SYS and RA-SYS classifications, showing strong overall performance, with high precision and recall in both cases. For LS-SYS, the model correctly identifies 1,481 true positives with only 53 false positives and 78 false negatives. Similarly, the RA-SYS model achieves 1,601 true positives, 59 false positives, and 107 false negatives. Notably, in both models, the number of false positives is lower than the number of false negatives, which is a preferable outcome in many classification tasks, especially when false alarms are more costly or disruptive than missing some true positives. Precision remains very high in both cases (around 96.5%), while recall is slightly lower for RA-SYS (93.7% vs. 95.0% for LS-SYS), indicating the models are conservative but reliable in assigning positive labels. This outcome suggests that the implicit 50% threshold used via argmax is well-calibrated for prioritizing precision without excessively compromising recall.

Table 8. LS-SYS confusion matrix

	Predicted Label		
	NOT LS-SYS	**LS-SYS**	**TOTAL**
NOT LS-SYS	5,197 (99%)	53 (1%)	5,250 (100%)
LS-SYS	78 (5%)	1,481 (95%)	1,559 (100%)

Table 9. RA-SYS confusion matrix

	Predicted Label		
	NOT RA-SYS	**RA-SYS**	**TOTAL**
NOT RA-SYS	5,192 (98,9%)	59 (1,1%)	5,251 (100%)
RA-SYS	107 (6,3%)	1601 (93,7%)	1,708 (100%)

Our methodology's design and performance metrics demonstrate that the classifiers employ a deliberately conservative decision boundary. By prioritizing the reduction of Type I errors (false positives), these models ensure that downstream analyses are conducted on a highly reliable set of domain-specific patents, even if this means excluding a small number of borderline cases. While this strategy may overlook some "grey area" documents whose semantic profiles fall outside the training distribution, it preserves the rigor of both seed-set expansion and the overall patent landscaping process. Moreover, by maintaining only unambiguous AI patents

[39] Looking at the reported results for LSTM models, they obtain the following F1 scores: 0.987 for the "browser" topic, 0.965 for the "operating system" topic, 0.991 for the "video codec" topic, and 0.982 for the "machine learning" topic.

in our seed and pruned patent sets, we substantially reduce noise in subsequent trend and network analyses. This focus on purity ensures that longitudinal studies of AI innovation accurately reflect genuine domain developments.

Upon completing the fine-tuning process, we saved the model checkpoints and uploaded the LS-SYS and RA-SYS classifiers to the Hugging Face Model Hub[40]. Each model repository includes comprehensive documentation, such as hyperparameter settings, evaluation metrics, and detailed training statistics. These publicly available models can be accessed through the Transformers library[41], allowing researchers and practitioners to easily reuse, evaluate, or adapt them for related tasks in patent analysis or AI landscape mapping.

5.2. Inference phase and examples

Finally, we turn to the inference stage, during which the fine-tuned models are employed to assess the degree of relatedness between patents in the expansions and those in the corresponding seed sets. In this context, the LS-SYS model is applied to predict on 477,293 patent applications in its Level 1 expansion and 3,586,480 in Level 2. Similarly, the RA-SYS model is used to classify 443,117 applications in Level 1 and 4,189,234 in Level 2. For the sake of the present analysis, we do not apply our models to patents residing outside the two expansions. Table 10 reports the number of patent applications classified as "positive" by the machine learning models across expansions.

Table 10. Results of the inference phase

	Level 1 expansion	Level 2 expansion
LS-SYS	214,792 (46%)	171,362 (4,8%)
RA-SYS	173,757 (40%)	179,769 (4,3%)

Overall, both classifiers exhibit consistent retention patterns, underscoring the robustness of our landscaping procedure: Level 1 expansions are characterized by substantially higher retention rates than Level 2 expansions. For the LS-SYS domain, 46 % of Level 1 filings are retained versus only 4.8 % from Level 2; the RA-SYS model shows similar rates: 40 % retention for Level 1 and 4.3 % for Level 2.

Interestingly, our results closely mirror those reported by Abood and Feltenberger (2018). This outcome is consistent with the structural logic of the landscape: by design, patents in Level 1 are expected to be more closely related (both technologically and semantically) to those in the seed sets, and therefore more likely to be retained by the classification models. In the Appendix, Table A3 presents illustrative patent examples, while

[40] The LS-SYS model can be accessed at https://huggingface.co/Fradalessandro/bert-for-patents-finetuned_ls-sys and the RA-SYS model at https://huggingface.co/Fradalessandro/bert-for-patents-finetuned_r-as

[41] In particular, after importing the Transformer library, users can load the pretrained tokenizer and classifier via the following code: AutoTokenizer.from_pretrained("repo") and AutoModelForSequenceClassification.from_pretrained("repo").

Table A4 quantifies the proportion of AI-related documents containing our listed keywords. Crucially, a substantial share of the patents flagged by our classifiers as AI-related contain none of these keywords, demonstrating the models' ability to generalize beyond simple lexical matches.

6. Results and descriptive analysis

In the present Section, we offer a detailed empirical analysis in which we further validate the results of our fine-tuned classifiers by benchmarking their outputs against prior literature reviewed in Subsection 2.2. We start by tracing the evolution of AI patenting over the past forty years and assessing the countries most active in AI patent production. We then analyze the sectors leading the AI race, distinguishing the domain-specific contributions of LS-SYS and RA-SYS innovations and juxtaposing our findings with established sectoral patterns. Building on the sectoral results, we map inter-sector collaboration networks in AI patents, identifying the key knowledge hubs. Furthermore, we rank the most prolific applicants in each domain, validating that our classifiers capture the principal firms driving AI innovation. Finally, we employ co-classification analyses to depict the technological space of AI patents, illustrating the main technological building blocks LS-SYS and RA-SYS subfields.

To avoid inflationary effects due to multiple applications protecting the same invention, we conduct the analysis at the DOCDB patent family level. A family is considered related to each AI domain if it includes at least one patent application flagged as AI-related by the corresponding machine learning model. This approach yields a set of AI-related DOCDB families, which may be associated with the LS-SYS domain, the RA-SYS domain, or both.

6.1. The evolution of AI patenting across time and space

In total, we identified 479,216 AI-related patent families having a priority date between 1980 and 2021[42]. Figure 4 displays the evolution of AI DOCDB patent families over time, along with their breakdown into the LS-SYS and RA-SYS domains. The data reveal a steady and substantial increase in AI-related patenting activity from the early 2000s onward, with a particularly sharp acceleration observed after 2010. The volume of LS-SYS-related families consistently outpaces that of RA-SYS from the period 2013-2015, following the Deep Learning revolution (Sejnowski, 2018; Souza et al., 2024). Overall, the emerging picture is consistent with those provided by previous studies (Alderucci et al., 2020; Van Roy et al, 2020; Dernis et al., 2021; Miric et al., 2023). The declining trend over the last two years may be related to the COVID-19 pandemic; however, we argue that such dynamics are primarily driven by publication lags and administrative processing delays in patent filings (Dass et al., 2017), which can lead to data truncation and incomplete coverage in the PATSTAT database[43].

[42] For each DOCDB patent family, we use the earliest filing year as a reference date.
[43] Considering the canonical 18-month lag from application to publication, additional administrative processing and indexing delays, the decline in counts for 2021 primarily reflects data truncation rather than a true reduction in patenting activity.

Figure 4. AI-related DOCDB patent families over time

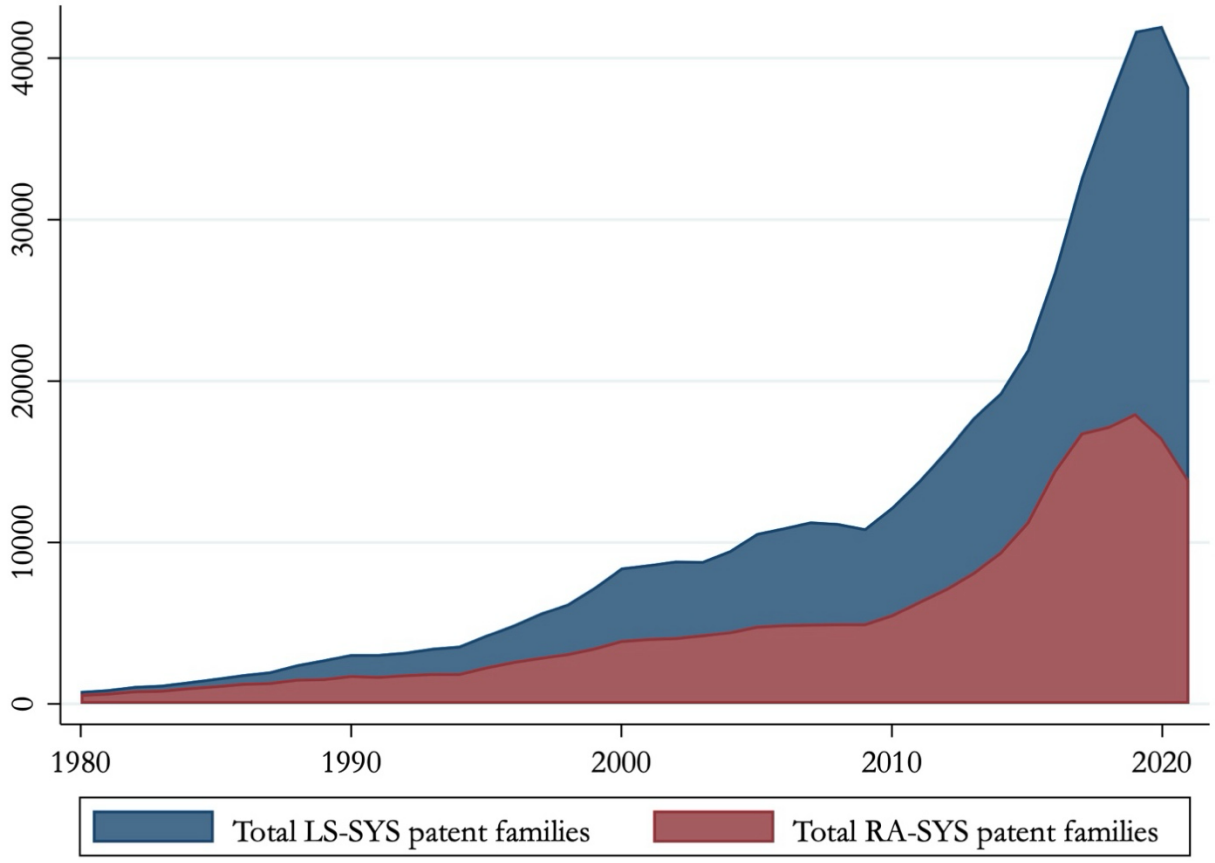


Figure 5 deepens the evolution over time of patent families classified as related to both LS-SYS and RA-SYS domains. The intersection of these two domains comprises innovations at the nexus of "thinking" and "acting," such as humanoid robots, next-generation autonomous vehicles (drones, self-driving cars, even submersibles), and intelligent autonomous systems in general. Prior to 2010, these overlapping families grew modestly; interestingly, beginning in 2013–2014, we observe a sharp inflection, with the absolute count of dual-domain families more than tripling within five years. This breakpoint coincides with the wider recognition and deployment of deep-learning breakthroughs, most notably AlexNet's 2012 ImageNet victory and the subsequent proliferation of convolutional and reinforcement-learning architectures in real-world systems (Sejnowski, 2018). As deep neural networks became both more accurate and computationally tractable, robotics researchers were able to integrate end-to-end perception, planning, and control pipelines that were previously infeasible. The steep post-2013 rise therefore reflects not just incremental improvements in individual AI subfields, but a genuine technological fusion: LS-SYS methods (e.g., deep vision and natural-language understanding) began to be embedded directly into RA-SYS platforms, giving rise to genuinely "intelligent" machines capable of contextual awareness, adaptive decision-making, and autonomous operation.

