The Journal of Technology Transfer
https://doi.org/10.1007/s10961-025-10189-8

RESEARCH

The artificial intelligence patent dataset (AIPD) 2023 update

Nicholas A. Pairolero[1] · Alexander V. Giczy[1,2] · Gerard Torres[3] · Tisa Islam Erana[3] · Mark A. Finlayson[1,3] · Andrew A. Toole[1]

Accepted: 21 January 2025

Abstract

The 2023 update to the Artificial Intelligence Patent Dataset (AIPD) extends the original AIPD to all United States Patent and Trademark Office (USPTO) patent documents (i.e., patents and pre-grant publications, or PGPubs) published through 2023, while incorporating an improved patent landscaping methodology to identify AI within patents and PGPubs. This new approach substitutes BERT for Patents for the Word2Vec embeddings used previously, and uses active learning to incorporate additional training data closer to the "decision boundary" between AI and not-AI to help improve predictions. We show that this new approach achieves substantially better performance than the original methodology on a set of patent documents where the two methods disagreed—on this set, the AIPD 2023 achieved precision of 68.18 percent and recall of 78.95 percent, while the original AIPD achieved 50 percent and 21.05 percent, respectively. To help researchers, practitioners, and policy-makers better understand the determinants and impacts of AI invention, we have made the AIPD 2023 publicly available on the USPTO's economic research web page.

Keywords Patent · Patent landscape · Artificial intelligence · AI · AI patent dataset · AIPD

✉ Nicholas A. Pairolero
Nicholas.Pairolero@uspto.gov

✉ Andrew A. Toole
Andrew.Toole@uspto.gov

Alexander V. Giczy
Alexander.Giczy1@USPTO.GOV

Gerard Torres
Gerard.Torres@USPTO.GOV

Tisa Islam Erana
tisla016@fiu.edu

Mark A. Finlayson
markaf@fiu.edu

1 United States Patent and Trademark Office, Alexandria, USA

2 Addx Corporation, Alexandria, USA

3 Florida International University, Miami, USA

Published online: 22 February 2025

JEL Classification O31 · O34 · C45 · L86

1 Introduction

The Artificial Intelligence Patent Dataset (AIPD) was publicly released by the United States Patent and Trademark Office (USPTO) in 2021 (Giczy et al., 2022). Since its release, the AIPD has significantly contributed to the understanding of artificial intelligence (AI) invention by encouraging research into both its determinants and impacts (e.g., Toole et al., 2020; Chattergoon & Kerr, 2022; Chowdhury et al., 2022; Gaske, 2023; Gomes et al., 2023; Liu et al., 2023; Park, 2024; Giczy et al., 2024; Gao et al., 2024; Rathi et al., 2024), and calling attention to the significant challenges associated with identifying AI within patent documents (Grashof & Kopka, 2023; Hotte et al., 2022; Montobbio et al., 2023). Subsequent to the creation of the AIPD, researchers have developed several new methodologies for identifying technologies disclosed in patent documents (Choi et al., 2022; Islam Erana & Finlayson, 2024; Krestel et al., 2021; Pelaez et al., 2024; Pujari et al., 2022; Yoo et al., 2023), allowing us to improve our approach while extending the original dataset to include all patent documents published through 2023.

The 2023 update of the AIPD (hereinafter called the "AIPD 2023") identifies which of 15.4 million U.S. patent documents (patents and pre-grant publications, or PGPubs) published from 1976 through 2023 contain AI (separately identified for the eight AI component technologies from the AIPD, including machine learning, vision, natural language processing, speech, evolutionary computation, AI hardware, knowledge processing, and planning and control). The update includes an additional 2.2 million patent documents published since January 2021 that were not included in the original 2021 release,[1] and has been publicly released on the USPTO's economic research webpage (https://www.uspto.gov/ip-policy/economic-research/research-datasets/artificial-intelligence-patent-dataset).

The original AIPD achieved state of the art performance relative to several benchmarks in the literature that were available at the time of its release (Giczy et al., 2022). The AIPD 2023 was created from the original AIPD framework and incorporates several improvements from the recent patent landscaping literature. For example, we now incorporate BERT for Patents (Devlin et al., 2018; Srebrovic & Yonamine, 2020) into our machine learning architecture (originally based on Abood & Feltenberger, 2018 and extended in Giczy et al., 2022 and Islam Erana & Finlayson, 2024). Additionally, we overcome a limitation of the Abood and Feltenberger (2018) "expansion method" used to create the training dataset for the original AIPD (Giczy et al., 2022) by including training observations closer to the "decision boundary" of AI and not-AI, thereby enabling the model to learn from patent documents that are more difficult to classify.[2] These observations were manually labeled and selected via an active learning model that sampled patent documents from close to the 50 percent prediction threshold (i.e., from the set of observations where the

[1] The AIPD 2023 dataset does not include 14,140 patent documents that were in the previous AIPD 2021 release. These documents were not included for several reasons, including that some were granted patents and PGPub that have since been withdrawn. See Supplementary Information E for information regarding withdrawn patents and PGPubs.

[2] Given an input consisting of likely true positive observations, the "expansion method" finds negative observations by randomly sampling from those that are far away from the true positives (i.e., to increase the likelihood that the observations are actually true negatives), leaving little interior training data from which the model can learn the true decision boundary.

model was most uncertain). Islam Erana and Finlayson (2024) shows the benefits of adopting these new approaches within the original AIPD framework. The final major change is that we've enriched the training dataset by including patent documents published since 2018 that express the new ways AI has been used to augment invention since the creation of the original AIPD.

Given the large number of differences between the AIPD 2023 and the original approach, we carefully analyzed the set of "disagreements" between the models. Overall, the number of disagreements varied across the AI component technologies, ranging from 123,810 patent documents in speech to 264,618 in machine learning and 809,066 in AI hardware. Notably, the disagreements far outnumbered those documents where both models agreed that the inventions contained AI. For example, 70.21 percent of the documents where at least one of the models predicted machine learning were disagreements. To better understand which model was "right" more often, we compared predictions for 1000 patents documents published in 2019 and manually reviewed and labeled 229 disagreements at the AI technology component level. In all but one of the eight components the AIPD 2023 achieved higher precision and recall on the manually reviewed documents, leading to greater F1 scores. When considered at the aggregate AI level (i.e., disagreements in at least one AI technology component), the AIPD 2023 achieved precision of 68.18 percent and recall of 78.95 percent, while the original AIPD achieved 50 percent and 21.05 percent, respectively.

Even though the AIPD 2023 component technology models have better F1 scores than the original approach, the AIPD 2023 produces a substantially greater number of AI predictions each year, which is consistent with training and evaluation metrics for each component technology that favor higher recall at the expense of precision. For researchers seeking greater continuity with the original AIPD, or those that prefer greater precision at the expense of recall (Grashof & Kopka, 2023), we show that increasing the AIPD 2023 prediction threshold for determining AI can produce an AI prediction volume that closely matches the original approach.[3] Further, we identify a threshold estimate in the AIPD 2023 for balancing precision and recall, which when used produces a more accurate estimate of the volume of AI. The AIPD 2023 release contains the raw model prediction scores, allowing researchers to choose the prediction threshold and thus the level of precision and recall that is most appropriate for their application. In addition, the dataset includes binary variables for several thresholds, including 50 percent, the threshold for balancing precision and recall, and the estimate that best reproduces the volume of AI from the original AIPD's 50 percent threshold.

As with the original dataset, our testing revealed that the AIPD 2023 is better for certain AI technology components than others. For example, the new predictions for evolutionary computation are substantially worse than those for the other AI component technologies (as revealed by model training metrics), a feature of the dataset that has not changed since the original AIPD. There are likely too few patent documents containing evolutionary computation in our training dataset to produce reliable predictions. Additionally, the AIPD 2023 model for AI hardware (i.e., hardware that is specifically designed to improve AI computation) achieved both worse precision and recall than the original AIPD in our manual evaluation. Although there are many potential reasons for this, one possibility is that

[3] Given a patent document, each AIPD component technology model produces a prediction, which can be interpreted as a probability between 0 and 1 (with 1 indicating AI and 0 indicating not-AI). The prediction threshold is the probability for which all predictions above it will be labeled AI and all below will be labeled not-AI.

our new training dataset contains annotations from several different reviewers, and labeling patent documents in AI hardware is difficult. AI software inventions are often described as being embedded in a physical hardware system, and general-purpose hardware improvements may improve AI computation as well as computation more generally. These nuances could make it more difficult for humans to consistently label AI hardware, and therefore reduce the overall quality of the predictions.