Figure 5. AI-related DOCDB patent families over time – domain overlap

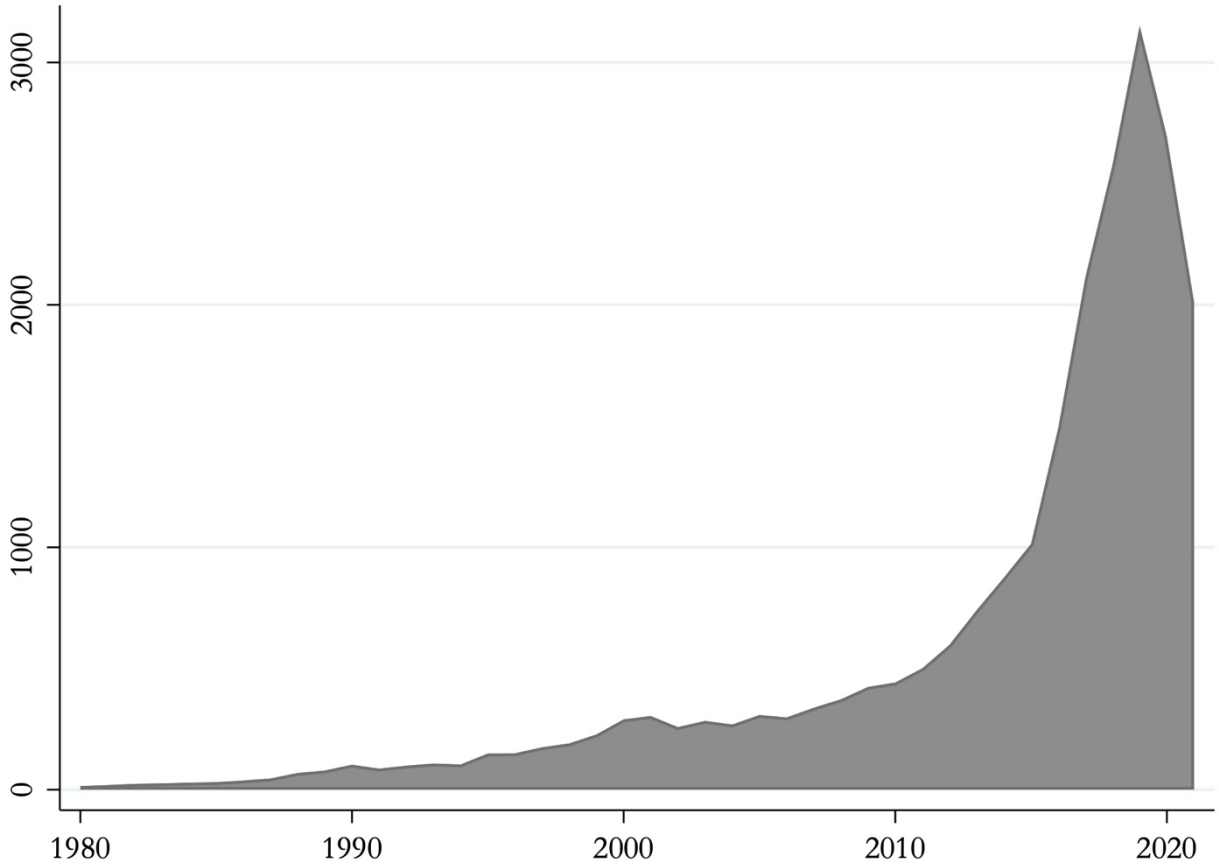
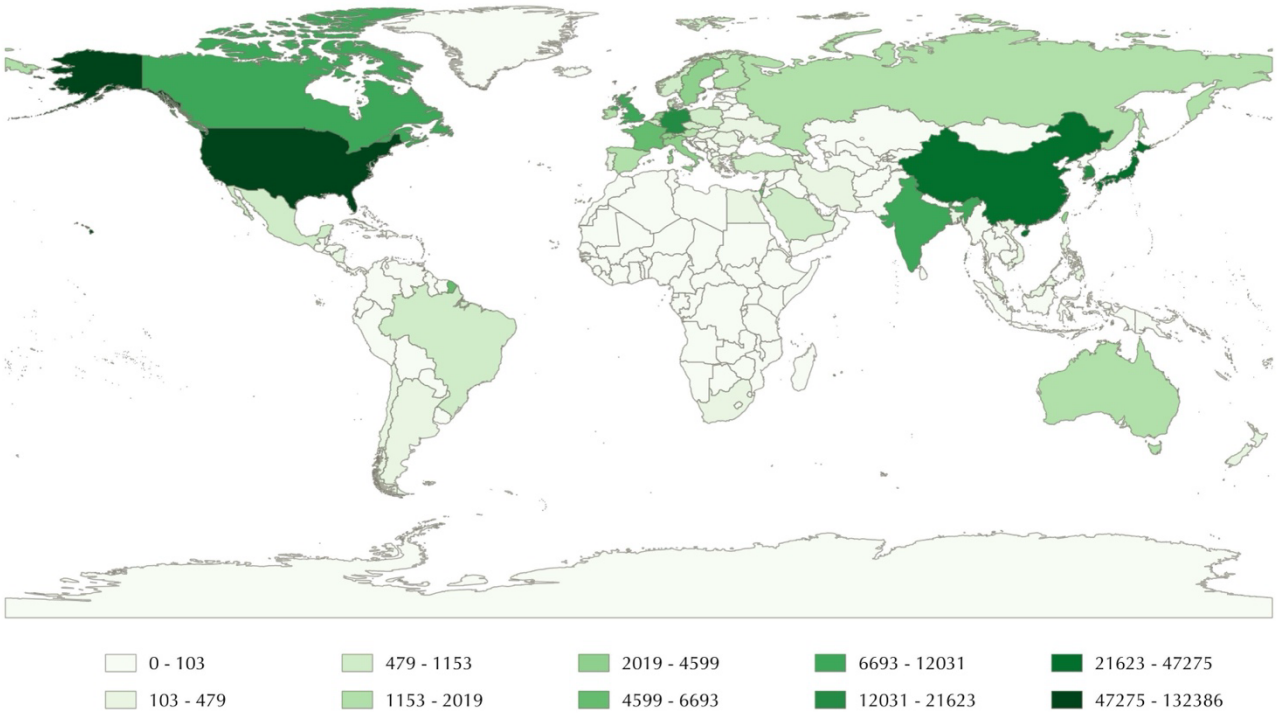


Figure 6 maps the geographic origins of AI-related DOCDB patent families over 2011–2021, using inventors' addresses. This choice ensures to capture the loci of the invention, as it often indicates a laboratory, a research establishment, or the place of residence of the inventor (Maraut et al., 2008; de Rassenfosse et al., 2019). To this end, we leveraged data contained in the USPTO PatentsView and OECD Regpat databases. The results of our machine learning classification models reaffirm that AI technologies are predominantly developed in the United States and Asia, while Europe appears to lag behind, despite some notable exceptions (such as Germany). This distribution mirrors earlier findings (Baruffaldi et al., 2020; Van Roy et al., 2020; Dernis et al., 2021; Gonzales, 2023; Miric et al., 2023; Santarelli et al., 2023; Damioli et al., 2024) and underscores two critical insights: first, global AI innovation remains concentrated in a handful of countries; second, Asia's footprint is especially pronounced, even within a universe that necessarily overweights filings at the USPTO, EPO, and WIPO. Indeed, due to the availability of georeferenced inventor data, it is important to recall that our patent universe includes only patent families in which at least one application is filed at the USPTO, EPO, or WIPO[44]. While this may lead to a relative overrepresentation of US- or EU-based patenting activity, the data nonetheless capture the prominent role of Asian countries in global AI patenting. Likewise,

[44] As a result, we therefore omit AI-related patents whose family members reside exclusively in national patent offices not captured in our dataset.

extending the analysis to applications filed at national offices (such as the Chinese, Japanese, or Korean patent offices) would further reinforce the leading position of Asia in this domain.

Figure 6. Production of AI patent families across countries – Period 2011-2021



6.2. Sectoral patterns of AI innovation

To delineate the sectoral contours of AI innovation, we follow established practice (Van Roy et al., 2020; Santarelli et al., 2023; Damioli et al., 2024, 2025; D'Alessandro et al., 2025) by linking each AI-related DOCDB family to its patent applicant and, through that entity, to an economic sector. Specifically, we enrich our AI-patent dataset with firm-level information from Moody's Orbis and Orbis IP, which provide comprehensive information at the firm level. By mapping each applicant to its principal activity under the NACE Rev. 2 two-digit taxonomy, we capture the industrial origin of AI inventions and, by extension, the underlying knowledge bases driving them. This matching exercise yielded successful assignments for 436,185 patent families (91% of our AI corpus). We then aggregate the two-digit NACE codes into broader sectoral classes (as detailed in Table A5 in the appendix) following the grouping scheme of Damioli et al. (2025). These consolidated classes enable us to compare AI patenting activity across key industries and to identify which sectors are leading or lagging in AI-related invention.

Figure 7. AI patenting activity across sectoral classes

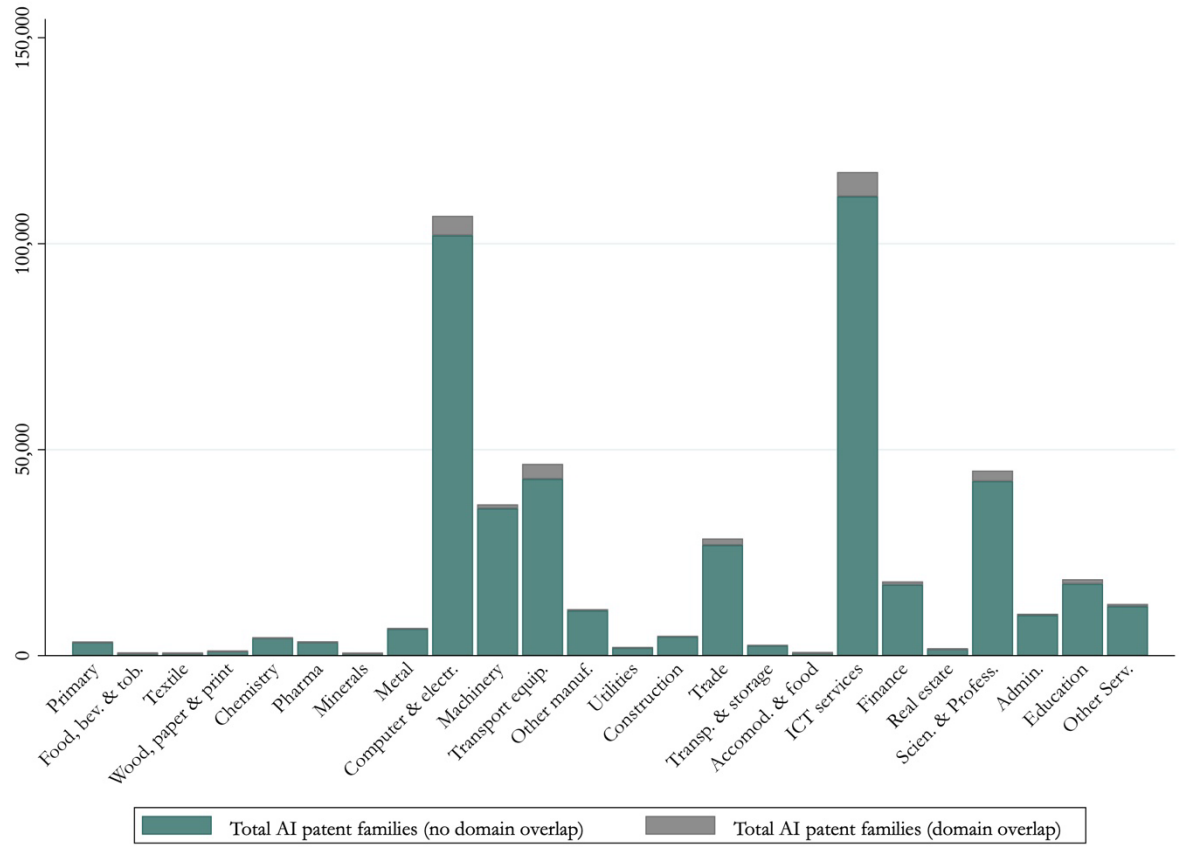
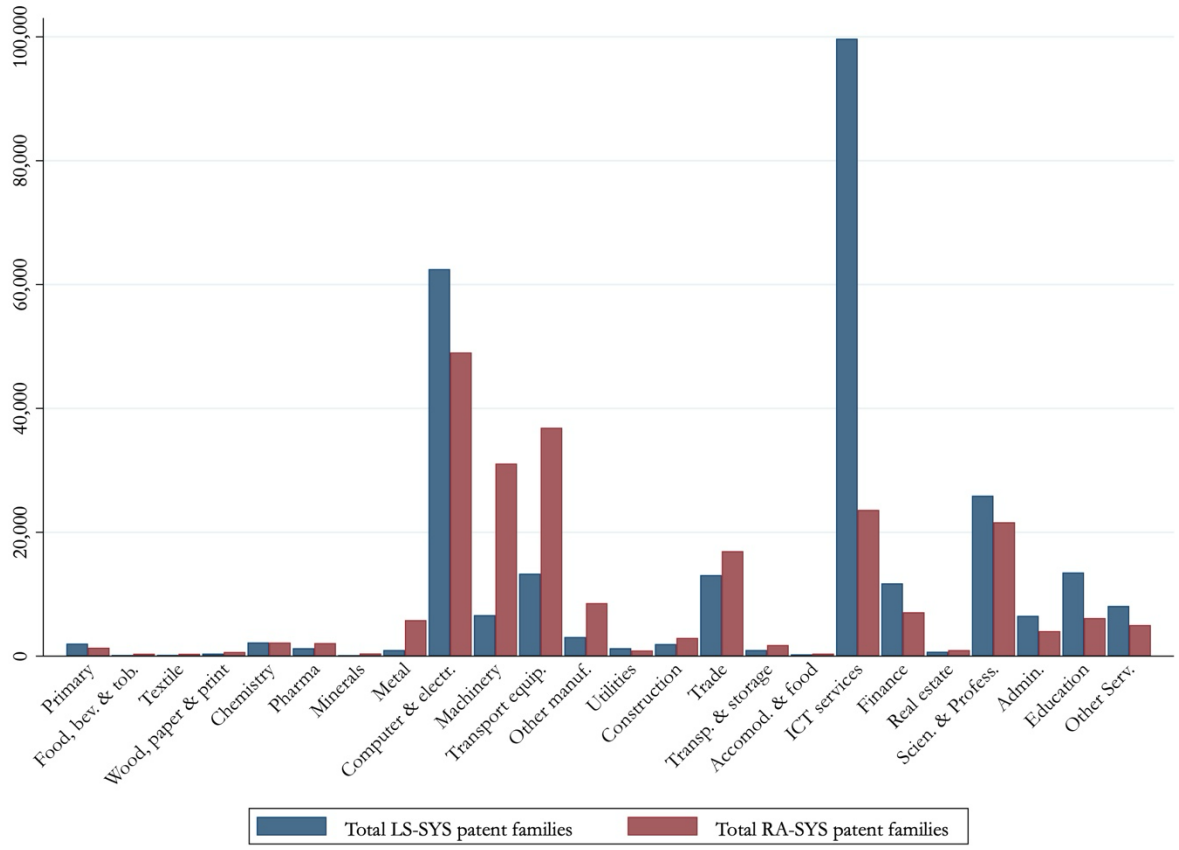


Figure 7 presents the distribution of AI-related DOCDB families across our aggregated sectoral classes. As in prior work (Van Roy et al., 2020; Dernis et al., 2021; Santarelli et al., 2023; Damioli et al., 2024, 2025; Calvino et al., 2024; D'Alessandro et al., 2025), we find that both ICT manufacturing and ICT services overwhelmingly lead the AI revolution, at least as far as patents are concerned. This evidence is consistent with the intuition that AI-related technologies have emerged from the broader ICT-related technological paradigm (Dosi, 1982 and 1988), following a path-dependent evolution. These aspects have been corroborated by recent empirical analysis (Lee & Lee, 2021; Igna & Venturini, 2023; Santarelli et al., 2023; Xiao and Boschma, 2023; D'Alessandro et al., 2025). This implies the presence of sectoral cumulativeness and increasing returns in AI inventions (Breschi & Malerba, 1997). However, significant contributions also emerge from non-ICT sectors—machinery, transportation, trade, and scientific & professional services—confirming the diversification documented by Damioli et al. (2025). The growing involvement of these industries may suggest that AI is coalescing into a standalone technological paradigm, extending beyond its ICT roots and reshaping innovation trajectories across the economy. To deepen our validation, we partition AI patent families into two mutually exclusive cohorts: namely, LS-SYS or RA-SYS patent families (green bars), and "dual-domain" patents (grey bars). Strikingly, these dual-domain innovations concentrate almost exclusively in four core industries: ICT manufacturing, ICT services, and transport equipment and, to a lesser extent, Scientific & Professional services. Such clustering underscores the inherently complex nature of "intelligent" robotics and next-generation autonomous systems: they draw concurrently on symbolic reasoning, data-driven learning,

perception, and actuation capabilities. In other words, these breakthrough inventions represent true knowledge recombinations that require the combination of heterogeneous technological modules (Weitzman, 1998).

Figure 8. LS-SYS and RA-SYS patenting activity across sectoral classes



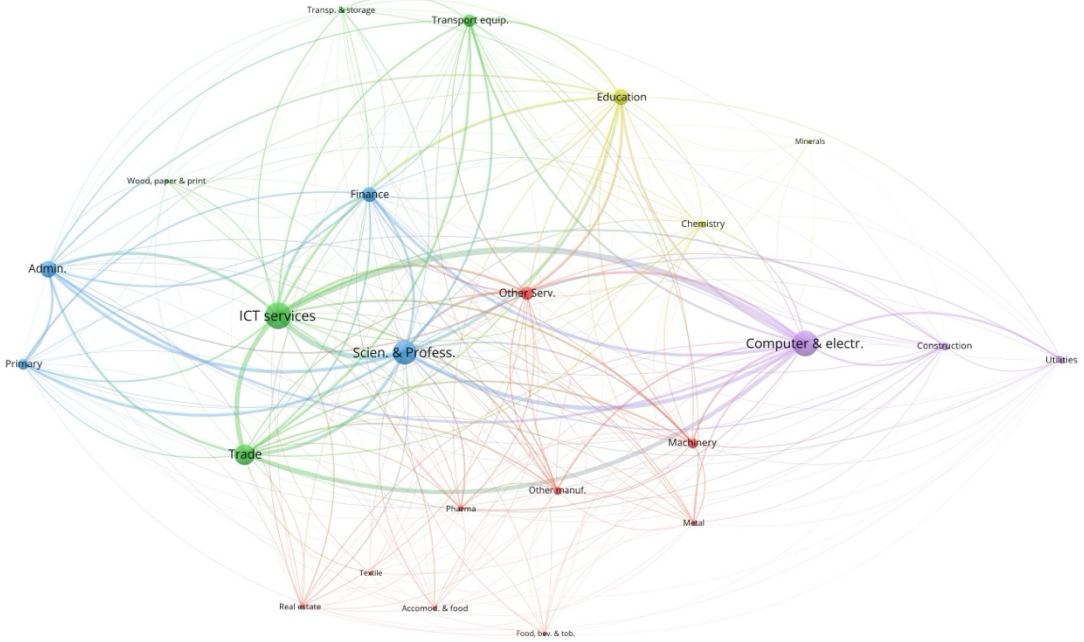
To unpack the sectoral drivers behind Figure 7, we disaggregate AI patents into LS-SYS and RA-SYS cohorts and trace their industrial origins. Results are provided in Figure 8. Consistent with the definition of the two domains, we find interesting results: (a) Core ICT manufacturing sectors feature prominently in both domains, underscoring their cross-domain strength in LS-SYS and RA-SYS software and hardware development. This highlights their pivotal role in supplying the physical platforms on which advanced reasoning and physical intelligence converge. (b) Core ICT service sectors, including software vendors, cloud-computing firms, and consulting groups, account for the lion's share of LS-SYS innovations. This reflects their expertise in developing and deploying algorithmic learning and symbolic-reasoning systems, which are inherently software-centric and rely on scalable, service-based infrastructures (Calvino & Fontanelli, 2023; Calvino et al., 2024). (c) Industrial machinery and transportation equipment sectors lead the RA-SYS domain, consistent with the idea that these sectors have both competencies and resources functional to create the physical embodiments, such as robots, autonomous vehicles, and intelligent machinery. These results further confirm earlier findings that robotics innovations are concentrated in traditional heavy industry and transport domains (UKIPO, 2014; Montobbio et al., 2022).

6.3. Sectoral patterns of AI collaboration

In this subsection, we turn to the co-applicant networks underlying AI patent families in order to reveal how sectors collaborate for the production of AI innovations. Building on the sectoral assignments from Subsection 6.2, we treat each DOCDB family as a collaboration event whenever it lists two or more distinct applicant firms. We then aggregate these events at the sectoral level, counting every time firms from Sector A and Sector B co-apply for the same patent. To better capture the nuances across domains, we distinguish collaborations in the LS-SYS domain from those in the RA-SYS realm. We employ VOSviewer's mapping and clustering routines, which position sectors closer together when they share similar collaboration profiles, draw edges whose thickness reflects collaboration frequency, and normalize the resulting co-occurrence data[45].

Figure 9 shows the results for the LS-SYS domain. Interestingly, we find a core collaboration quartet, composed of ICT services, Science & Professional services, Trade, and ICT manufacturing, reflecting intensive, multi-party partnerships to develop learning-and-symbolic-systems technologies. Moreover, as the VOS visualization allows for dimensionality reduction[46], it implies that spatial proximity on the map encodes similarity in collaboration patterns. ICT services, Science & Professional services, and Trade cluster together, suggesting a shared, software-oriented knowledge base, whereas ICT manufacturing sits marginally apart, signaling its complementary role in supplying hardware and systems-integration expertise. Finally, sectors such as Finance, Transportation & Storage, Admin, Education, and Utilities maintain thinner ties to the core, hinting at nascent cross-industry diffusion of LS-SYS methods.

Figure 9. Sectoral collaborations – LS-SYS domain



[45] In particular, after feeding VOSviewer and an edge list where rows define the number of co-occurrences of sector A and B, the software applies a normalization called "association strength", which takes into account size effects. For more details, see van Eck, N. J., & Waltman, L. (2009).

[46] These methods are similar to Multidimensional scaling (MDS) techniques, see van Eck et al. (2010)

Figure 10. Sectoral collaborations – RA-SYS domain

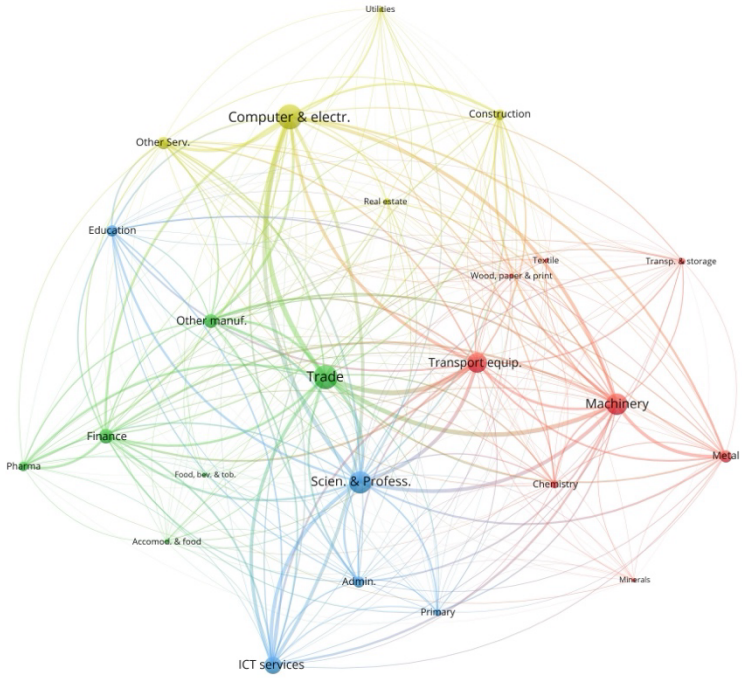


Figure 10 depicts the RA-SYS collaboration network, whose sectoral co-patenting activity forms a distinct and heterogeneous landscape. We find a core "heavy industry" cluster, encompassing Machinery and Transport Equipment. In this respect, it is possible to speculate that these two sectors share a similar knowledge base in factory automation and next-generation vehicles. Indeed, their sectoral partnerships appear to be rather similar: they tend to cooperate with sectors presumably supplying critical complementary knowledge. Notably, the Trade sector occupies a pivotal position, exhibiting robust ties not only to the heavy-industry core but also to ICT and professional-services clusters. This may suggest that Trade firms orchestrate the fusion of diverse technological modules, ranging from sensor networks to control algorithms, underscoring the inherent complexity of RA-SYS innovations in this sector.

To more precisely quantify inter-sectoral dependencies, we follow Calvino and Fontanelli (2025) by computing, for each ordered pair of sectors (i, j), the conditional probability:

$$P_{i|j} = c_{ij}/n_j \tag{1}$$

Where c_{ij} is the number of patent families co-filed by firms in sectors i and j, and n_j represents the total number of patent families involving sector j. In this respect, we estimate the probability of observing a firm from sector i, conditional on observing the presence of a firm from sector j in a given patent family. By evaluating this conditional probability for every (i, j) combination, we obtain an asymmetric square matrix in which each row i captures the strength of sector i's "knowledge-supply" to the various column sectors. Higher values of $P_{i|j}$ indicate that, conditional on sector j participating in a patent, sector i is more likely to be its collaborator, reflecting a stronger flow of knowledge and expertise from i to j. A low probability, instead, means that sector i does not serve as a "knowledge pool" for sector j.

Figure 11 shows the result for the LS-SYS domain, and the resulting matrix reveals a highly concentrated "knowledge-supply" structure within the latter field. When any sector j co-files an LS-SYS patent, there is an average probability of 25% that it does so alongside an ICT manufacturing firm and a 24% chance alongside an ICT services provider, underscoring these two sectors as near-universal hubs of LS-SYS components and expertise. Trade and Science & Professional Services also emerge as secondary knowledge conduits, with an average probability ranging from 10 to 20%. In contrast, all the other sectors show consistently low conditional probabilities, indicating that they rarely act as suppliers of LS-SYS know-how. Together, these patterns point to strong increasing-returns and concentration dynamics. Figure 12's conditional probability matrix for RA-SYS paints a far more polycentric landscape than the LS-SYS domain. While the ICT manufacturing sector remains indispensable, exhibiting an average conditional probability of almost 20%, the Machinery and Transport Equipment rows also display high average conditional probabilities (17% and 14%). This reciprocity underscores that heavy-industry players consume but also supply critical expertise to other industries, such as robotics hardware, sensor integration, and control systems. At the same time, Trade and Science & Professional Services continue to act as major bridges, with an average probability of 16% and 17%, respectively. In contrast, ICT Services' conditional probabilities fall relative to LS-SYS (now 8%), indicating a reduced upstream role for pure software developers in robotics and autonomous systems projects.

Figure 11. Sectoral interdependencies – LS-SYS domain

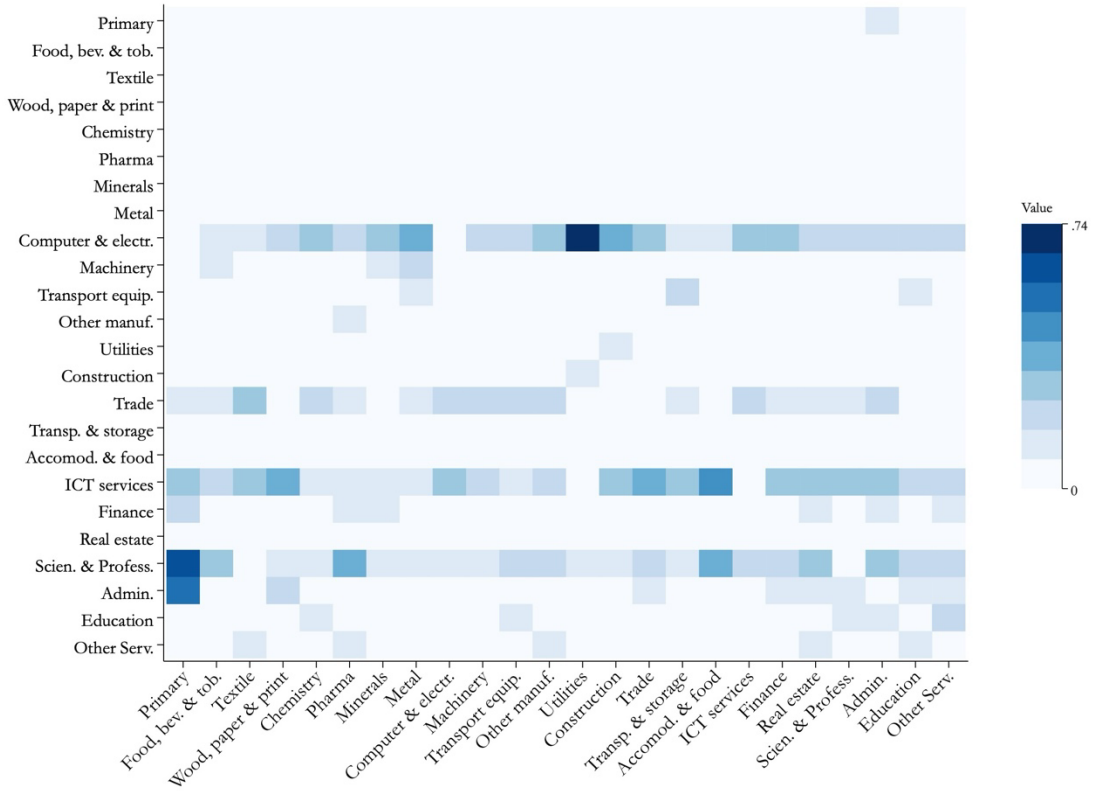
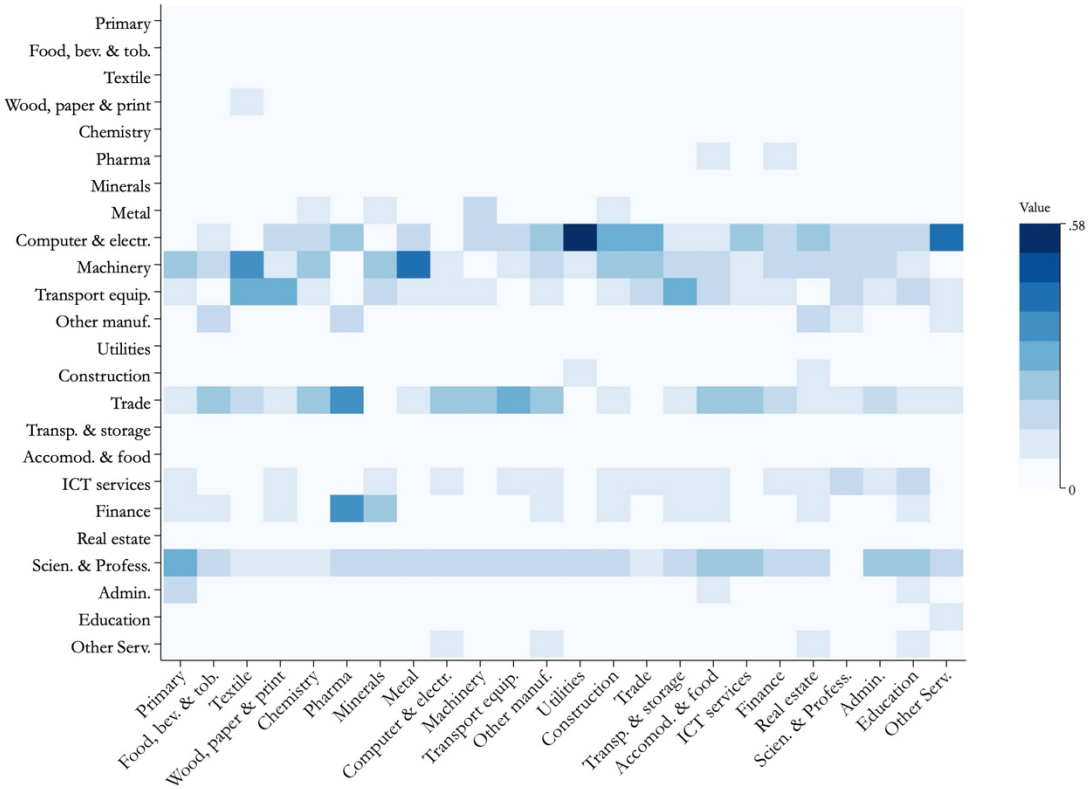