The article proceeds as follows: first, we provide a brief overview of the model used to produce the original AIPD, and describe the literature that uses this dataset or relies on the article that describes it, Giczy et al. (2022). Second, we identify the differences between the approach used for the AIPD 2023 relative to the original AIPD model, followed by a description of the evaluation sample and the performance results obtained from it. Next, we provide several extensions, which include an analysis of the impact of adjusting the prediction threshold for AI, and more information on the set of disagreements between the new and original approaches. We conclude by describing several practical challenges associated with implementing a machine learning approach such as the one we used, as well as highlighting potentially promising areas for future research in this area. More information on the dataset, including how it may be used with publicly available patent data from PatentsView,[4] is available in the Supplementary Information.

2 Background

2.1 Definition of AI

We adopt the eight component technology-based definition of AI from the original AIPD (Giczy et al., 2022). Giczy et al. (2022) provides detailed descriptions of each component, which we summarize here. The first component technology, *machine learning*, contains computational methods that learn from data. The second, *computer vision,* extracts information from visual inputs to understand images and videos, and the third, *speech,* processes audio to understand sequences of words. The field of *evolutionary computation*, the fourth component, consists of biologically inspired methods, like genetic algorithms, which are procedures that improve themselves by selecting the best alternative from a set of randomly generated mutations. The fifth component, *knowledge processing,* consists of approaches to represent information and extract new facts from existing knowledge bases. The field of *natural language processing* is the sixth component technology and contains methods to understand language. *AI hardware* includes physical hardware designed to execute AI software, and finally, the field of *planning and control* consists of systems that generate plans to reach specified goals.

As described in Giczy et al. (2022) and Toole et al. (2019), our definition of AI is broad and encompasses earlier and more general technologies beyond the deep learning and large language models that are currently most associated with AI. However, for researchers interested in a narrower definition of AI, the dataset can be subset to only include AI from specific component technologies (like machine learning).

[4] PatentsView is a publicly accessible data visualization platform supported by the USPTO's Office of the Chief Economist that contains several research datasets on U.S. patents and PGPubs (see https://patentsview.org/).

2.2 Original AIPD methodology

The original AIPD was created using a multi-step deep learning approach based on the automated patent landscaping methodology of Abood and Feltenberger (2018). In the first step, patent classification/keyword queries were created to identify patent documents within each of the eight AI component technologies. The queries were designed to maximize precision, or the likelihood that retrieved documents were truly in the component technologies. To construct these queries, Giczy et al. (2022) intersected the set of documents returned as potential positives from four classification systems: United States Patent Classification (USPC), Cooperative Patent Classification (CPC), International Patent Classification (IPC), and the Clarivate Derwent World Patent Index.[5] This approach was used since a document identified as being in a given component technology by each of the classification systems would have a higher likelihood of being a true positive than a document only returned by one of the classification systems. The documents returned by the component technology queries formed the positive example "seed sets."

Next, the expansion method of Abood and Feltenberger (2018) was used to identify negative example "anti-seed" documents. This expansion approach used technology classifications, citations, and patent family[6] relationships to find documents that were "far enough away" from the seed documents to be likely true negatives.[7] Observations in the anti-seed sets did not cite seed documents, were not related to seed documents through family relationships, and had classifications that were generally different than those associated with seed documents. The seed and anti-seed sets formed the training datasets for each AI component technology model. Giczy et al. (2022) manually evaluated the quality of the seed and anti-seed sets, finding high accuracy in both sets (around 92%).

The deep learning architecture was based on the best performing model of Abood and Feltenberger (2018), consisting of separate long short-term memory (LSTM) neural networks for patent application claims and abstracts (using Word2Vec for the 300-dimensional word embedding inputs) and a dense neural network to process patent citations (both backward and forward) (which were one-hot encoded as inputs). The outputs of these separate networks, one for abstracts, claims and citations, were then concatenated into a 664-dimensional vector and input into a network consisting of several additional neural layers. Figure 4 of Giczy et al. (2022) provides a schematic of this architecture.

Giczy et al. (2022) showed that the approach used in the original AIPD achieved superior performance when identifying AI relative to alternatives in the literature, based on a holdout set of 368 patent documents that were manually annotated by USPTO patent

[5] For more information on the Derwent World Patent Index, see https://clarivate.com/intellectual-property/patent-intelligence/derwent-world-patents-index/.

[6] We use the "national family" definition of patent families, which groups applications emanating from a given jurisdiction that share at least one priority. More information is available at https://www.wipo.int/edocs/mdocs/aspac/en/wipo_ip_bkk_12/wipo_ip_bkk_12_www_238983.pdf

[7] The Abood and Feltenberger (2018) expansion method expands away from a seed set in two stages. The first stage, called the L1 expansion, identifies all family members of seed documents, then all the forward and backward citations of this enlarged set of documents, and then once again the family members of the resulting documents in the previous citation step (family-citation-family). In addition, L1 conducts a CPC-expansion from the seed set, which identifies all documents with CPC codes whose relative frequency in the seed set is at least 50 times the relative frequency of the codes within the entire set of patent documents. The second stage, called the L2 expansion, begins with the L1 documents and performs the citation and family expansions on them. Anti-seed documents are randomly sampled from all remaining documents not in the seed, L1, and L2 sets.

examiners. More information on the methodology and evaluation of the original AIPD is available in Giczy et al. (2022).

2.3 Use of the AIPD

Since the release of the AIPD in 2021, the dataset has been downloaded 5226 times, and the article describing the AIPD (Giczy et al., 2022) has been referenced over 50 times by a variety of studies in the economics, management, computer science, and legal literatures.[8] Some of these studies have used the AIPD directly, while others have used information in Giczy et al. (2022) to inform their research methods or as a resource for supporting material. Table A1 in Supplementary Information A shows that these uses are the primary ways the AIPD has been used, with the addition of several articles that benchmark the AIPD against other AI classification methods to assess how input datasets on AI affect applied results (e.g., like the degree to which AI is a "general purpose technology" or GPT) (Hötte et al., 2022; 2023; 2024).

Beyond scientific impact, the AIPD has been used broadly to stimulate policy discussions between the U.S. Federal Government and various stakeholders, including at several events associated with the USPTO's AI and Emerging Technology Partnership,[9] as well as the Office of the Director of National Intelligence Science and Technology Partnership.[10] Additionally, the AIPD was used in the USPTO's 2022 report to Congress on patent eligible subject matter in the United States (Vidal, 2022) to document how recent changes in patent law might affect upstream AI investments that support invention, as well as downstream innovation and commercialization opportunities in AI (Frumkin et al., 2024; Toole & Pairolero, 2020). The release of the AIPD 2023 should continue to support this research and policy activity by improving the quality of the underlying dataset and extending it through the end of 2023.

3 AIPD 2023 methodology

To create the AIPD 2023, we used the same machine learning approach as the original release but incorporated several improvements. Table C1 in the Supplementary Information briefly summarizes these changes, which center on an improved algorithmic framework and better training data.

The first improvement is that the machine learning models now use BERT for Patents (Srebrovic & Yonamine, 2020) rather than Word2Vec text embeddings. In a 2018 article, researchers at Google showed that BERT out-performed existing approaches, including Word2Vec, on several natural language processing benchmark tasks (Devlin et al., 2018). Within the patent landscaping context, Islam Erana and Finlayson (2024) shows that BERT for Patents achieves superior performance over Word2vec by a significant margin when incorporated into the machine learning architecture used to produce the AIPD (Abood &

[8] As of May, 2024.

[9] For more information, see https://www.uspto.gov/initiatives/artificial-intelligence/ai-and-emerging-technology-partnership-engagement-and-events.

[10] For more information, see https://www.dni.gov/index.php/who-we-are/organizations/policy-capabilities/in-step-the-intelligence-science-technology-partnership.

Feltenberger, 2018; Giczy et al., 2022).[11,12] Figure C1 in Supplementary Information C illustrates the updated AIPD 2023 model architecture.

The second major improvement over the original AIPD is updated training data. We used the original AIPD training data as a base, but extended it by: (1) adding newly labeled data closer to the "decision boundary" between AI and not-AI, (2) adding patent documents that were manually labeled by USPTO patent examiners when evaluating the original AIPD, and (3) adding AI patent documents published after 2019. Figure B1 in the Supplementary Information illustrates the updated training data construction.

The "decision boundary" documents were selected using active learning via a support vector machine (SVM) supervised machine learning model to identify and annotate documents that were close to the 50% prediction threshold between AI and not-AI (i.e., those documents for which the active learning model was the most uncertain) (Islam Erana & Finlayson, 2024). These documents were from years 1976–2018, with 90% of the documents being from 2018. Graduate students in AI at Florida International University (FIU) annotated this data set, resulting in 1147 documents across the eight AI component technologies.[13] The number of patent documents used for training from this source is shown in Table 1 in the "Decision Boundary" columns.

We also included in the training data the 800 patent documents that were previously annotated by USPTO examiners during the original AIPD evaluation. These documents were randomly sampled from the AIPD: 216 from the original seed training set, 216 from the original anti-seed training set, and 368 from all patent documents not in the seed or anti-seed sets. USPTO patent examiners specialized in AI labeled which of these 800 documents contained each of the eight AI component technologies (more information on this process is available in Giczy et al., 2022). For training the updated AIPD models we selected only those patent documents where two annotators agreed on whether the document was AI in the component technology or not, i.e., we did not include those patent documents that required adjudication by a third examiner. The number of training documents from this source is shown in the "Examiner Annotated" columns of Table 1.

While incorporating the decision boundary annotations into the training data improves model performance (Islam Erana & Finlayson, 2024), the most recent document in this training set was published in 2018. To capture the new ways AI has been used in invention

[11] BERT has also been shown to improve model performance across a variety of other patent related tasks, including prior art search (Chikkamath et al., 2024; Vowinckel & Hahnke, 2023) and citation prediction (Ghosh et al., 2024).

[12] Additionally, we removed the citation part of the deep learning model architecture. Citations were previously added to the model via "one-hot encoding," i.e., each citation corresponded to an element in an n-dimension vector, where the element was a 1 if a document included a given citation and 0 otherwise. Consistent with Abood and Feltenberger (2018), the maximum dimension of this vector was set to 50,000, which corresponded to 50,000 distinct citations. Given that the 78,654 patents in our training dataset have citations to 760,564 distinct U.S. patents (a number that does not include citations to PGPubs without patents or forward citations), the maximum of 50,000 would not have come close to covering the total (less than 6.6 percent); moreover, increasing the maximum dimension would have complicated the neural network. Islam Erana and Finlayson (2023) also excluded citations in its comparisons. It did, however, incorporate CPC codes of cited patent documents, a promising approach that groups citations by technology area but a feature that we did not include in our models.

[13] Using the training dataset from the original AIPD and a small set of positive and negative examples labeled by FIU researchers to initiate the active learning model, the SVM was retrained every 10 new annotations selected near the 50 percent prediction threshold (using the uncertainty sampling method of Lewis and Gale 1994) to continually improve its understanding of the decision boundary. More information on this procedure is available in Lewis and Gale (1994) and Islam Erana and Finlayson (2024).

since then, we added additional positive observations published from 2019 through 2023 to the seed set. These documents were obtained from search queries updated from the ones used for the original AIPD.[14] The queries were designed to be narrow, i.e., with very high precision. Moreover, to be conservative and not to overwhelm the previous training data, each seed set was increased by only 10 percent.[15] Table B1 in Supplementary Information B provides the queries used to add these additional documents to the seed sets, and Table B2 shows how many were added using this approach.

The training data summarized in Table 1 is imbalanced across several dimensions. First, there were differing numbers of positive and negative observations in each AI component technology. Second, the decision boundary documents and examiner annotations were far outnumbered by the seed/anti-seed, even though the former two may contain more information about how to classify AI. We accounted for these imbalances by weighting the observations during training, as specified in Table 1, so that each group of documents (regardless of the number of documents in them) received approximately equal weight.[16]

As with the original AIPD, we trained one model for each AI component technology.[17] To estimate performance, each model was trained five times using an 80/20 train/test split.[18] We averaged the resulting "test" performance metrics over the five runs by training epoch (i.e., the number of complete passes through the data during training) and used this information to determine the optimal number of training epochs to use for the final models (using all the training data).[19] Table 2 shows both validation (Panel a) and final model (Panel b) metrics (accuracy, precision, recall, and the F1 measure), as well as the number of epochs used to train the final models ("Epoch" column in Panel b). As expected, the validation F1 scores (Panel a) are usually lower than the final model scores (Panel b) but the differences are not generally substantial. In most component technologies, such as machine learning and natural language processing, final model F1 scores are within two standard deviations of the validation metrics, but in some others, such as planning and control, the differences are larger.

As seen in Panel b of Table 2, the final training F1 scores ranged from a high of 0.947 for speech, to a low of 0.306 for evolutionary computation. Similar to the original AIPD, evolutionary computation continues to be a challenging component technology to identify

[14] The original queries used to define the seed sets in the AIPD were not directly re-useable due to substantial changes in the Cooperative Patent Classification (CPC) system for classifying AI inventions. The updated queries followed the same approach however, identifying likely true positives using several classification systems, including CPC, the United States Patent Classification (USPC), the International Patent Classification (IPC), and Derwent World Patent Index classification system.

[15] Due to the small number of seed documents in evolutionary computation, all 2019–2023 documents from the updated query were added.

[16] The weights were set such that, for each AI component, the number of documents times the weight approximately equaled the number of anti-seed documents, i.e., the largest number among all the training data groups. The weight was capped at about 700 to reduce unnecessary influence from any set having a very small number of documents.

[17] Additional methodological details are provided in Supplementary Information C.

[18] In TensorFlow and Keras, "test" is referred to as "validation," i.e., the subset of data withheld during model training and used to evaluate model performance after each training epoch.

[19] We used stratified splits for the 80/20 training runs (where the strata were seed, anti-seed, positive decision boundary, negative decision boundary, positive examiner annotated, and negative examiner annotated) and trained the models for a maximum of 40 training epochs for each run. We selected the number of epochs to use based on how the average F1 score (over 5 runs) changed, picking the number of epochs that approximately maximized F1 (so as to avoid overfitting). For the final models we used all the training data (i.e., no 80/20 split) with the selected number of epochs from the previous step.

Table 1 Number of documents and sample weights for each source of training data

AI component technology	Metric	Seed/Anti-seed		Decision Boundary		Examiner Annotated	
		Seed	Anti-seed	Positive	Negative	Positive	Negative
Machine learning	Number	1045	14,957	31	1116	103	598
	Sample weight	14.3	1.0	482.5	13.4	145.2	25.0
Evolutionary computation	Number	101	14,964	35	1112	2	797
	Sample weight	148.2	1.0	427.5	13.5	700.0	18.8
Natural language processing	Number	1182	14,956	19	1128	54	709
	Sample weight	12.7	1.0	787.2	13.3	277.0	21.1
Vision	Number	879	14,958	59	1088	24	751
	Sample weight	17.0	1.0	253.5	13.7	623.2	19.9
Speech	Number	828	14,964	19	1128	24	762
	Sample weight	18.1	1.0	787.6	13.3	623.5	19.6
Knowledge processing	Number	725	14,966	76	1071	60	626
	Sample weight	20.6	1.0	196.9	14.0	249.4	23.9
Planning and control	Number	1587	14,960	287	860	50	551
	Sample weight	9.4	1.0	52.1	17.4	299.2	27.2
AI hardware	Number	2885	14,955	49	1098	21	756
	Sample weight	5.2	1.0	305.2	13.6	712.1	19.8

A single document may be classified in more than one AI component technology. Seed documents include those added from 2019–2023 (see Supplementary Information B). We edited the raw training data to exclude overlapping documents such that, for the same AI component technology, the document remained in only one set, with an order of precedence of: examiner annotated, decision boundary, and seed/anti-seed. Additionally, we removed documents without both abstract and claims text following text pre-processing

in patent data. Giczy et al. (2022) suggests this may be the result of too few positive observations in the training data, a characteristic that has not changed since the original analysis. A final observation is that precision is lower than recall in each component technology, suggesting that the models may favor returning relevant AI documents at the expense of higher false positive rates when using a prediction threshold of 50 percent. We provide more discussion on the tradeoff between precision and recall in the "Extensions and discussion" section below.