Figure 12. Sectoral interdependencies – RA-SYS domain



6.4. Firms and AI patenting activity

In the Appendix (Tables A6–A8), we report the fifty most prolific patent applicants for LS-SYS, RA-SYS, and the subset referring to the "domain overlap" DOCDB families. Within the Learning and Symbolic Systems domain (Table A6), often termed "narrow AI", ICT firms overwhelmingly dominate the patenting activity. As in Miric et al. (2023), at the summit stands International Business Machines, whose Watson platform has given rise to dozens of patents on natural-language reasoning, knowledge graphs, and deep learning. Close behind are Samsung, the Microsoft group, and Google, which lead the "on-device" AI domain. Chinese technology leaders, such as Baidu, Tencent, Huawei, Ping An Technology, and Alibaba, also feature prominently among the top fifty, reflecting China's rapid ascent in AI research and development. Three distinct innovation archetypes emerge beyond these headline players: first, imaging specialists (e.g., Fujifilm, Xerox, Fujitsu, Canon, Siemens Healthcare, Philips, Nuance) that fuse machine learning with domain expertise to advance medical and industrial imaging solutions. Second, pure-software pioneers (e.g., Microsoft, Google, SAP, Adobe, Meta, Oracle), who are driving breakthroughs in natural language processing, symbolic reasoning, and deep learning frameworks. Third, integrated hardware–software conglomerates (e.g., IBM, Intel, Nvidia, Apple, General Electric) offering end-to-end AI platforms that combine custom silicon, software stacks, and specialized models. Notably, patenting activity also extends to non-ICT firms, such as Amazon's cloud-based AI services and financial institutions like Bank of America, underscoring the diffusion of narrow-AI methods beyond the ICT realm.

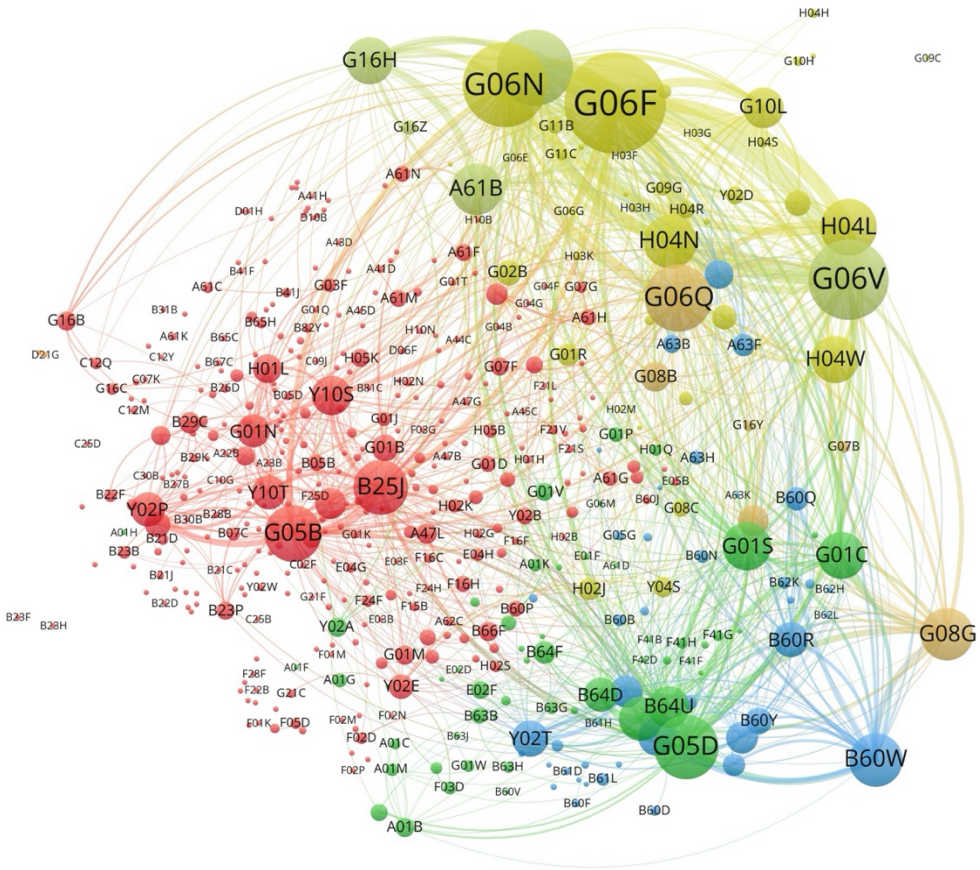
Table A7 vividly illustrates the sectoral breadth of RA-SYS patenting, anchored by a manufacturing-heavy core that spans industrial-robotics specialists, automotive, electronics conglomerates, aerospace contractors, and niche robotics innovators. At the top is the industrial-robotics specialist Fanuc Corporation, which claims to supply more than a hundred different robots for industrial automation applied in welding, assembling, painting, vision inspection, and many other fields. Furthermore, we also find legacy carmakers like Honda, Toyota, Hyundai, and Ford, who have aggressively expanded into robotics and autonomous systems research. For instance, Honda's ASIMO humanoid robot, first revealed in the early 2000s, catalyzed a wave of patents on human-robot interaction, while Toyota's early "Partner Robot" filings foreshadow today's AI-driven mobility solutions. General Motors and Nissan similarly leverage their vehicle platforms to patent advanced driver-assistance and autonomous-driving modules. Furthermore, electronics giants like Samsung and LG leverage sensor and display expertise to develop service robots and smart-home devices, just as Applied Materials and Tokyo Electron patent wafer-handling robots critical to chip fabrication. In aerospace and defense, Boeing's autonomous flight-control systems and Honeywell's foundational drone-guidance patents underpin modern UAV platforms, while DJI's Phantom series further democratizes aerial robotics for photography and inspection. Even non-traditional players make their mark: Amazon, probably thanks to the acquisition of Kiva Systems, spurred a suite of patents on warehouse automation and delivery drones, while healthcare robotics is exemplified by Intuitive Surgical's da Vinci and Covidien's Hugo systems, which mark the frontier in robotic-assisted surgery systems. The overall picture is rather consistent with the evidence provided in prior studies (UKIPO, 2014; Montobbio et al., 2022; Savin et al., 2022).

Table A8 highlights the fifty firms whose patent portfolios most frequently span both the "thinking" (LS-SYS) and "acting" (RA-SYS) dimensions of AI, revealing a richly heterogeneous ecosystem where legacy incumbents and agile newcomers alike co-drive innovation at the intersection of these two subfields. Once again, at the apex sits International Business Machines, whose Watson AI has been embedded in many autonomous systems technologies. Close behind are electronics powerhouses LG and Samsung, each coupling advanced sensor arrays with proprietary AI stacks to power service robots and smart-home platforms. Traditional automotive manufacturers also feature prominently: General Motors, Honda, and Toyota have long infused deep-learning engines into self-driving research and humanoid robotics. Remarkably, several young ventures and spin-offs have staked their claim in this cross-domain space. Among all, two firms belonging to the Alphabet realm, namely, Waymo, which evolved from Google's seminal self-driving car project, and X Development, born as Google's self-driving car arm and now active in more than 20 different field, committed to bringing "sci-fi ideas into reality to help solve some of the world's hardest problems." The Uber Advanced Technology Group similarly patents across delivery drones and freight automation, while Zoox, an Amazon subsidiary, develops purpose-built robotaxi platforms that blend in-house AI learning with bespoke vehicle designs. A final example is Argo AI, which was founded by two scientists from Google and Uber's automated driving programs. Together, the emerging picture reveals a dynamic ecosystem in which hardware manufacturers, software architects, automakers, and startup pioneers all converge, melding learning algorithms with robotic embodiments to drive the next frontier of intelligent machines.

6.5. AI Technological Space

Finally, we validate our classifications by mapping the underlying technological building blocks of AI patents using CPC codes. As a first general assessment, Figure 13 illustrates the VOSviewer-generated technological space formed by AI patent families, following the conceptualization proposed by Hidalgo et al. (2007) and Hidalgo and Hausmann (2009). In this network visualization, each node corresponds to a unique 4-digit CPC code: its size reflecting both frequency in our AI corpus and the number of co-occurrence links, while edge thickness denotes the strength of associations between codes (van Eck & Waltman, 2009). Given the characteristics of the VOS mapping technique, this approach allows for an intuitive interpretation of the technological clustering within the AI technological space. Notably, the VOS mapping reveals a distinction between the two core AI domains, as was the case for the AI scientific knowledge space (Figure 3). In the top-right region of the map, one can observe a dense and cohesive cluster representing the Learning and Symbolic Systems [47] domain, characterized by a concentration of frequently occurring CPC codes with strong interconnections. In contrast, the lower-left and lower-right regions of the map capture the Robotics and Autonomous Systems domain[48], which is more spatially dispersed. This clear spatial separation of CPC clusters provides compelling, orthogonal evidence that our machine-learning models have successfully distinguished the two core AI subfields.

Figure 13. AI patent families' CPC codes co-occurrence map



[47] Colored in yellow.
[48] Robotics-related CPC codes are colored in red, while Autonomous Systems ones are in blue/green.

Finally, Tables A9, A10, and A11 in the appendix show the top-10 4-digit CPC codes associated with LS-SYS, RA-SYS, and "domain overlap" patent families, respectively. In Table A9, which captures the overall LS-SYS technological landscape, reveals a pronounced concentration in general-purpose computing and data-processing classes, with a strong concentration in G06F (Electric digital data processing, 46.8%) and G06N (Computing arrangements based on specific computational models, 29%), both of which cover core software-based and machine learning-driven innovations. Notably, image- and video-processing classes, such as G06V (17.5%), G06T (17.2%), H04L (9.7%), and H04N (6.5%), feature prominently, corroborating our firm-level findings that a substantial subset of top patentees are imaging specialists leveraging pattern-recognition techniques in medical, industrial, and multimedia applications. Furthermore, speech and voice recognition technologies also seem to be highly represented (G10L, 8.7%). Finally, the appearance of G16H (5.9%) and A61B (5.3%) signals the growing penetration of learning and symbolic methods into medical diagnostics and healthcare workflows.

By contrast, Table A10 reflects the more heterogeneous and spatially dispersed nature of the RA-SYS domain, corresponding to the lower-left and right regions of the map in Figure 13. In this field, the top three codes are B25J (Manipulators; 13.5%), G05D (Control of non-electric variables; 12.7%), and G05B (General control systems; 10.1%), reflecting the hardware-centric core of RA-SYS innovation. General digital processing (G06F, 9.6%) and image-processing codes (G06T, 7.4%; G06V, 7.2%) appear next, highlighting the importance of on-board computation and perception. Vehicle drive control (B60W, 8.2%) and traffic systems (G08G, 7.1%) underscore the transportation focus, while G01S (Radio navigation and positioning, 6.5%) points to sensing and localization functionalities crucial for autonomous platforms. Finally, the repeated appearance of A61B (Diagnosis; surgery; identification) illustrates its cross-domain relevance, particularly as a key application field at the intersection of healthcare, robotics, and data processing. This list, however, indicates a stronger orientation toward control systems, robotics, and physical interaction with the environment; the overall diversity and lower shares align with the broader and less dense distribution of nodes in the RA-SYS cluster on the map.

Lastly, Table A11 merges these two profiles, illustrating the technological convergence of "thinking" and "acting" by analyzing the "domain overlap" AI patent families. Here, general data processing (G06F, 31.9%), image recognition (G06V, 29.8%), image-data processing (G06T, 26.7%), and computational modelling techniques (G06N, 24.7%) remain dominant, reflecting the learning and symbolic reasoning backbone of intelligent systems. At the same time, control-oriented and hardware-centric classes, such as G05D (19.2%), G05B (14.9%), B25J (12.4%), and B60W (15.1%), feature prominently, indicating that these AI algorithms are deeply integrated with mechanical and vehicular control technologies. Together, these tables confirm that patents spanning both domains combine robust computational capabilities with sophisticated control and actuation modules, validating the cross-domain classification achieved by our models.