4 Evaluation

Given the large number of methodological differences between the original AIPD and the 2023 update, we conduct a series of analyses to identify how these changes affect the predictions. Our first analysis focuses on the "disagreements" between the two approaches, or the set of documents predicted as AI by either the AIPD 2023 or the original AIPD (using a 50% prediction threshold) but not both. Table 3 shows for each AI component the number of disagreements of two types: AI predicted in the 2023 AIPD but not the original AIPD, and AI predicted in the original AIPD but not the 2023 update. In addition, the table includes the total number of disagreements, the total number of AI predictions (either AI in the AIPD 2023 or original AIPD, or both), and the percentage of disagreements relative to

Table 2 Training metrics for each AI component technology model

(a) Validation metrics from 5 × 20/80 split training for final number of training epochs

AI component	Validation accuracy		Validation precision		Validation recall		Validation F1	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
Machine learning	0.984	0.005	0.831	0.049	0.935	0.036	0.865	0.020
Evolutionary computation	0.986	0.006	0.208	0.033	0.274	0.021	0.225	0.022
Natural language processing	0.993	0.003	0.906	0.028	0.965	0.019	0.929	0.019
Vision	0.977	0.009	0.725	0.085	0.888	0.028	0.776	0.054
Speech	0.987	0.007	0.796	0.100	0.942	0.023	0.845	0.060
Knowledge processing	0.958	0.015	0.568	0.131	0.843	0.021	0.648	0.080
Planning and control	0.927	0.014	0.628	0.060	0.802	0.037	0.686	0.027
AI hardware	0.943	0.004	0.786	0.025	0.856	0.032	0.811	0.011

(b) Final model training metrics

AI component	Epochs	Accuracy	Precision	Recall	F1
Machine learning	29	0.987	0.831	0.973	0.884
Evolutionary computation	27	0.976	0.267	0.406	0.306
Natural language processing	30	0.991	0.877	0.977	0.912
Vision	32	0.992	0.870	0.976	0.908
Speech	31	0.998	0.940	0.963	0.947
Knowledge processing	28	0.975	0.668	0.945	0.755
Planning and control	25	0.968	0.848	0.954	0.890
AI hardware	30	0.962	0.741	0.986	0.832

In Panel a, an 80/20 train/test split was used five times, and validation metrics were averaged across all five runs for the number of training epochs used in each of the final models. In Panel b, all training data was used to train a final model without a train/test split; hence, the table shows only the training metrics. All metrics are based on a 50% threshold between AI (positive result) and not-AI (negative result) in each AI component

the total number of AI predictions in each component. Notably, the percentage of disagreements is substantial, nearly two thirds or higher in each component technology, ranging from a low of 62.59 percent in planning and control, to a high of 95.35 percent in evolutionary computation. In machine learning, 70.21 percent of the positive predictions from both models are disagreements.

4.1 Manual evaluation

To better understand which model is more likely to be "right" on the set of disagreements, we annotated 229 disagreements at the component technology level for documents published in 2019.[20] We selected these documents for several reasons: first, to evaluate

[20] Since the 15.4 M patent documents in our analysis were sorted by publication data and divided into 1,000 document subsets, we selected one of the subsets that were published in 2019. After running predictions using the new models and consolidating across all eight AI components, we compared those predictions to the ones from the original AIPD models; 229 disagreements at the component technology level

Table 3 Summary statistics on the "disagreements" between the AIPD 2023 and the original AIPD

	AI in AIPD 2023 (50%) but not original AIPD (50%)	Not-AI in AIPD 2023 (50%) but AI in original AIPD (50%)	Total disagreements	Total AI predictions	Percentage of disagreements out of predictions
Machine learning	181,134	83,484	264,618	376,870	70.21%
Evolutionary computation	156,115	39,413	195,528	205,069	95.35%
Natural language processing	236,017	35,715	271,732	393,485	69.06%
Vision	409,096	146,565	555,661	806,991	68.86%
Speech	86,743	37,067	123,810	178,731	69.27%
Knowledge processing	248,410	494,460	742,870	1,111,018	66.86%
Planning and control	422,654	425,267	847,921	1,354,646	62.59%
AI hardware	622,026	187,040	809,066	1,107,293	73.07%
Any AI	1,113,051	332,239	1,445,290	2,628,504	54.99%

Includes all patent documents published between 1976 and 2020 and having predictions from both the updated AIPD 2023 and the original AIPD. A patent document is labeled AI within a given component technology if its prediction is greater than or equal to 50%. Total disagreements are when one model (AIPD 2023 or original AIPD) predicts AI and the other does not. Total AI predictions is either model predicts AI. The difference between the total number of AI predictions and the total disagreements in each component technology is the number of agreements (i.e., both agree AI or not AI in that component). The percentage of disagreements is relative to the total number of AI predictions in that component performance on recent data; second, to assess a period of time that experienced a rapid increase in positive AI predictions (see Fig. 1 below); and perhaps the most important, to examine an area where we might expect to see an improvement in performance (since the decision boundary observations were primarily chosen from those published in 2018). However, for this last reason, any improvement in performance should be considered an upper bound, rather than a population level difference.

We split the documents among three annotators who each labeled the documents for the AI component technology source of disagreement (i.e., those that disagreed for machine learning, natural language processing, etc.), with approximately 80 total annotations for each annotator, about 10 from each AI component (229 annotations total, with almost 30 total from each AI component).[21] The objective of this analysis was to assess which of the two approaches had better performance on these "disagreements."

Table 4 shows several performance metrics, including precision, recall, and the F1 measure (wherein predictions were based on a 50% threshold), from the viewpoint of each model—the 2023 update in the top panel and the original AIPD in the bottom panel. From the perspective of the AIPD 2023, precision is the share of documents labeled as AI by the 2023 update in which this model was correct. Recall is the share of true AI documents, as determined by the annotators, in which the 2023 update was correct. In contrast, the bottom panel of Table 4 shows these metrics from the perspective of the original AIPD.[22] For every AI component technology but one (AI hardware), including "any AI" (i.e., whether the patent document is predicted as AI in at least one AI component technology), the 2023 update has higher precision and recall, which results in higher F1 measures. For example, the AIPD 2023 update achieved 100 percent precision and 65 percent recall in machine learning, while the original model achieved 81.82 percent precision, and only 34.62 percent recall. While the number of annotations in each component technology was not large, when taken together these results suggest that the AIPD 2023 provides better predictive performance than the original model.[23]

The annotation analysis does reveal a weakness in the AIPD 2023, however—the predictions for AI hardware were generally worse than the original model, with an F1 score of only 0.333 on the sample of disagreements. Although it is difficult to understand precisely why, one possibility is that AI inventions are often described in patent documents as being executed on or embedded within hardware. This description makes it difficult to differentiate inventions that are directed toward hardware specifically designed to improve AI systems from general AI software which also describes how the software might be implemented on computer hardware. Moreover, the distinction between hardware specifically designed to improve AI systems and hardware that can be used to more generally improve computation may be hard to distinguish. For example, quantum computers may

Footnote 20 (continued)

using a 50% threshold for both models: 161 documents where the updated model predicted AI in any of the components but the original model didn't, and 36 vice versa.

[21] We attempted to label 30 disagreements total in each AI component technology, but two components in our sample—machine learning and speech—did not have 30 disagreements (at 28 and 21 disagreements, respectively).

[22] From the perspective of the original AIPD, precision is the share of documents labeled AI by the *original AIPD* in which this model was correct. Recall is the share of true AI documents, as determined by the annotators, in which the *original AIPD* was correct.

[23] Example documents from the evaluation, including those that were AI in the original AIPD but are not in the AIPD 2023, and vice versa, are discussed in Supplementary Information C.

enable faster AI training as well as improved execution of other algorithms (e.g., within cryptography), the latter not fitting into our definition of AI hardware (see Giczy et al., 2022). The new training dataset used in the AIPD 2023 included data annotated by many different reviewers, including FIU AI graduate students (for the decision boundary training set) and USPTO patent examiners (for the examiner annotated training set), potentially bringing these definitional challenges to the forefront of the analysis for more challenging components such as AI hardware.

As a final note, it is important to remember that the evaluation of the original AIPD revealed that annotating AI documents is challenging, even for human experts. In that analysis, USPTO examiners achieved 0.348 precision and 0.816 recall, resulting in an F1 score of 0.488 on a random sample of patent documents selected from outside the training set (see Giczy et al., 2022). This issue has not been resolved with the 2023 update—we used the same categorical-based definition of AI as before, as well as the same definitions for each AI component technology. Disagreements between annotators in a second manual review of 300 randomly sampled documents from some of the more difficult cases in the AIPD 2023 (i.e., those that were labeled AI in the update but not-AI in the original AIPD) revealed one potential reason for this disagreement—many USPTO patent applications describe the transmission and manipulation of data through programmable logic.[24] It is challenging to identify when these processes rise to the level of AI, especially in broader components such as planning and control when the data is used to form a plan and control a system, from more basic logical processes, e.g., receiving an input signal from a device and manipulating the signal to produce a desired result.