7. Conclusions

Artificial intelligence has rapidly sparked a new technological revolution, reshaping economic structures, scientific discovery, and competitive dynamics across industries and geographies. Its dual role, as both a GPT and GP-IMI, underpins an urgent need for robust methods to trace AI's evolution and diffusion. Scholars have deployed a variety of approaches (e.g., mining scientific publications and patents, analyzing labor-market signals, and scraping corporate websites) to approximate where and how AI technologies originate, propagate, and translate into real-world applications. Yet each method carries trade-offs in timeliness, coverage, and semantic precision, and static keyword- or classification-code schemes struggle to keep pace with AI's rapidly shifting lexicon and research frontiers.

In this paper, we introduce a fully reproducible, machine-learning–driven framework to map AI innovation through patent data, targeting two foundational subfields: Learning and Symbolic Systems and Robotics and Autonomous Systems. Building on recent advances in domain-specific pre-trained transformers, we fine-tuned two BERT for Patents models (Srebrovic & Yonamine, 2020) using seed sets derived from a deep search leveraging an extended list of weighted n-grams and targeted heuristics, coupled with a negative-sampling procedure grounded in the automated patent landscaping procedure (Abood and Feltenberger, 2018). This hybrid pipeline operates on English-language titles and abstracts across three patent offices (USPTO, EPO, WIPO) over a forty-year horizon, delivering classifiers that dynamically adapt to emergent AI subdomains.

Our empirical analysis validates the results of our models and unveils the contours of AI patenting since 1980. We document a pronounced inflection in post-2012 filings, corresponding to the Deep Learning Revolution, with LS-SYS patent families outpacing RA-SYS ones. Geographically, the United States and select Asian nations anchor global AI production, while Europe lags, save for outliers such as Germany. As far as sectoral patterns of AI innovation are concerned, ICT-intensive industries dominate LS-SYS inventions, whereas RA-SYS patents spring from a more polycentric constellation, including machinery, transport equipment, and trade integrators. Collaboration networks further reveal ICT manufacturing and services as central knowledge hubs for LS-SYS, contrasted by a heavy-industry core augmented by professional and trade sectors in RA-SYS. Conditional-probability matrices and CPC-based technological maps corroborate these domain distinctions and highlight the rich interplay of computation and control in the subset of "domain overlap" patents. Finally, firm-level insights offer a fine-grained perspective on who is driving AI innovation today. In LS-SYS, top patentees reflect the primacy of software and algorithm developers, alongside imaging specialists, and an emerging cohort of Chinese leaders. Within RA-SYS, industrial-robotics players sit alongside automotive manufacturers and aerospace firms, underscoring hardware's centrality. The "domain-overlap" set features true cross-domain pioneers, paired with agile spinouts and startups, illustrating how hardware and software innovation converge at the AI frontier.

The contributions of this work are threefold. First, we systematically assess existing AI-tracking methodologies, clarifying their complementary strengths and motivating the emergence of ML-powered approaches as the new dominant paradigm in AI-related research. Second, we develop and publicly release two fine-tuned BERT for Patents classifiers, along with an extended list of n-grams, that form a reusable toolkit

for AI monitoring. Third, we provide an empirical analysis in which we validate our models and describe the evolution of the AI patenting activity across a four-decade horizon, detailing its temporal inflection, geographic clustering, sectoral contours, collaboration dynamics, and firm-level leadership.

We highlight several limitations and avenues for future research. First, by confining our analysis to titles and abstracts, we may overlook nuanced disclosures residing in full texts. Yet, extensive evidence from patent analytics research confirms that abstracts reliably convey the core inventive contributions of most filings. Second, our Patent Universe includes only DOCDB families with at least one filing at the USPTO, EPO, or WIPO, as well as English-only publications, potentially undercounting inventions protected solely at national offices (e.g., CNIPA, JPO, KIPO). However, these three authorities together represent the world's largest innovation markets, and patentees routinely seek coverage there for high-value AI inventions. On the methodological front, in place of an entirely classification-code-based seed-sampling strategy (as in Giczy et al., 2022), we employ a hybrid approach that combines detailed n-grams with human-inspired heuristics. As a result, our seed-set construction requires that each initial example must contain at least one "tier-one" keyword drawn from our weighted n-gram lexica. While these heuristics were designed to maximize precision and reduce noise, they lead to an underrepresentation of patents lacking those explicit terms during training. In practice, however, the extensive coverage of our tier-one lists (86 tier-one terms for LS-SYS; 20 for RA-SYS), combined with secondary scoring heuristics, ensures comprehensive coverage of established AI concepts. Even in manual coding, human experts likely rely on these same hallmark terms to identify AI content. To broaden both seed sets, we leave to future work the inclusion of patents lacking tier-one keywords; however, while increasing the number of patents defined as AI-related, this will unlikely alter the dominant patterns we documented. In this context, a natural extension would be to adopt weak-supervision frameworks (Ratner et al., 2017; Boecking et al., 2021), in which large training corpora are programmatically generated by combining multiple labeling functions (such as keyword matches, technology-class filters, and citation-link heuristics) and then reconciling their outputs into a unique high-quality label. Finally, we have not fine-tuned our models on "borderline" patents, namely, those whose relevance scores cluster around the decision threshold of 50%. Incorporating borderline cases through active-learning schemes (Pairolero et al., 2025) would further enhance model generalization and alleviate contextual undercoverage bias, an especially critical step when deploying these classifiers across the entire patent universe. We reserve these aspects for future studies.

References

Abood, A., & Feltenberger, D. (2018). Automated patent landscaping. Artificial Intelligence and Law, 26(2), 103-125.

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. Journal of Labor Economics, 40(S1), S293-S340.

Acs, Z. J., Braunerhjelm, P., Audretsch, D. B., & Carlsson, B. (2009). The knowledge spillover theory of entrepreneurship. Small Business Economics, 32(1): 15-30.

Aghion, P., & Howitt, P. (1992). A Model of Growth Through Creative Destruction. Econometrica, 60(2), 323–351.

Agrawal, A., McHale, J., & Oettl, A. (2019). Finding needles in haystacks: Artificial intelligence and recombinant growth. In Agrawal, A., Gans, J., & Goldfarb, A. (eds.), The economics of artificial intelligence: an agenda. University of Chicago Press.

Agrawal, A., McHale, J., & Oettl, A. (2023). Superhuman science: How artificial intelligence may impact innovation. Journal of Evolutionary Economics, 33(5), 1473-1517.

Agrawal, A., McHale, J., & Oettl, A. (2024). Artificial intelligence and scientific discovery: A model of prioritized search. Research Policy, 53(5), 104989.

Alderucci, D., Branstetter, L., Hovy, E., Runge, A., & Zolas, N. (2020). Quantifying the impact of AI on productivity and labor demand: Evidence from US census microdata. In Allied social science associations—ASSA 2020 annual meeting.

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

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and unsupervised learning for data science, 3-21.

Audretsch, D. B., Keilbach, M. C., & Lehmann, E. E. (2006). Entrepreneurship and economic growth. Oxford University Press.

Babina, T., Fedyk, A., He, A., & Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. Journal of Financial Economics, 151, 103745.

Baruffaldi, S., van Beuzekom, B., Dernis, H., Harhoff, D., Rao, N., Rosenfeld, D., & Squicciarini, M. (2020). Identifying and measuring developments in artificial intelligence: Making the impossible possible. OECD Science, Technology and Industry Working Papers, 2020(5), 1-68.

Bianchini, S., Müller, M., & Pelletier, P. (2022). Artificial intelligence in science: An emerging general method of invention. Research Policy, 51(10)

Boecking, B., Neiswanger, W., Xing, E., & Dubrawski, A. (2020). Interactive weak supervision: Learning useful heuristics for data labeling. arXiv preprint arXiv:2012.06046.

Borgonovi, F., Calvino, F., Criscuolo, C., Nania, J., Nitschke, J., O'Kane, L., Samek, L., & Seitz, H. (2023). Emerging trends in AI skill demand across 14 OECD countries. OECD Artificial Intelligence Papers, No. 2, OECD Publishing, Paris.

Braunerhjelm, P., Acs, Z.J., Audretsch, D. B., & Carlsson, B. (2009). The missing link: Knowledge diffusion and entrepreneurship in endogenous growth. Small Business Economics, 34(2): 105-125.

Breschi, S. & Malerba, F. (1997). Sectoral Innovation Systems: Technological Regimes, Schumpeterian Dynamics and Spatial Boundaries. In Edquist, C. (ed.), Systems of Innovation. Technologies, Institutions and Organizations, Routledge.

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies 'Engines of growth'?. Journal of econometrics, 65(1), 83-108.

Burning Glass Technologies (2019). Mapping the genome of jobs: The burning glass skills taxonomy. Boston.

Calvino, F., & Fontanelli, L. (2023). A portrait of AI adopters across countries. Documents de travail de l'OCDE sur la science, la technologie et l'industrie.

Calvino, F., & Fontanelli, L. (2025). Decoding AI: Nine facts about how firms use artificial intelligence in France (No. 2025/13). LEM Working Paper Series.

Calvino, F., Dernis, H., Samek, L., & Ughi, A. (2024). A sectoral taxonomy of AI intensity. OECD Artificial Intelligence Papers, No. 30, OECD Publishing, Paris.

Calvino, F., Samek, L., Squicciarini, M., & Morris, C. (2022). Identifying and characterising AI adopters: A novel approach based on big data. OECD Science, Technology and Industry Working Papers, No. 2022/06, OECD Publishing, Paris.

Carbonara, E., & Santarelli, E. (2023). Artificial intelligence and robots: A threat or an opportunity for SMEs and entrepreneurship?. In Carbonara, E., & Tagliaventi, M. (eds.). SMEs in the digital era. Cheltenham: Edward Elgar, pp. 104-121.

Choi, S., Lee, H., Park, E., & Choi, S. (2022). Deep learning for patent landscaping using transformer and graph embedding. Technological Forecasting and Social Change, 175, 121413.

Chung, P., & Sohn, S. Y. (2020). Early detection of valuable patents using a deep learning model: case of semiconductor industry. Technological Forecasting and Social Change 158, 120146.

Cicerone, G., Faggian, A., Montresor, S., & Rentocchini, F. (2023). Regional artificial intelligence and the geography of environmental technologies: does local AI knowledge help regional green-tech specialization?. Regional Studies, 57(2), 330-343.

Cockburn, I. M., Henderson, R., & Stern, S. (2019). The impact of artificial intelligence on innovation. In Agrawal, A., Gans, J., & Goldfarb, A. (eds.), The economics of artificial intelligence: an agenda. University of Chicago Press.

Colombelli, A., D'Amico, E., & Paolucci, E. (2023). When computer science is not enough: universities knowledge specializations behind artificial intelligence startups in Italy. The Journal of Technology Transfer, 48(5), 1599-1627.

D'Alessandro, M., Santarelli, E., & Vivarelli, M. (2025). The KSTE + I approach and the advent of AI technologies: Evidence from European regions. Journal of Technology Transfer.

Dahlke, J., Beck, M., Kinne, J., Lenz, D., Dehghan, R., Wörter, M., & Ebersberger, B. (2024). Epidemic effects in the diffusion of emerging digital technologies: evidence from artificial intelligence adoption. Research Policy, 53(2), 104917.

Dahlke, J., Schmidt, S., Lenz, D., Kinne, J., Dehghan, R., Abbasiharofteh, M., ... & Rammer, C. (2025). The WebAI Paradigm of Innovation Research: Extracting Insight From Organizational Web Data Through AI.

Damioli, G., Van Roy, V., Vértesy, D., & Vivarelli, M. (2024). Drivers of employment dynamics of AI innovators. Technological Forecasting and Social Change, 201, 123249.

Damioli, G., Van Roy, V., Vertesy, D., & Vivarelli, M. (2025). Is artificial intelligence leading to a new technological paradigm?. Structural Change and Economic Dynamics, 72, 347-359.

Dass, N., Nanda, V., & Xiao, S. C. (2017). Truncation bias corrections in patent data: Implications for recent research on innovation. Journal of Corporate Finance, 44, 353-374.

Davidsson, P., & Sufyan, M. (2023). What does AI think of AI as an external enabler (EE) of entrepreneurship? An assessment through and of the EE framework. Journal of Business Venturing Insights, 20, e00413.

de Rassenfosse, G., Kozak, J., & Seliger, F. (2019). Geocoding of worldwide patent data. Scientific Data, 6(1), 260.

Dernis, H., Calvino, F., Moussiegt, L., Nawa, D., Samek, L., & Squicciarini, M. (2023). Identifying artificial intelligence actors using online data. OECD Science, Technology and Industry Working Papers, No. 2023/01, OECD Publishing, Paris.

Dernis, H., Moussiegt, L., Nawa, D., & Squicciarini, M. (2021). Who develops AI-related innovations, goods and services?. OECD Science, Technology and Industry Policy Papers.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Dosi, G. (1982). Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. Research Policy, 11(3), 147-162.

Dosi, G. (1988). Sources, Procedures, and Microeconomic Effects of Innovation. Journal of Economic Literature, 26, 1120-1171.

Dunham, J., Melot, J., & Murdick, D. (2020). Identifying the development and application of artificial intelligence in scientific text. arXiv preprint arXiv:2002.07143.

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

Giczy, A. V., Pairolero, N. A., & Toole, A. A. (2022). Identifying artificial intelligence (AI) invention: A novel AI patent dataset. The Journal of Technology Transfer, 47(2), 476-505.

Gonzales, J. T. (2023). Implications of AI innovation on economic growth: a panel data study. Journal of Economic Structures, 12(1), 13.

Grashof, N., & Kopka, A. (2023). Artificial intelligence and radical innovation: an opportunity for all companies?. Small Business Economics, 61(2), 771-797.

Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. Proceedings of the national academy of sciences, 106(26), 10570-10575.