Such definitional aspects of forming a patent landscape are under-researched but are potentially very important for improving model performance. As previously discussed, we used active learning to identify training data near the decision boundary between AI and not-AI for the different AI component technologies. The ultimate success of active learning depends on the ability of human annotators to consistently label documents near the boundary. The difficulty of human experts to label these cases consistently would place an upper bound on the efficacy of this approach.

5 Extensions and discussion

5.1 Adjusting the prediction threshold to better identify the volume of AI

Figure 1 shows the number of patent documents published in each year from 1976 to 2021 that were predicted to be AI using the 50 percent prediction threshold in the original AIPD

[24] In this second manual review exercise, each of three annotators were given 100 documents randomly sampled from the documents predicted to be AI in the AIPD 2023 and not-AI in the original AIPD, and each document was reviewed by only one annotator. While this analysis cannot reveal anything about recall from the perspective of the AIPD 2023 (since all selected documents were predicted as AI), it can determine precision (as the share of documents accurately predicted to be AI). Reflecting the challenges associated with identifying AI within this set of potentially more challenging documents, precision varied widely across the three reviewers, from a high of 59 percent to a low of 20 percent. Overall precision was 38.67 percent, which is very similar to the overall precision determined on the evaluation set of non-training documents in the original AIPD (i.e., 40.54 percent). As a final note, precision is different on this set than the first manual review sample drawn from 2019, as described above, because (1) the 300 documents did not include the other set of disagreements (i.e., those predicted to be AI in the original AIPD and not-AI in the AIPD 2023), and (2) the annotators were not the same.

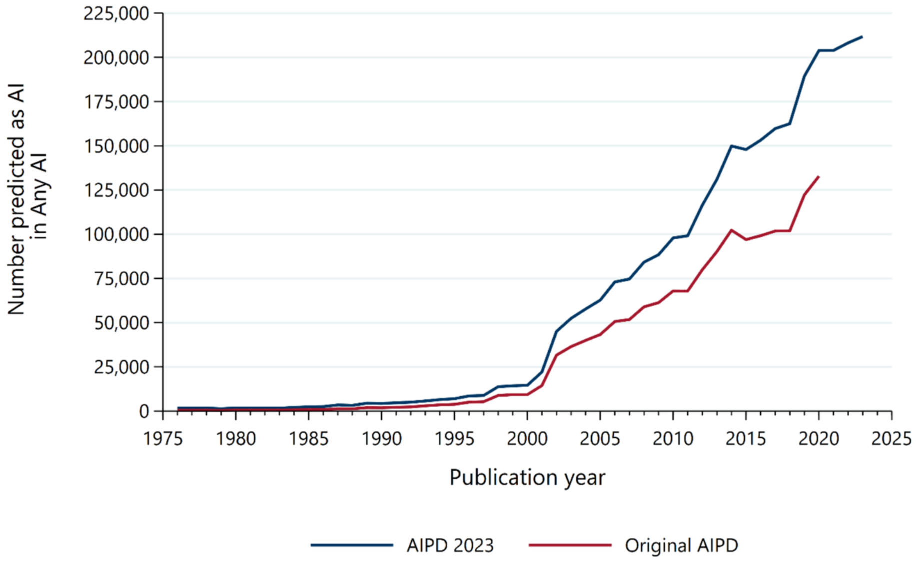

Fig. 1 The number of USPTO patent documents published each year that were predicted to be AI by the original AIPD and the 2023 AIPD update. The original AIPD runs through the end of 2020 and the AIPD 2023 update through the end of 2023. The figure uses a 50% prediction threshold. A document is predicted as "Any AI" if it is predicted as AI in any one of the eight AI component technologies

and from 1976 to 2023 in the 2023 update (also using a 50 percent prediction threshold). Most noticeably, the number of documents predicted to be AI in the AIPD 2023 is substantially higher each year: about 50 percent higher relative to the original AIPD. The models produce similar trends however, with the exception of between 2015 and 2018, where the original AIPD is relatively flat while the AIPD 2023 increases slightly. Exploring the predictions by AI component reveals that the new models consistently predicted more AI than the original models in each component technology, except for knowledge processing where the new model predicted less AI, and planning/control where the two approaches predicted about the same.

One way to adjust the overall number of AI predictions is to change the probability threshold for determining AI. In the prior section we used a 50 percent threshold—those documents with predictions of at least 50 percent were labeled AI in that component while those strictly less than 50 percent were determined not to be AI in that component.[25] Raising the threshold generally increases precision and lowers recall since only documents reaching the new, higher probability threshold are predicted to be AI, but conversely a greater number of true AI documents with intermediate probabilities are missed. Researchers may favor greater precision or recall depending on their application, or they may seek to replicate an existing analysis with extended data from the AIPD 2023 that more closely aligns with the original AIPD.[26]

[25] If a document is predicted as AI in any component then it is identified as being "any AI" (as in Giczy et al., 2022 and Toole et al., 2020).

[26] For example, Grashof and Kopka (2023) prefers the WIPO keyword/classification approach for identifying AI invention in patent documents because of its higher precision. Rather than switching to a method like this, researchers can increase the prediction threshold for AI to increase precision.

Table 4 Performance statistics on a sample of 229 manually-reviewed "disagreements" at the component technology level in a set of 1000 patent documents published in 2019

(a) From the perspective of the AIPD 2023

AI Component	True positive	True negative	False negative	False positive	Total	Precision	Recall	F1
ML	17	2	9	0	28	1.0000	0.6538	0.7907
NLP	17	1	1	11	30	0.6071	0.9444	0.7391
Vision	21	1	2	6	30	0.7778	0.9130	0.8400
Speech	14	0	0	7	21	0.6667	1.0000	0.8000
KR	14	2	12	2	30	0.8750	0.5385	0.6667
Planning	15	2	10	3	30	0.8333	0.6000	0.6977
Hardware	5	5	7	13	30	0.2778	0.4167	0.3333
Any AI	15	4	4	7	30	0.6818	0.7895	0.7317

(b) From the perspective of the original AIPD

AI Component	True positive	True negative	False negative	False positive	Total	Precision	Recall	F1
ML	9	0	17	2	28	0.8182	0.3462	0.4865
NLP	1	11	17	1	30	0.5000	0.0556	0.1000
Vision	2	6	21	1	30	0.6667	0.0870	0.1538
Speech	0	7	14	0	21	–	0.0000	0.0000
KR	12	2	14	2	30	0.8571	0.4615	0.6000
Planning	10	3	15	2	30	0.8333	0.4000	0.5405
Hardware	7	13	5	5	30	0.5833	0.5833	0.5833
Any AI	4	7	15	4	30	0.5000	0.2105	0.2963

Sample consists of patent documents published in 2019 where the original AIPD and 2023 update disagree across at least one of the eight AI component technologies. True and false positives and negatives are based on the perspective from the model noted above each sub table. Each document was reviewed by a single reviewer. Precision for speech in Panel b is not defined since there are no true positives or false positives

From the perspective of accurately predicting the volume of AI, one would like to balance precision and recall since:

$$N_{AI} \cdot Recall = M_{AI} \cdot Precision \quad (1)$$

where N_{AI} is the true volume of AI and M_{AI} is the volume of AI predicted by the model (both sides of the equation are the number of true AI documents predicted by the model). If precision equals recall, then the model accurately predicts the true volume of AI. As discussed above, the AIPD 2023 had better performance overall than the original AIPD. However, recall was higher than precision for the AIPD 2023 in both the training statistics (Table 2) and the manual evaluation (Table 4, Panel a). If this relationship between recall and precision extends to the population of patent documents (i.e., *recall* > *precision*), then the number of documents predicted to be AI in the AIPD 2023 would be biased upward (since $M_{AI} = \frac{Recall}{Precision} \cdot N_{AI} > N_{AI}$).

From the perspective of the original AIPD, Fig. 8 in Giczy et al. (2022) shows precision, recall, and F1 estimated from a holdout sample of 368 patent documents for every

AI prediction threshold. In this figure, precision and recall were relatively balanced at the threshold of 50 percent (at 40.5 percent for precision and 37.5 percent for recall) and were equal at a threshold of 35 percent. Unfortunately, we cannot reproduce the analysis that adjusts the prediction threshold in Giczy et al. (2022) for the AIPD 2023 since we used the original AIPD holdout documents to train the AIPD 2023 models (i.e., the "examiner annotated" training data). However, we can more accurately determine the volume of AI with the AIPD 2023 by using the 35 percent threshold with the original dataset to determine which threshold in the AIPD 2023 would be necessary to replicate a prediction volume that balances precision and recall.[27] To accomplish this task, we analyze different thresholds for the AIPD 2023 to calibrate the prediction volume to that from the original AIPD at a 35 percent threshold.