Hidalgo, C. A., Klinger, B., Barabási, A. L., & Hausmann, R. (2007). The product space conditions the development of nations. Science, 317(5837), 482-487.

High-level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI. European Commission.

Igna, I., & Venturini, F. (2023). The determinants of AI innovation across European firms. Research Policy, 52(2), 104661.

Jha, M., Qian, J., Weber, M., & Yang, B. (2024). ChatGPT and corporate policies (No. w32161). National Bureau of Economic Research.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. Journal of Economic Literature, 61(4), 1281-1317.

Lamperti, F. (2024). Unlocking machine learning for social sciences: The case for identifying Industry 4.0 adoption across business restructuring events. Technological Forecasting and Social Change, 207, 123627.

Lee, J. S., & Hsiang, J. (2020). Patent classification by fine-tuning BERT language model. World Patent Information, 61, 101965.

Lee, J., & Lee, K. (2021). Is the fourth industrial revolution a continuation of the third industrial revolution or something new under the sun? Analyzing technological regimes using US patent data. Industrial and Corporate Change, 30(1), 137-159.

Mann, K., & Püttmann, L. (2023). Benign effects of automation: New evidence from patent texts. Review of Economics and Statistics, 105(3), 562-579.

Maraut, S., Dernis, H., Webb, C., Spiezia, V., & Guellec, D. (2008). The OECD REGPAT database: a presentation. OECD.

Martínez, C. (2010). Patent families: When do different definitions really matter?. Scientometrics, 86(1), 39-63.

Miric, M., Jia, N., & Huang, K. G. (2023). Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. Strategic Management Journal, 44(2), 491-519.

Montobbio, F., Staccioli, J., Virgillito, M. E., & Vivarelli, M. (2022). Robots and the origin of their labour-saving impact. Technological Forecasting and Social Change, 174, 121122.

Nilsson, N. (2010). The Quest for Artificial Intelligence: A History of Ideas and Achievements. Cambridge: Cambridge University Press.

Pairolero, N. A., Giczy, A. V., Torres, G., Islam Erana, T., Finlayson, M. A., & Toole, A. A. (2025). The artificial intelligence patent dataset (AIPD) 2023 update. The Journal of Technology Transfer, 1-24.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. In Proceedings of the VLDB Endowment, 11(3), 269–282.

Santarelli, E., Staccioli, J., & Vivarelli, M. (2023). Automation and related technologies: a mapping of the new knowledge base. The Journal of Technology Transfer, 48(2), 779-813.

Savin, I., Ott, I., & Konop, C. (2022). Tracing the evolution of service robotics: Insights from a topic modeling approach. Technological Forecasting and Social Change, 174, 121280.

Schumpeter, J. A. (1939). Business cycles: A theoretical, historical and statistical analysis of the capitalist process. McGraw-Hill, New York.

Sejnowski, T. J. (2018). The deep learning revolution. MIT press.

Souza, D., Geuna, A., & Rodríguez, J. (2024). How Small is Big Enough? Open Labeled Datasets and the Development of Deep Learning. arXiv preprint arXiv:2408.10359.

Squicciarini, M. & Nachtigall, H. (2021). Demand for AI skills in jobs: Evidence from online job postings. OECD Science, Technology and Industry Working Papers, No. 2021/03, OECD Publishing, Paris.

Srebrovic, R., & Yonamine, J. (2020). Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery. White paper.

Trajtenberg, M. (2019). Artificial intelligence as the next gpt. In Agrawal, A., Gans, J., & Goldfarb, A. (eds.), The economics of artificial intelligence: an agenda. University of Chicago Press.

United Kingdom Intellectual Property Organization (UKIPO) (2014). Eight Great Technologies - Robotics and Autonomous Systems: A Patent Overview. London: UK Intellectual Property Office.

van Eck, N. J. & Waltman, L. (2023). VOSviewer Manual.

van Eck, N. J., & Waltman, L. (2007). Bibliometric mapping of the computational intelligence field. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 15(05), 625-645.

van Eck, N. J., & Waltman, L. (2009). How to Normalize Cooccurrence Data? An Analysis of Some Well-Known Similarity Measures. Journal of the American Society for Information Science and Technology 60(8), 1635-1651.

van Eck, N. J., & Waltman, L. (2014). Visualizing bibliometric networks. Measuring scholarly impact: Methods and practice. Springer International Publishing, 285-320.

van Eck, N.J., Waltman, L., Dekker, R., & Van den Berg, J. (2010). A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. Journal of the American Society for Information Science and Technology, 61(12), 2405–2416.

Van Roy, V., Vertesy, D., Damioli, G. (2020). AI and Robotics Innovation. In Zimmermann, K. (ed.), Handbook of Labor, Human Resources and Population Economics. Springer.

Vannuccini, S., & Prytkova, E. (2023). Artificial Intelligence's new clothes? A system technology perspective. Journal of Information Technology.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Waltman, L., van Eck, N.J., & Noyons, E.C.M. (2010). A unified approach to mapping and clustering of bibliometric networks. Journal of Informetrics, 4(4), 629–635.

Weitzman, M. L. (1998). Recombinant growth. The Quarterly Journal of Economics, 113(2), 331-360.

World Intellectual Property Organization (WIPO) (2019). WIPO technology trends 2019– Artificial intelligence.

Xiao, J., & Boschma, R. (2023). The emergence of artificial intelligence in European regions: the role of a local ICT base. Annals of Regional Science, 71, 747–773.

Appendix

Table A1. Relevant CPC codes – LS-SYS Level 1 Expansion

CPC codes	Technological area
G05B13/0265, G05B13/027	Adaptive control systems, using learning criterion or neural network only
G06F18/211, G06F18/2113, G06F18/213, G06F18/214, G06F18/2148, G06F18/2155, G06F18/217, G06F18/2178, G06F18/24, G06F18/241, G06F18/2411, G06F18/2413, G06F18/24133, G06F18/2414, G06F18/24143, G06F18/2415, G06F18/2431, G06F18/253, G06F18/285	Pattern recognition, feature extraction, generating training patterns, classification techniques
G06F30/27	Computer-aided design using machine learning
G06F2207/4824	Neural network arrangements for processing data
G06N3/006, G06N3/02, G06N3/04, G06N3/042, G06N3/043, G06N3/044, G06N3/0442, G06N3/045, G06N3/0455, G06N3/0464, G06N3/047, G06N3/0475, G06N3/048, G06N3/049, G06N3/063, G06N3/065, G06N3/08, G06N3/082, G06N3/084, G06N3/086, G06N3/088, G06N3/09, G06N3/10, G06N3/105, G06N3/126	Computing arrangements based on biological models, e.g. social simulations or particle swarm optimization, neural networks, fuzzy logic, using electronic, learning means, backpropagation, supervised or unsupervised learning
G06N5/01, G06N5/04, G06N5/045, G06N5/046	Computing arrangements using knowledge-based models, e.g. dynamic search techniques, heuristics, dynamic trees, branch-and-bound, inference or reasoning models
G06N7/01	Computing arrangements based on probabilistic networks
G06N20/00, G06N20/10, G06N20/20	Machine learning
G06T3/4046	Geometric image transformations in the plane of the image using neural networks
G06T2207/20081, G06T2207/20084	Indexing scheme for image analysis or image enhancement by training a model
G06V10/454, G06V10/764, G06V10/774, G06V10/776, G06V10/82, G06V10/95	Arrangements for image or video recognition or understanding, e.g. convolutional neural networks, classification models, pattern learning
G06V30/19173	Character recognition; Recognizing digital ink; Document-oriented image-based pattern recognition using classification methods
G06V2201/03	Recognition of patterns in medical or anatomical images
G10L15/16	Speech recognition using neural networks
G10L25/30	Speech or voice analysis techniques using neural networks

Table A2. Relevant CPC codes – RA-SYS Level 1 Expansion

CPC codes	Technological Area
A47L9/2852, A47L11/24, A47L11/4011, A47L11/4061, A47L2201/00, A47L2201/04	Robotic cleaning machines, accessories and methods for regulating their displacement
A61B34/30, A61B34/32, A61B34/35, A61B34/37, A61B34/70, A61B34/74, A61B34/76, A61B2034/2059, A61B2034/301, A61B2034/302, A61B2034/305	Computer-aided surgery comprising surgical robots operating autonomously, for telesurgery, mechanical position encoders, robotic arms
B05B13/0431	Robots or articulated arms for applying liquids or other fluent materials to surfaces of objects or other work by spraying
B25J5/00, B25J5/005, B25J5/007, B25J5/02	Manipulators mounted on wheels or on carriages
B25J9/00, B25J9/0003, B25J9/0006, B25J9/0009, B25J9/0081, B25J9/0084, B25J9/0087, B25J9/0093, B25J9/0096, B25J9/023, B25J9/04, B25J9/042, B25J9/046, B25J9/06, B25J9/08, B25J9/10, B25J9/102, B25J9/104, B25J9/126, B25J9/16, B25J9/1602, B25J9/1605, B25J9/1607, B25J9/161, B25J9/1612, B25J9/162, B25J9/1628, B25J9/163, B25J9/1633, B25J9/1638, B25J9/1641, B25J9/1651, B25J9/1653, B25J9/1656, B25J9/1661, B25J9/1664, B25J9/1666, B25J9/1669, B25J9/1671, B25J9/1674, B25J9/1676, B25J9/1679, B25J9/1682, B25J9/1684, B25J9/1687, B25J9/1689, B25J9/1692, B25J9/1694, B25J9/1697	Programme-controlled manipulators, e.g. home robots, exoskeletons, characterised by multiple movable arms, controlled using specific programmes such as fuzzy logic, neural networks, control loops, motion or trajectory planning
B25J11/00, B25J11/0005, B25J11/005, B25J11/008, B25J11/0085, B25J11/009	Manipulators not otherwise provided for, such as manipulators having means for high-level communication with users, e.g. speech generator, face recognition means, for mechanical processing tasks, for service tasks
B25J13/00, B25J13/003, B25J13/006, B25J13/02, B25J13/06, B25J13/08, B25J13/085, B25J13/086, B25J13/088, B25J13/089	Controls for manipulators by means of audio-responsive input, hand grip control means, by means of sensing devices
B25J15/00, B25J15/0009, B25J15/0019, B25J15/0052, B25J15/04, B25J15/0616, B25J15/08	Gripping heads, comprising multi-articulated fingers, e.g. resembling a human hand, multiple end effectors
B25J17/00, B25J17/02	Joints and wrist joints

B25J18/00	Arms
B25J19/00, B25J19/0025, B25J19/0029, B25J19/005, B25J19/0075, B25J19/02, B25J19/021, B25J19/022, B25J19/023, B25J19/04, B25J19/06	Accessories fitted to manipulators, e.g. for monitoring, for viewing, for sensing, for balancing, for safety
B60W60/0011	Drive control systems specially adapted for autonomous road vehicles, involving planning systems to avoid obstacles
B62D57/02, B62D57/024, B62D57/032	Vehicles characterised by having other propulsion or other ground- engaging means than wheels or endless track, such as ground-engaging propulsion means, e.g. walking members
B65G1/0492	Storage devices with cars adapted to travel in storage aisles
G05B19/4155, G05B19/4182, G05B19/42, G05B19/423, G05B19/425, G05B2219/39082, G05B2219/40202, G05B2219/40298, G05B2219/45083, G05B2219/50391	Programme-control systems characterised by programme execution, recording and playback systems, robotics vision and touch
G05D1/0016, G05D1/0022, G05D1/0027, G05D1/0038, G05D1/0044, G05D1/0088, G05D1/02, G05D1/021, G05D1/0212, G05D1/0214, G05D1/0217, G05D1/0219, G05D1/0221, G05D1/0223, G05D1/0225, G05D1/0227, G05D1/0231, G05D1/0234, G05D1/0238, G05D1/024, G05D1/0242, G05D1/0246, G05D1/0248, G05D1/0251, G05D1/0255, G05D1/0257, G05D1/027, G05D1/0272, G05D1/0274, G05D1/0276, G05D1/028, G05D1/0291, G05D1/0297, G05D2201/02, G05D2201/0203, G05D2201/0207, G05D2201/0208, G05D2201/0209, G05D2201/0211, G05D2201/0215, G05D2201/0216, G05D2201/0217	Control of position, course, altitude or attitude of land, water, air or space vehicles, e.g. using automatic pilots characterised by predetermined rules or autonomous decision making
G06N3/008	Computing arrangements based on physical entities controlled by simulated intelligence so as to replicate intelligent life forms, e.g. based on robots replicating pets or humans in their appearance or behaviour
Y10S901/01, Y10S901/02, Y10S901/09, Y10S901/46, Y10S901/47	Mobile robots, arm motion controller, closed loop systems, sensing and optical devices
Y10T74/20305, Y10T74/20317	Machine element or mechanism such as robotic arms, including electric motors