Figure 2 summarizes this analysis. An AIPD 2023 threshold of 81% (blue dashed line) matches the original AIPD at a 35% threshold from 2000 to 2014, while an AIPD 2023 threshold of 90% (bright blue long dash line) matches 2017 and after. An AIPD 2023 threshold of 86% is the midpoint between these two threshold estimates (green dash-dot line) and appears to split the difference. Thus, researchers could select one of these AIPD 2023 thresholds, either the upper bound (81%), lower bound (90%), or midpoint (86%) to obtain a prediction volume that more closely balances precision and recall. Importantly however, while modifying thresholds does adjust the volume of AI predicted by the models in aggregate, i.e., for "any AI," it does not identify the same U.S. patent documents as AI.

In addition, researchers who would like to replicate the prediction volumes from the original AIPD at a 50% threshold might select a threshold of 93% for the AIPD 2023 (see Fig. C2 in Supplementary Information C).

5.2 More information on the "disagreements" with the original AIPD

The manual evaluation discussed above compared the AIPD 2023 to the original AIPD on a set of patent documents published in 2019 where the two approaches disagreed (using 50 percent thresholds for both), finding that the 2023 update had better performance (see Table 4). To better understand these differences, we further analyzed how the predicted probabilities from both models differed on the set of disagreements. Figure C3 in the Supplementary Information shows machine learning prediction scores from the AIPD 2023 on the y-axis relative to the absolute difference between the prediction scores of the AIPD 2023 and original model on the x-axis for those documents where there is a disagreement at the 50% threshold. The figure reveals that when the two models disagree, they disagree substantially (the figures for the other AI component technologies are similar, and are available upon request). For example, the greatest density of disagreements occurs when the AIPD 2023 predicts AI with near certainty (i.e., close to 1.0), and the original AIPD predicts not-AI with near certainty (i.e., near 0.0, thus resulting in an absolute difference close to 1.0), and vice versa. In other words, the models are not disagreeing most where

[27] One caveat is that the precision and recall estimates provided from the original AIPD were from a random sample of patent documents outside of the training set, and therefore are not population estimates. However, given that the seed and anti-seed sets were only 0.07 and 0.88 percent of the population, respectively, a simple random sample of the size annotated for the original AIPD would have overwhelmingly contained non-training documents (i.e., if the 368 documents had been drawn randomly, the expected number of seed documents would have been 0.23 while the anti-seed would have been 3.22). Therefore, the precision and recall estimates from the non-training set in the original AIPD closely approximate the overall population estimates obtainable from a random sample of similar size.

one or the other is uncertain (i.e., where one or both models predict near 50 percent), but where they are each almost completely certain on the outcome (which is wrong for one of the models).

Figure C3 also illustrates that there is a large degree of variability in the relative predictions, since almost all combinations of valid values are present (thereby forming a completely filled in "arrow" shape). Table 5 provides more information on this variability for each component technology by presenting the percentage of documents that are in each of four sections of the arrow in Fig. C3: (1) upper right, which represents positive AI predictions in the AIPD 2023 at 0.90 or higher while the original AIPD predicts AI at 0.10 or lower[28]; (2) lower right, which represents the opposite; (3) the "tip" of the arrow figure, where both models have a relatively high degree of uncertainty, i.e., the AIPD 2023 predicts at between 0.40 and 0.60 and the original model is within 0.20 of the AIPD 2023 prediction; and (4) the remainder of the figure not in (1), (2) or (3).

Columns (1) and (2) in Table 5 quantify our observation that a small area of Fig. C3, i.e., those disagreements where both models were very certain in their predictions, form a substantial share of the overall disagreements. For example, this small area accounts for nearly 50 percent of the disagreements for speech, and about 42 percent for machine learning. In AI hardware, the share is smaller but still large, at about 20 percent. Further, the area where the AIPD 2023 was most uncertain and the original AIPD was also generally uncertain, i.e., the arrow tip in Column (3), contains very few disagreements (ranging from a high of 1.12 percent to a low of 0.20 percent).

These findings, combined with the overall large number of disagreements observed in Table 3, show that the models are highly sensitive to the underlying data used for training and the approach used to embed the text (the two major differences between the original AIPD and the AIPD 2023). Augmenting the training data by including annotated observations where the active learning model was most uncertain (as well as the examiner annotations and the new AI publications since 2019) and using BERT for Patents instead of Word2Vec dramatically moved the decision boundary, resulting in a new model that disagreed substantially with the previous approach. Despite these large changes, the performance improvement revealed in Table 4 emphasizes the importance of selecting an appropriate embedding approach and generating high quality training data; for example, by using active learning to generate data that allows the model to better learn the location of the decision boundary.

5.3 Comparison to other AI patent datasets

Giczy et al. (2022) benchmarked the original AIPD against several alternatives in the literature, including the Cockburn et al. (2019) and WIPO (2019) patent classification and keyword approaches, finding that the AIPD model significantly outperformed these other methods. The key finding was that these other approaches achieved high precision by specifying narrow queries to identify AI but suffered disproportionately in recall, thereby achieving relatively low F1 scores. By comparison, the original AIPD had lower precision, but disproportionately higher recall, resulting in a higher F1 score that, although not as high as that achieved by USPTO examiners, was much closer than the other approaches.

[28] Since in the first zone the absolute difference between the AIPD 2023 and original predictions are 0.90 or higher, if the AIPD 2023 predicts AI at 0.90 and above, then the original AIPD must predict AI at between 0.0 and 0.10.

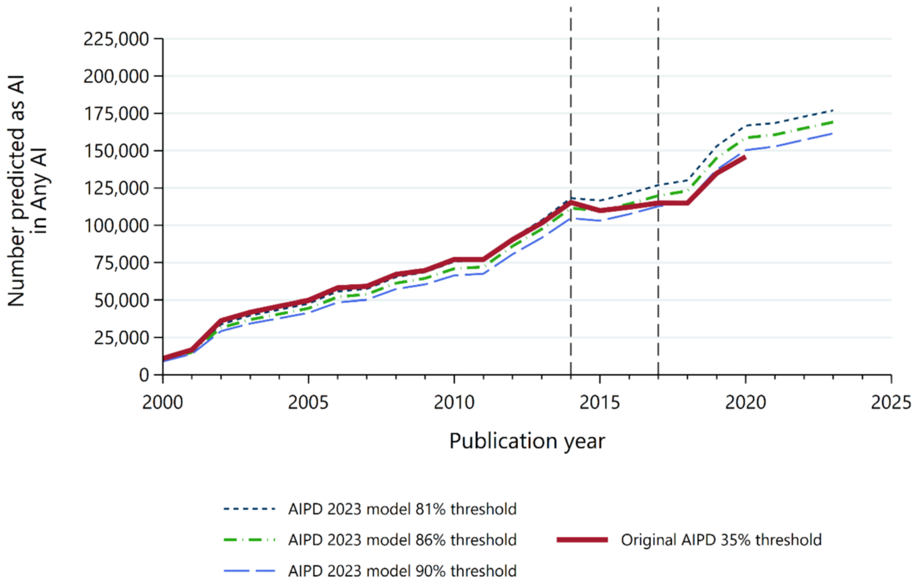

Fig. 2 The number of USPTO patent documents published each year between 1976 and 2023 that were predicted to be AI comparing the 2023 update with varying prediction thresholds to the original AIPD at a 35% threshold

Since the publication of the original AIPD in 2021, an influential AI patent dataset produced by the Center for Security and Technology (CSET) (Thomas & Murdict, 2020) has been used in several policy analyses, including Stanford's AI Index (Maslej et al., 2024; Zhang et al., 2022) and the National Science Foundation's Invention, Knowledge Transfer and Innovation report (Robbins, 2024).[29,30] CSET's approach for identifying AI patents differs from ours in two significant ways. First, CSET's definition of AI relies on the Association for Computing Machinery approach that categorizes AI along 35 dimensions, including AI techniques (e.g., machine learning and logic programming), functional applications (e.g., language processing and computer vision), and application fields (e.g., life sciences and banking/finance). Our definition of AI overlaps significantly with the ACM taxonomy, but we do not use the same AI categories and do not classify directly into AI application fields, preferring to use our algorithm to find AI wherever it exists across technologies (see Toole et al., 2019).