Table A3. LS-SYS and RA-SYS examples – inference results

Application ID	DOCDB family ID	Title	Abstract	Pos. Prob.	Pred. Class
524300358	68991627	Generation method and device of speech recognition model, and electronic equipment	The invention relates to a generation method and device of a speech recognition model, and electronic equipment for improving the accuracy and a recognition effect of model recognition. The method comprises the following steps of acquiring training samples, wherein each training sample comprises a speech frame sequence and a corresponding labeled text sequence; serving the speech frame sequence as an input characteristic of an encoder, serving a speech coding frame of the speech frame sequence as an output characteristic of the encoder, and training the encoder; and serving the speech coding frame as the input characteristic of a decoder, serving the labeled text sequence corresponding to the speech frame sequence as the output characteristic to train the decoder so as to obtain a current predicted text sequence, serving the speech coding frame as the input characteristic of the decoder sampling the labeled text sequence corresponding to the speech frame sequence and the current predicted text sequence according to a preset probability, serving the sequence obtained by combining the same as the output characteristic, and re-training the decoder.	0.988	LS-SYS (Level 1 Expansion)
441645254	53400372	Method for improving tracking using dynamic background compensation with centroid compensation	A method for tracking an object across a number of image frames comprises identifying a region containing the object in a first image frame to be stored as an exemplar view of the object. An appearance model (modified Exemplar View histogram is created based on the region in the first image frame and a background region in a second image frame, and the method determines at least one of a location and size of a predicted region for tracking the object in the second image frame using the appearance model. The method corrects at least one of the determined location and size of the predicted region in the second image frame in accordance with at least one of the location and size of the region in the first image frame corresponding to the exemplar view of the object.	0.816	LS-SYS (Level 2 Expansion)
489853140	61155532	Automatic screen state detecting robot, method and computer readable storage medium	The invention discloses an automatic screen state detecting robot. The automatic screen state detecting robot comprises a memory and a processor; the memory stores an automatic screen state detecting program; when the program is executed by the processor, the following steps are implemented: controlling a robot to move to each preset area of each piece of service equipment of an unmanned site area; if the robot moves into the preset area of a piece of service equipment, detecting whether a circuit fault of the display screen of the service equipment occurs; if the circuit fault of the display screen of the service equipment does not occur, controlling the service equipment to display an image according to preset display parameters, and analyzing the image displayed by the display screen soas to analyze that whether the abnormity of a preset type of the display screen of the service equipment occurs […].	0.974	RA-SYS (Level 1 Expansion)
457454093	51790809	Graph-based navigation using lighting effects	Methods, lighting control systems, mobile computing devices and computer-readable media are described herein for graph-based navigation through an environment. In various embodiments, a graph including a plurality of nodes and a plurality of edges may be provided, e.g., by a lighting control system, to mobile computing devices such as smart phones and tablets carried through the environment. Each node may correspond to a location of a lighting effect produced by a particular light source within the environment. Each edge may correspond to a path between two nodes. Data indicative of a path travelled through the graph by one or more mobile computing devices between a first node corresponding to a first location and a second node corresponding to a second location may be used to update the graph. Mobile computing devices may utilize the graph to calculate optimal paths and/or guide users through the environment.	0.73	RA-SYS (Level 2 Expansion)

Table A4. AI-related patents and keyword matches

Technological AI area	Keywords scoring 1	Keywords scoring 0.75	Keywords scoring 0.5
LS-SYS	89,967 (23.30%)	46,425 (12%)	7,169 (1.86%)
RA-SYS	85,568 (24.2%)	20,384 (5.77%)	2,240 (0.63%)

Note: For each AI domain, the table reports the number and percentage of model-classified AI patents whose titles or abstracts include at least one n-gram from our lexicon with a relevance score of 1, 0.75, or 0.5. The first column lists the two domains, while columns 2–4 show the absolute counts and shares of documents matching each keywords group.

Table A5. Sectoral classes

Sectoral class	Class Description	NACE Rev. 2 codes
Primary	Agriculture, forestry and fishing	01 to 03
	Mining and quarrying	05 to 09
Food, bev. & tob.	Manufacture of food products, beverages and tobacco products	10 to 12
Textile	Manufacture of textiles, apparel, leather and related products	13 to 15
Wood, paper & print	Manufacture of wood and paper products, and printing	16 to 18
Chemistry	Manufacture of coke, and refined petroleum products	19
	Manufacture of chemicals and chemical products	20
	Manufacture of rubber and plastics products	22
Pharma	Manufacture of pharmaceuticals, medicinal chemical and botanical products	21
Minerals	Manufacture of other non-metallic mineral products	23
Metal	Manufacture of basic metals	24
	Manufacture of fabricated metal products, except machinery and equipment	25
Computer & electr.	Manufacture of computer, electronic and optical products	26
	Manufacture of electrical equipment	27
Machinery	Manufacture of machinery and equipment not elsewhere classified	28
Transport equip.	Manufacture of motor vehicles, trailers and semi-trailers	29
	Manufacture of other transport equipment	30
Other manuf.	Other manufacturing, and repair and installation of machinery and equipment	31 to 33
Utilities	Electricity, gas, steam and air-conditioning supply	35
	Water supply, sewerage, waste management and remediation	36 to 39
Construction	Construction of building, civil engineering and specialised construction activities	41 to 43
Trade	Wholesale and retail trade, repair of motor vehicles and motorcycles	45 to 47
Transp. & storage	Transportation and storage	49 to 53
Accomod. & food	Accommodation and food service activities	55 and 56
ICT services	Publishing, audiovisual and broadcasting activities	58 to 60
	Telecommunications	61
	IT and other information services	62 and 63
Finance	Financial and insurance activities	64 to 66
Real estate	Real estate activities	68
Scien. & Profess.	Legal, accounting, management, architecture, engineering, technical testing and analysis activities	69 to 71
	Scientific research and development	72
	Other professional, scientific and technical activities	73 to 75
Admin.	Administrative and support service activities	77 to 82
Education	Education	85
Other Serv.	Public administration and defense, compulsory social security	84
	Human health services	86
	Residential care and social work activities	87 and 88
	Arts, entertainment and recreation	90 to 93
	Other services	94 to 96
	Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use	97 and 98
	Activities of extra-territorial organisations and bodies	99

Table A6. LS-SYS patent families by Applicant

Firm	Pat. Families	Firm	Pat. Families
International Business Machines Corp.	17,611	Robert Bosch GmbH	1,443
Samsung Electronics Co., LTD.	6,036	Baidu Netcom S&T Co., LTD.	1,368
Microsoft Corporation	4,905	Sap SE	1,360
Microsoft Technology Licensing, LLC	4,762	Meta Platforms, Inc.	1,324
Google LLC	4,618	Apple Inc.	1,229
NEC Corporation	2,840	LG Electronics Inc.	1,210
Siemens AG	2,797	Oracle International Corporation	1,199
Intel Corp.	2,736	AT&T Inc.	1,159
Fujitsu Limited	2,710	Nippon Telegraph and Telephone Corp.	1,144
Koninklijke Philips N.V.	2,598	Tencent technology Co., ltd.	1,124
Ping An Technology Co., LTD.	2,400	Bank of America Corporation	1,077
Huawei Technologies Co., LTD.	2,212	Korea Electronics & Telecom Co., LTD	977
Hitachi LTD.	2,127	Altaba Inc.	964
Sony Group Corporation	2,052	Alibaba Group Holding Limited	962
Toshiba Corporation	1,969	TATA Consultancy Services Limited	947
Canon Inc.	1,781	Cisco Technology, Inc.	905
Hewlett-Packard Development Co LP	1,774	Accenture Global Solutions Limited	904
Adobe Inc.	1,735	Nuance Communications, Inc.	816
General Electric Company	1,706	Nvidia Corporation	815
E.T.R.I.	1,693	LM Ericsson AB	782
Qualcomm Incorporated	1,598	Baidu Online Network Technology LTD.	772
Panasonic Holdings Corporation	1,514	Siemens Healthcare GmbH	769
Mitsubishi Electric Corporation	1,497	Fujifilm Business Innovation Corp.	748
Xerox Corporation	1,495	Capital One Services LLC	746
Amazon.com, Inc.	1,478	Mitsubishi Electric Research Laboratories	742

Note: The table displays the top 50 most active firms of LS-SYS patent families. By focusing on patent families rather than individual patent applications, this approach reduces inflation caused by family size and minimizes the influence of strategic patenting behavior. As a result, it offers a more accurate representation of firms actively developing technologies related to the “Learning and Symbolic Systems” domain.

Table A7. R-AS patent families by Assignee

Firm	Pat. Families	Firm	Pat. Families
Fanuc corporation	3,422	Nissan Motor Co., LTD.	908
Honda Motor Co., LTD.	2,848	Olympus Corporation	882
Toyota Motor Corporation	2,710	Koninklijke Philips N.V.	877
Samsung Electronics Co., LTD.	2,200	Huawei Technologies Co., LTD.	855
SZ DJI Technology Co., LTD.	2,101	Yaskawa Electric Corporation	842
The Boeing Company	2,063	Caterpillar Inc.	813
Hyundai Motor Company	2,024	Deere & Company	777
LG Electronics Inc.	1,994	Kawasaki Heavy Industries LTD	769
International Business Machines Corp.	1,964	Honeywell International Inc.	754
Robert Bosch GmbH	1,938	Applied Materials Inc.	739
Ford Global Technologies LLC	1,858	Google LLC	708
Ford Global Technologies Inc.	1,812	Komatsu MFG Co., LTD.	708
Siemens AG	1,776	Amazon.com, Inc.	687
Panasonic Holdings Corporation	1,679	Intel Corp.	682
Mitsubishi Electric Corporation	1,611	Hon Hai Precision Ind. Co., LTD.	677
Hitachi LTD	1,545	Volkswagen AG	672
Kia Corporation	1,408	Intuitive Surgical Operations, Inc.	667
GM Global Technology Operations	1,388	Fuji Corporation	666
General Electric Company	1,239	Honeywell Int.	655
Sony Group Corporation	1,225	Fujitsu Limited	655
Denso Corporation	1,085	Komatsu LTD.	650
Tokyo Electron Limited	978	Covidien LP	629
Canon Incorporated	960	Covidien Limited	620
Seiko Epson Corporation	927	Kia Co., LTD.	618
Toshiba Corporation	924	LG Electronics Co., LTD.	596

Note: The table displays the top 50 most active firms of RA-SYS patent families. By focusing on patent families rather than individual patent applications, this approach reduces inflation caused by family size and minimizes the influence of strategic patenting behavior. As a result, it offers a more accurate representation of firms actively developing technologies related to the “Robotics and Autonomous Systems” domain.

Table A8. Overlapping AI patent families by Assignee

Firm	Pat. Families	Firm	Pat. Families
International Business Machines Corp.	473	The Boeing Company	145
LG Electronics Inc.	457	Kia Corporation	142
Samsung Electronics Co., LTD.	414	Qualcomm Incorporated	139
GM Global Technology Operations	336	Baidu Netcom S&T Co., LTD.	134
Honda Motor Co., LTD.	314	Uber Advanced Technology Center LLC	134
Intel Corp.	277	Toyota Research Institute, Inc.	129
Baidu USA LLC	274	Uber Technologies, Inc.	128
Google LLC	268	Korea Electronics & Telecom Co., LTD.	126
Fanuc Corporation	264	Mitsubishi Electric Research Laboratories	123
Ford Global Technologies LLC	247	General Electric Company	123
Ford Global Technologies Inc.	243	Koninklijke Philips N.V.	116
Robert Bosch GmbH	238	Baidu Online Network Technology LTD.	115
Microsoft Technology Licensing, LLC	238	LG Electronics Co., LTD.	113
Hitachi LTD	205	Zoox, Inc.	111
Siemens AG	200	Apple Inc.	103
Hyundai Motor Company	196	Fujitsu Limited	99
Sony Group Corporation	195	General Motors Company	99
Mitsubishi Electric Corporation	193	Honeywell International Inc.	99
Microsoft Corporation	192	Toshiba Corporation	99
E.T.R.I.	189	HRL Laboratories, LLC	98
Toyota Motor Corporation	181	Amazon.com, Inc.	95
Waymo LLC	175	TATA Consultancy Services Limited	94
Huawei Technologies Co., LTD.	166	Rockwell Automation Technologies Inc.	94
X Development LLC	152	Argo AI Holdings, LLC	93
Nvidia Corporation	152	Panasonic Holdings Corporation	91

Note: The table displays the top 50 most active firms of overlapping AI patent families, defined as the intersection of the LS-SYS and RA-SYS domains. By focusing on patent families rather than individual patent applications, this approach reduces inflation caused by family size and minimizes the influence of strategic patenting behavior. As a result, it offers a more accurate representation of firms actively developing technologies pertaining to both AI domains.

Table A9. Relevant Learning and Symbolic Systems CPC codes

CPC 4-digit	Description	Share
G06F	Electric digital data processing	46.8%
G06N	Computing arrangements based on specific computational models	29%
G06V	Image or video recognition or understanding	17.5%
G06T	Image data processing or generation, in general	17.2%
G06Q	Information and communication technology [ICT] specially adapted for administrative, commercial, financial, managerial or supervisory purposes; systems or methods specially adapted for administrative, commercial, financial, managerial or supervisory purposes, not otherwise provided for	14%
H04L	Transmission of digital information, e.g. telegraphic communication	9.7%
G10L	Speech analysis techniques or speech synthesis; speech recognition; speech or voice processing techniques; speech or audio coding or decoding	8.7%
H04N	Pictorial communication, e.g. television	6.5%
G16H	Healthcare informatics, i.e. information and communication technology [ICT] specially adapted for the handling or processing of medical or healthcare data	5.9%
A61B	Diagnosis; surgery; identification	5.3%

Note: The table displays the ten most relevant 4-digit CPC codes within the overall sample of "Learning and Symbolic Systems" patent families resulting from the deep learning identification procedure. The share is computed as the ratio between the number of patent families being assigned to the focal CPC code and the total number of patent families. This classification mitigates inflation due to family size, and it is less influenced by strategic patenting activities, thereby providing a clearer depiction of technological classes related to the domain "Learning and Symbolic Systems".

Table A10. Relevant Robotics and Autonomous Systems CPC codes.

CPC 4-digit	Description	Share
B25J	Manipulators; chambers provided with manipulation devices	13.5%
G05D	Systems for controlling or regulating non-electric variables	12.7%
G05B	Control or regulating systems in general; functional elements of such systems; monitoring or testing arrangements for such systems or elements	10.1%
G06F	Electric digital data processing	9.6%
B60W	Conjoint control of vehicle sub-units of different type or different function; control systems specially adapted for hybrid vehicles; road vehicle drive control systems for purposes not related to the control of a particular sub-unit	8.2%
A61B	Diagnosis; surgery; identification	8.1%
G06T	Image data processing or generation, in general	7.4%
G06V	Image or video recognition or understanding	7.2%
G08G	Traffic control systems	7.1%
G01S	Radio direction-finding; radio navigation; determining distance or velocity by use of radio waves; locating or presence- detecting by use of the reflection or reradiation of radio waves; analogous arrangements using other waves	6.5%

Note: The table displays the ten most relevant 4-digit CPC codes within the overall sample of "Robotics and Autonomous Systems" patent families resulting from the deep learning identification procedure. The share is computed as the ratio between the number of patent families being assigned to the focal CPC code and the total number of patent families. This classification mitigates inflation due to family size, and it is less influenced by strategic patenting activities, thereby providing a clearer depiction of technological classes related to the domain "Robotics and Autonomous Systems".