The second major difference is that CSET uses patent classifications and keywords to identify AI, similar to the approaches used by WIPO (2019) and Cockburn et al. (2019). Therefore, we might expect CSET's approach to favor precision over recall, and as a result underreport the true volume of AI. Figure 1 in Thomas and Murdict (2020) reveals this to be the case, finding that just over 10,000 patents and around 65,000 applications were published *worldwide* in 2020. By comparison, the original AIPD at the threshold of 35 percent

[29] Code for implementing the CSET approach has been made available on GitHub at https://github.com/georgetown-cset/1790-ai-patent-data.

[30] The CSET AI data is also used by Our World in Data (https://ourworldindata.org/grapher/artificial-intelligence-granted-patents-by-industry) and articles in the popular media, including by Axios (e.g., https://www.axios.com/local/san-francisco/2024/04/03/silicon-valley-patents-ai-chatgpt).

Table 5 Distribution of prediction scores between the AIPD 2023 and original AIPD models for the set of disagreements using the 50% threshold

AI Component	Zone	(1) Upper right	(2) Lower right	(3) Arrow tip	(4) All others
	AIPD 2023 prediction score	0.9 and above	0.1 and below	Between 0.6 and 0.4	Remaining
	Absolute difference AIPD 2023 and original AIPD	0.9 and above	0.9 and above	0.2 and below	Remaining
Machine learning		28.80%	13.22%	0.29%	57.69%
Evolutionary computation		22.01%	1.14%	0.37%	76.48%
Natural language processing		35.81%	1.56%	0.38%	62.25%
Vision		24.69%	5.63%	0.60%	69.09%
Speech		36.81%	11.24%	0.20%	51.75%
Knowledge processing		8.25%	27.64%	0.52%	63.59%
Planning and control		16.21%	17.51%	0.49%	65.79%
AI hardware		17.31%	2.96%	1.12%	78.62%

See Fig. C3 for a visualization of the regions provided in Columns (1)–(4) for machine learning.

(to balance precision and recall) has nearly 150,000 U.S. PGPubs and patents published in 2020 from the USPTO alone. Despite this fact, the CSET approach has at least one major advantage—it's easily extendable to worldwide patent datasets, whereas the AIPD is significantly more computationally intensive and is currently only available for USPTO publications.

5.4 Practical challenges associated with patent landscaping

We faced several practical challenges when updating the AIPD, which we hope by discussing here will lower the barriers for other economics, legal, and business researchers considering using these or similar methods. First, we required significant computational resources to train our models and execute predictions. The server we used had 112 CPUs, 1.47 TB RAM, and eight NVIDIA A100-40 GB GPUs. Due to the amount of data required to train the models, where we processed BERT for Patents text sub-word tokens through long short-term memory (LSTM) neural networks, we used only the CPUs for training. Training the final models took an average of approximately 1.3 h for each AI technology component model for a total time of approximately 10.6 h. To run model predictions, we divided the approximately 15.4 M patent documents into "shards" of 1000 documents each, and then used one GPU to execute the predictions for groups of 200–800 shards at a time. A group of 400 shards took about 24 h to run in a single GPU enabled process, and we used up to four GPUs and processes simultaneously. The predictions took a total of about 11 calendar days of near constant processing using this parallel approach. A significant time-consuming portion of the process was converting text into BERT for Patents embeddings; since we used all the sub-word tokens of the embedding, these files were very large and could not be reasonably kept beyond their immediate use, particularly for predictions.[31]

A second challenge is associated with evolving patent classification systems, especially for emerging technologies such as AI. New classification symbols may be created, to include new symbols split from old ones or the creation of subordinate symbols, and some symbols may be retired. It is thus important to distinguish whether the classification symbols in patent data are from when the document was published (or granted as a patent) or have been updated to the most recent classification schema and symbols. In our analysis, this challenge affected how we updated the training data beyond 2019, requiring us to modify the classification-based seed set queries originally used to identify documents likely to contain AI (see Supplementary Information B). Many of the original classifications had been replaced, while new symbols had been added, making it challenging to update the training dataset in a way consistent with the scope of the original AIPD queries.

6 Conclusion

The AIPD 2023 extends the AIPD to all USPTO patent and PGPubs through 2023, while also improving the underlying methodology used for identifying AI patent documents. The major methodological changes include the use of BERT for Patents to embed patent

[31] For example, the files for training each AI component technology model were approximately 68–79 GB in size. Running predictions for a shard of 1000 patent documents required approximately 2 GB for each of abstract text and claims text (4 GB total); we kept the embeddings for a shard in memory and ran predictions for all eight classification models before discarding them. Given 15.4 M documents, saving all embedding files to disk would have required over 200 TB.

document abstracts and claims, as well as new training data selected through active learning to better identify where the decision boundary exists between AI and non-AI. In addition to results supporting this method from existing AI patent landscaping research (Islam Erana & Finlayson, 2024), our manual evaluation shows that this new method performs better than the original AIPD approach.

Our study reveals several important insights beyond this overall finding. First, identifying AI in patent documents remains difficult, even for human experts. The research community and policymakers would benefit from greater exploration into the sources of these difficulties; for example, do we need better definitions of AI or better guidelines for human annotators when creating training datasets?[32] Although we found that the AIPD 2023 had better predictions than the original AIPD, the F1 score of 0.73 at the overall AI level on the set of disagreements leaves room for improvement. Improved annotation would translate directly into improved landscaping performance. Further, the AIPD 2023 model for AI hardware performed substantially worse than the original model, perhaps because it may be an especially challenging area of technology to identify and our training dataset included annotations from many different reviewers. Beyond AI, researchers could create strategies to employ when developing training datasets for technology areas of various annotation difficulty.

In addition, a promising area of future research would be to explore model performance with abstracts and titles alone, as these are readily available in European Patent Office's (EPO) Worldwide Patent Statistical Database (PATSTAT)[33] and therefore would allow researchers to extend this approach to patent documents published worldwide. Importantly, this model should require fewer computational resources than our current approach since it would only rely on abstract/title text and not claims. Relatedly, the machine learning architecture could be simplified by taking advantage of text summarization embedding vectors, e.g., [CLS] tokens from BERT for Patents, as opposed to using individual sub-word tokens in complex LSTM networks (see Ghosh et al., 2024 for an implementation of such an approach based on Bert for Patents[34]). By characterizing these tradeoffs, researchers could make better decisions regarding the costs and benefits of different approaches to patent landscaping.

Our analysis revealed the importance of selecting an appropriate prediction threshold for a given application. From the perspective of accurately predicting the volume of AI, researchers should try and balance precision and recall as much as possible. However, this threshold may not be appropriate for other applications, e.g., when assessing diffusion, a researcher might be more concerned about increasing recall at the expense of precision to better assess the reach of a given technology. To the best of our knowledge, very little applied research exists that explores the impact of adjusting precision and recall within applied applications. While likely highly dependent on each application and therefore difficult to characterize, greater exploration into this issue would improve

[32] A related issue is what should be considered an AI invention. AI is diffusing broadly across technologies (Toole et al., 2020) and AI may be but one element out of many that combine to form an invention. AI may be a novel element of the invention or it may not be. Researchers may prefer to define an AI invention in different ways, depending on their application. The AIPD does not consider these alternatives—an invention contains AI if the abstract or claims describe the use of at least one of the AI component technologies. This is an important area of further research.

[33] Patent application titles are included in PATSTAT table TLS202, and abstracts in TLS203. See PATSTAT Global Data Catalog, available at https://www.epo.org/en/searching-for-patents/business/patstat.

[34] Ghosh et al. (2024) used the mean of the output layer embedding tokens, finding it outperformed the BERT [CLS] token.

the evidence base derived from the identification of specific technologies within patent data.

Finally, in recent years, the economics, management, and legal research communities have begun using generative AI within the research process itself (see Korinek, 2023). While we did not use generative AI to update the AIPD, these methods appear promising but also introduce new challenges. For example, how might researchers ensure the generative AI system uses a given technology definition, and if documents are labeled at different times, ensure the system consistently uses the same definition of technology? As we described earlier, these problems also exist with human labelers, but they are perhaps harder to solve with generative AI as it can be difficult to assess the reasons for its decision-making.