Table A11. Relevant AI domain overlap CPC codes.

CPC 4-digit	Description	Share
G06F	Electric digital data processing	31.9%
G06V	Image or video recognition or understanding	29.8%
G06T	Image data processing or generation, in general	26.7%
G06N	Computing arrangements based on specific computational models	24.7%
G05D	Systems for controlling or regulating non-electric variables	19.2%
B60W	Conjoint control of vehicle sub-units of different type or different function; control systems specially adapted for hybrid vehicles; road vehicle drive control systems for purposes not related to the control of a particular sub-unit	15.1%
G05B	Control or regulating systems in general; functional elements of such systems; monitoring or testing arrangements for such systems or elements	14.9%
B25J	Manipulators; chambers provided with manipulation devices	12.4%
G06Q	Information and communication technology [ICT] specially adapted for administrative, commercial, financial, managerial or supervisory purposes; systems or methods specially adapted for administrative, commercial, financial, managerial or supervisory purposes, not otherwise provided for	10%
G01C	Measuring distances, levels or bearings; surveying; navigation; gyroscopic instruments; photogrammetry or videogrammetry	9.9%

Note: The table displays the ten most relevant 4-digit CPC codes within the sample of "domain overlap" patent families, defined as the intersection of the LS-SYS and RA-SYS domains. The share is computed as the ratio between the number of patent families being assigned to the focal CPC code and the total number of patent families. This classification mitigates inflation due to family size, and it is less influenced by strategic patenting activities, thereby providing a clearer depiction of technological classes related to the domain patents that span both domains.

Working Papers

Dipartimento di Politica Economica

1. *Innovation, jobs, skills and tasks: a multifaceted relationship.* M. Piva, M. Vivarelli. Vita e Pensiero, maggio 2018 (ISBN 978-88-343-3654-0)
2. *A bridge over troubled water: Interdisciplinarity, Novelty, and Impact.* M. Fontana, M. Iori, F. Montobbio, R. Sinatra. Vita e Pensiero, settembre 2018 (ISBN 978-88-343-3793-6)
3. *Concordance and complementarity in IP instruments*. M. Grazzi, C. Piccardo, C. Vergari. Vita e Pensiero, gennaio 2019 (ISBN 978-88-343-3879-7)
4. *Sustainable finance, the good, the bad and the ugly: a critical assessment of the EU institutional framework for the green transition*. L. Esposito, E.G. Gatti, G. Mastromatteo. Vita e Pensiero, febbraio 2019 (ISBN 978-88-343-3892-6)
5. *Technology and employment in a vertically connected economy: a model and an empirical test.* G. Dosi, M. Piva, M.E. Virgillito, M. Vivarelli. Vita e Pensiero, giugno 2019 (ISBN digital edition [PDF]: 978-88-343-4008-0)
6. *Testing the employment impact of automation, robots and AI: A survey and some methodological issues.* L. Barbieri, C. Mussida, M. Piva, M. Vivarelli. Vita e Pensiero, settembre 2019 (ISBN digital edition [PDF]: 978-88-343-4052-3)
7. *A new proposal for the construction of a multi-period/multilateral price index*. C.R. Nava, A. Pesce, M.G. Zoia. Vita e Pensiero, ottobre 2019 (ISBN digital edition [PDF]: 978-88-343-4114-8)
8. *Lo Stato Sociale: da "lusso" a necessità*. L. Campiglio. Vita e Pensiero, febbraio 2020 (ISBN digital edition [PDF]: 978-88-343-4184-1)
9. *Robots and the origin of their labour-saving impact*. F. Montobbio, J. Staccioli, M.E. Virgillito, M. Vivarelli. Vita e Pensiero, marzo 2020 (ISBN digital edition [PDF]: 978-88-343-4196-4)
10. *Business visits, technology transfer and productivity growth*. M. Piva, M. Tani, M. Vivarelli. Vita e Pensiero, marzo 2020 (ISBN digital edition [PDF]: 978-88-343-4210-7)
11. *Technology, industrial dynamics and productivity: a critical survey*. M. Ugur, M. Vivarelli. Vita e Pensiero, settembre 2020 (ISBN digital edition [PDF]: 978-88-343-4406-4)
12. *Back to the past: the historical roots of labour-saving automation*. J. Staccioli, M.E. Virgillito. Vita e Pensiero, novembre 2020 (ISBN digital edition [PDF]: 978-88-343-4473-6)
13. *The present, past, and future of labor-saving technologies.* J. Staccioli, M.E. Virgillito. Vita e Pensiero, dicembre 2020 (ISBN digital edition [PDF]: 978-88-343-4479-8)
14. *Why Do Populists Neglect Climate Change? A Behavioural Approach*. L.A. Lorenzetti. Vita e Pensiero, dicembre 2020 (ISBN digital edition [PDF]: 978-88-343-4483-5)
15. *Relative wages, payroll structure and performance in soccer. Evidence from Italian Serie A (2007-2019)*. C. Bellavite Pellegrini, R. Caruso, M. Di Domizio. Vita e Pensiero, gennaio 2021 (ISBN digital edition [PDF]: 978-88-343-4490-3)
16. *Robots, AI, and Related Technologies: A Mapping of the New Knowledge Base*. E. Santarelli, J. Staccioli, M. Vivarelli. Vita e Pensiero, gennaio 2021 (ISBN digital edition [PDF]: 978-88-343-4499-6)
17. *Detecting the labour-friendly nature of AI product innovation*. G. Damioli, V. Van Roy, D. Vertesy, M. Vivarelli. Vita e Pensiero, aprile 2021 (ISBN digital edition [PDF]: 978-88-343-4600-6)
18. *Circular Economy Approach: The benefits of a new business model for European Firms*. C. Bellavite Pellegrini, L. Pellegrini, C. Cannas. Vita e Pensiero, luglio 2021 (ISBN digital edition [PDF]: 978-88-343-4817-8)
19. *The impact of cognitive skills on investment decisions. An empirical assessment and policy suggestions.* L. Esposito, L. Marrese. Vita e Pensiero, luglio 2021 (ISBN digital edition [PDF]: 978-88-343-4822-2)
20. *"Thinking of the end of the world and of the end of the month": the Impact of Regenerative Agriculture on Economic and Environmental Profitability*. L.A. Lorenzetti, A. Fiorini. Vita e Pensiero, ottobre 2021 (ISBN digital edition [PDF]: 978-88-343-4898-7)

21. *Labour-saving automation and occupational exposure: a text-similarity measure*. F. Montobbio, J. Staccioli, M.E. Virgillito, M. Vivarelli. Vita e Pensiero, novembre 2021 (ISBN digital edition [PDF]: 978-88-343-5089-8)

22. *Climate reputation risk and abnormal returns in the stock markets: a focus on large emitters*. G. Guastella, M. Mazzarano, S. Pareglio, A. Xepapadeas. Vita e Pensiero, novembre 2021 (ISBN digital edition [PDF]: 978-88-343-5092-8)

23. *Carbon Boards and Transition Risk: Explicit and Implicit exposure implications for Total Stock Returns and Dividend Payouts*. M. Mazzarano, G. Guastella, S. Pareglio, A. Xepapadeas. Vita e Pensiero, novembre 2021 (ISBN digital edition [PDF]: 978-88-343-5093-5)

24. *Innovation and employment: a short update*. M. Vivarelli. Vita e Pensiero, gennaio 2022 (ISBN digital edition [PDF]: 978-88-343-5113-0)

25. *AI technologies and employment. Micro evidence from the supply side*. G. Damioli, V. Van Roy, D. Vertesy, M. Vivarelli. Vita e Pensiero, gennaio 2022 (ISBN digital edition [PDF]: 978-88-343-5119-2)

26. *The Effect of External Innovation on Firm Employment*. G. Arenas Díaz, A. Barge-Gil, J. Heijs, A. Marzucchi. Vita e Pensiero, febbraio 2022 (ISBN digital edition [PDF]: 978-88-343-5146-8)

27. *The North-South divide: sources of divergence, policies for convergence*. L. Fanti, M.C. Pereira, M.E. Virgillito. Vita e Pensiero, maggio 2022 (ISBN digital edition [PDF]: 978-88-343-3524-4)

28. *The empirics of technology, employment and occupations: lessons learned and challenges ahead*. F. Montobbio, J. Staccioli, M.E. Virgillito, M. Vivarelli. Vita e Pensiero, novembre 2022 (ISBN digital edition [PDF]: 978-88-343-5383-7)

29. *Cognitive biases and historical turns. An empirical assessment of the intersections between minds and events in the investors' decisions*. L. Esposito, L. Malara. Vita e Pensiero, gennaio 2023 (ISBN digital edition [PDF]: 978-88-343-5420-9)

30. *Interaction between Ownership Structure and Systemic Risk in the European financial sector*. C. Bellavite Pellegrini, R. Camacci, L. Pellegrini, A. Roncella. Vita e Pensiero, febbraio 2023 (ISBN digital edition [PDF]: 978-88-343-5446-9)

31. *Was Robert Gibrat right? A test based on the graphical model methodology*. M. Guerzoni, L. Riso, M. Vivarelli. Vita e Pensiero, marzo 2023 (ISBN digital edition [PDF]: 978-88-343-5457-5)

32. *A North-South Agent Based Model of Segmented Labour Markets. The Role of Education and Trade Asymmetries*. L. Fanti, M.C. Pereira, M.E. Virgillito. Vita e Pensiero, maggio 2023 (ISBN digital edition [PDF]: 978-88-343-5529-9)

33. *Innovation and the Labor Market: Theory, Evidence and Challenges*. N. Corrocher, D. Moschella, J. Staccioli, M. Vivarelli. Vita e Pensiero, giugno 2023 (ISBN digital edition [PDF]: 978-88-343-5580-0)

34. *The Effect of Economic Sanctions on World Trade of Mineral Commodities. A Gravity Model Approach from 2009 to 2020*. R. Caruso, M. Cipollina. Vita e Pensiero, dicembre 2023 (ISBN digital edition [PDF]: 978-88-343-5686-9)

35. *Education and Military Expenditures: Countervailing Forces in Designing Economic Policy. A Contribution to the Empirics of Peace*. A. Balestra, R. Caruso. Vita e Pensiero, gennaio 2024 (ISBN digital edition [PDF]: 978-88-343-5757-6)

36. *Vulnerability to Climate Change and Communal Conflicts: Evidence from Sub-Saharan Africa and South/South-East Asia*. S. Balestri, R. Caruso. Vita e Pensiero, maggio 2024 (ISBN digital edition [PDF]: 978-88-343-5829-0)

37. *Assessing changes in EU innovation policy programs: from SME instrument to EIC accelerator for start-up funding*. M. del Sorbo, C. Faber, M. Grazzi, F. Matteucci, M. Ruß. Vita e Pensiero, luglio 2024 (ISBN digital edition [PDF]: 978-88-343-5860-3)

38. *AI as a new emerging technological paradigm: evidence from global patenting.* G. Damioli, V. Van Roy, D. Vertesy, M. Vivarelli. Vita e Pensiero, settembre 2024 (ISBN digital edition [PDF]: 978-88-343-5873-3)

39. *The KSTE+I approach and the AI technologies*. F. D'Alessandro, E. Santarelli, M. Vivarelli. Vita e Pensiero, settembre 2024 (ISBN digital edition [PDF]: 978-88-343-5880-1)

40. *Quo Vadis Terra? The future of globalization between trade and war*. L. Esposito, E.G. Gatti, G. Mastromatteo. Vita e Pensiero, settembre 2024 (ISBN digital edition [PDF]: 978-88-343-5895-5)

41. *The Agents of Industrial Policy and the North-South Convergence: State-Owned Enterprises in an International-Trade Macroeconomic ABM.* L. Fanti, M.C. Pereira, M.E. Virgillito. Vita e Pensiero, ottobre 2024 (ISBN digital edition [PDF]: 978-88-343-5909-9)

42. *The impact of US elections on US defense industry: Firm-level evidence from 1996 to 2022*. A. Balestra, R. Caruso. Vita e Pensiero, January 2025 (ISBN digital edition [PDF]: 978-88-343-5937-2)

43. *Forecasting the Impact of Extreme Weather Events on Electricity Prices in Italy: A GARCH-MIDAS Approach with Enhanced Variable Selection*. M. Guerzoni, L. Riso, M.G. Zoia. Vita e Pensiero, January 2025 (ISBN digital edition [PDF]: 978-88-343-5938-9)

44. *The Theoretical Properties of Novel Risk-Based Asset Allocation Strategies using Portfolio Volatility and Kurtosis*. M.D. Braga, L. Riso, M.G. Zoia. Vita e Pensiero, January 2025 (ISBN digital edition [PDF]: 978-88-343-5939-6)

45. *Sustainable Finance in the New Geo-Political Era: A Difficult Balancing Act*. L. Esposito, M. Cocco. Vita e Pensiero, February 2025 (ISBN digital edition [PDF]: 978-88-343-5940-2)

46. *New technologies and employment: the state of the art*. M. Vivarelli, G. Arenas Díaz. Vita e Pensiero, March 2025 (ISBN digital edition [PDF]: 978-88-343-5941-9)

47. *Leveraging Knowledge Networks: Rethinking Technological Value Distribution in mRNA Vaccine Innovations*. R. Mastrandrea, F. Montobbio, G. Pellegrino, M. Riccaboni, V. Sterzi. Vita e Pensiero, March 2025 (ISBN digital edition [PDF]: 978-88-343-5991-4)

48. *A Twin Transition or a policy flagship? Emergent constellations and dominant blocks in green and digital technologies*. L. Nelli, M.E. Virgillito, M. Vivarelli. Vita e Pensiero, April 2025 (ISBN digital edition [PDF]: 978-88-343-5992-1)

49. *The role of business visits in fostering R&D investment*. M. Vivarelli, M. Piva, M. Tani. Vita e Pensiero, June 2025 (ISBN digital edition [PDF]: 978-88-343-5993-8)

50. *A Deep Learning procedure for the identification of Artificial Intelligence technologies in patent data*. F. D'Alessandro. Vita e Pensiero, June 2025 (ISBN digital edition [PDF]: 978-88-343-5994-5)