Policymakers continue to show strong interest in AI invention (Giczy et al., 2024; U.S. Executive Order, 2023; USPTO, 2024), and for good reason. Using the AIPD 2023, we find that in the five years from 2018 to 2023, the number of annual U.S. AI patent applications increased by 33%, rising from almost 76,000 in 2018 to just over 101,000 by 2023. Across the eight AI technology components, the largest growth over this five-year period was in machine learning, rising from nearly 12,000 applications in 2018 to almost 29,000 in 2023, a 142% increase. How to best incentivize future growth and manage its impacts is an important area of research, and publicly available resources like the AIPD have been used to shed light on these issues. To help researchers, practitioners, and policy-makers better understand the determinants and impacts of AI invention, we have made the AIPD 2023 publicly available on the USPTO's economic research web page (https://www.uspto.gov/ip-policy/economic-research/research-datasets/artificial-intelligence-patent-dataset). More information on the dataset is available in Supplementary Information D, and Supplementary Information E provides helpful information on how researchers may link the AIPD 2023 to other patent data fields, such inventors, assignees, and their locations, using publicly available data from the USPTO-sponsored PatentsView data platform (www.patentsview.org).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10961-025-10189-8.

Acknowledgements The views expressed are those of the individual authors and do not necessarily reflect official positions of the Office of the Chief Economist or the U.S. Patent and Trademark Office.

Author contributions Methodology: N.P., A.G., G.T., T.IE., M.F., A.T. Codebase: N.P., A.G., T.IE., M.F. Writing: N.P., A.G., G.T., T.IE., M.F., A.T. Review/comment: N.P., A.G., G.T., T.IE., M.F., A.T.

Data availability The AIPD 2023 is available on the USPTO's economic research webpage (https://www.uspto.gov/ip-policy/economic-research/research-datasets/artificial-intelligence-patent-dataset).

Declarations

Conflict of interest The authors declare no conflict of interest.

References

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

Chattergoon, B., & Kerr, W. R. (2022). Winner takes all? Tech clusters, population centers, and the spatial transformation of US invention. *Research Policy, 51*(2), 104418.

Chikkamath, R., Rastogi, D., Maan, M., & Endres, M. (2024). Is your search query well-formed? A natural query understanding for patent prior art search. *World Patent Information, 76*, 102254.
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.
Chowdhury, F., Link, A. N., & van Hasselt, M. (2022). Public support for research in artificial intelligence: A descriptive study of US department of defense SBIR projects. *The Journal of Technology Transfer, 47*(3), 762–774.
Cockburn, I.M., Henderson, R., and Stern, S. 2019. The Impact of Artificial Intelligence on Innovation. The Economics of Artificial Intelligence: An Agenda, p.115.
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.
Frumkin, J., Pairolero, N. A., Tesfayesus, A., & Toole, A. A. (2024). Patent eligibility after Alice: Evidence from USPTO patent examination. *Journal of Economics & Management Strategy*. https://doi.org/10.1111/jems.12592
Gao, X., Wu, Q., Liu, Y., & Yang, R. (2024). Pasteur's quadrant in AI: Do patent-cited papers have higher scientific impact? *Scientometrics*, *129*, 909–932
Gaske, M. R. (2023). Regulation priorities for artificial intelligence foundation models. *Vanderbilt Journal of Entertainment and Technology Law, 26*, 1.
Ghosh, M., Erhardt, S., Rose, M. E., Buunk, E., & Harhoff, D. (2024). PaECTER: Patent-level representation learning using citation-informed transformers. arXiv preprint arXiv:2402.19411.
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. https://doi.org/10.1007/s10961-021-09900-2
Giczy, A. V., Pairolero, N. A., & Toole, A. A. (2024). Discovering value: Women's participation in university and commercial AI invention. *Nature Biotechnology, 42*, 26–29. https://doi.org/10.1038/s41587-023-02075-1
Gomes, O., Mihet, R., & Rishabh, K. (2023). A growth and innovation model of the modern data economy.
Grashof, N., & Kopka, A. (2023). Artificial intelligence and radical innovation: An opportunity for all companies? *Small Business Economics, 61*(2), 771–797.
Hötte, K., Tarannum, T., Verendel, V., & Bennett, L. (2022). Exploring artificial intelligence as a general purpose technology with patent data–a systematic comparison of four classification approaches. arXiv preprint arXiv:2204.10304.
Hötte, K., Tarannum, T., Verendel, V., & Bennett, L. (2023). *AI technological trajectories in patent data*.
Hötte, K., Tarannum, T., Verendel, V., & Bennett, L. (2024). *Measuring artificial intelligence*.
Islam Erana, T., & Finlayson, M. (2024). Automated neural patent landscaping in the small data regime. arXiv:2407.08001.
Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature, 61*(4), 1281–1317.
Krestel, R., Chikkamath, R., Hewel, C., & Risch, J. (2021). A survey on deep learning for patent analysis. *World Patent Information, 65*, 102035.
Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. In B. W. Croft & C. J. van Rijsbergen (Eds.), *SIGIR'94* (pp. 3–12). Springer.
Liu, Y., Liu, C., & Huang, T. (2023). A new index measuring occupational exposure to artificial intelligence. Available at SSRN 4515629.
Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J., Shoham, Y., Wald, R., & Clark, J. (2024). The AI index 2024 annual report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, California.
Montobbio, F., Staccioli, J., Virgillito, M. E., & Vivarelli, M. (2023). The empirics of technology, employment and occupations: Lessons learned and challenges ahead. *Journal of Economic Surveys*.
Park, J. (2024). Analyzing the direct role of governmental organizations in artificial intelligence innovation. *The Journal of Technology Transfer, 49*, 1–29.
Pelaez, S., Verma, G., Ribeiro, B., & Shapira, P. (2024). Large-scale text analysis using generative language models: A case study in discovering public value expressions in AI patents. *Quantitative Science Studies, 5*, 1–26.
Pujari, S., Strötgen, J., Giereth, M., Gertz, M., & Friedrich, A. (2022). Three real-world datasets and neural computational models for classification tasks in patent landscaping. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 11498–11513.

Rathi, S., Majumdar, A., & Chatterjee, C. (2024). Did the COVID-19 pandemic propel usage of AI in pharmaceutical innovation? New evidence from patenting data. *Technological Forecasting and Social Change, 198*, 122940.

Robbins, C. (2024). *Invention, knowledge transfer and innovation*. National Science Foundation.

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

Thomas, P., & Murdick, D. (2020). Patents and artificial intelligence: A primer. CSET Data Brief (September 2020). Center for Security and Emerging Technology (CSET), Georgetown University, Washington, District of Columbia, United States. p. 20200038.

Toole, A. A., & Pairolero, N. A. (2020). Adjusting to alice: USPTO patent examination outcomes after Alice Corp. v. CLS Bank International. United States Patent and Trademark Office. IP Data Highlights. Number 3, April.

Toole, A., Pairolero, N., Giczy, A., Forman, J., Pulliam, C., Such, M., & Rifkin, B. (2020). Inventing AI: Tracing the diffusion of artificial intelligence with US patents. United States Patent and Trademark Office. IP Data Highlights, Number 5, October.

Toole, A. A., Pairolero, N. A., Forman, J. Q., & Giczy, A. V. (2019). The promise of machine learning for patent landscaping. *Santa Clara High Technology Law Journal, 36*, 433.

United States Executive Order. (2023). Safe, secure, and trustworthy development and use of artificial intelligence. *Federal Register, 88*, 14110.

United States Patent and Trademark Office (USPTO). (2024). Guidance on use of artificial intelligence-based tools in practice before the united states patent and trademark office. *Federal Register, 89*, 25609.

Vidal, K. (2022). Report to congress, patent eligible subject matter: Public views on the current jurisprudence in the United States. United States Patent and Trademark Office. https://www.uspto.gov/sites/default/files/documents/USPTO-SubjectMatterEligibility-PublicViews.pdf.

Vowinckel, K., & Hähnke, V. D. (2023). SEARCHFORMER: Semantic patent embeddings by siamese transformers for prior art search. *World Patent Information, 73*, 102192.

WIPO (World Intellectual Property Office). (2019). *WIPO technology trends 2019–artificial intelligence*. World Intellectual Property Organization.

Yoo, Y., Jeong, C., Gim, S., Lee, J., Schimke, Z., & Seo, D. (2023). A novel patent similarity measurement methodology: Semantic distance and technological distance. arXiv preprint arXiv:2303.16767.

Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, T., Manyika, J., Ngo, H., Niebles, J., Sellitto, M., Sakhaee, E., Shoham, Y., Clark, J., & Perrault, R. (2022). The AI index 2022 annual report. AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, Sanford, California.