Tracing Technological Leadership in AI: A Patent-Based Comparison of US and Global Assignees

Tily (Hongyu) Tu and Qiao Yue
University of Waterloo
200 University Ave W, Waterloo, ON, Canada N2L 3G1
{h9tu,yue.qiao}@uwaterloo.ca

Abstract

This study examines the global positioning of US assignees in the technological evolution of machine learning (ML) within AI, leveraging patent text analytics. Using over 87 000 AI-related patents from the PatentsView datasets, we apply topic modeling to patent abstracts to extract semantically meaningful keyword embeddings. For each year, we classify assignees into three categories: leaders, followers, and laggards, based on their alignment with industry-wide topic trajectories, drawing on the technology lifecycle framework. Focusing on ML patents, we compare individual assignee trajectories to aggregate industry trends and examine the temporal role of US vs. foreign assignees. Our results suggest that US assignees are more likely to exhibit leadership in early-phase developments but tend to converge with global trends over time. This work offers a scalable framework for measuring technological positioning and sheds light on the evolving global landscape of AI innovation.

1 Introduction

In the global race to dominate artificial intelligence (AI), questions of who leads technological innovation have become central to policymakers and scholars alike. Among various AI subfields, machine learning (ML) stands out as a core technological domain in which organizations actively compete for strategic advantage. Patents offer a valuable window into this race by capturing emerging ideas and the organizations advancing them. As noted by Cockburn et al. (2018), AI is reshaping the innovation process itself, making it increasingly important to understand where, how, and by whom cutting-edge developments are occurring.

Previous research has extensively used patent counts, forward citations, and classification codes to assess innovation performance (Cockburn et al., 2018; Zhu et al., 2022). More recently, textual analysis of patent documents has gained traction as a way to extract richer, semantically grounded insights. Zhu et al. (2022) demonstrates that features derived from patent abstracts, such as diversity and novelty, can meaningfully predict patent outcomes. Similarly, topic modeling has been widely applied to trace technological trends within patent corpora (Tseng et al., 2007). However, this body of work has largely focused on mapping technologies themselves, rather than positioning assignee types relative to industry-wide trends.

What remains missing is a scalable, text-based method to classify organizational positioning within the technology lifecycle. While Younge and Kuhn (2016) developed a vector space model to measure semantic similarity between patents using information retrieval techniques and large-scale USPTO data, their method focuses on pairwise comparisons and improving classification accuracy, rather than identifying where firms stand on emerging or declining technology trends. Specifically, their framework does not distinguish between leaders, followers, and laggards, nor does it track how such roles evolve. The absence of such a method leaves open critical questions about which types of organizations drive frontier innovation and how these dynamics vary across national contexts.

Addressing this gap is essential for understanding the global distribution of innovation capacity. It also informs discussions around national competitiveness, the diffusion of emerging knowledge, and the role of institutional context in shaping innovation strategies. In this study, we apply topic modeling to ML-related patent abstracts from the PatentsView. Using topic distributions, we classify each assignee per year as a leader, follower, or laggard based on their semantic proximity to the global trend.

Our primary contributions are threefold:

- We offer a semantic, text-driven framework for tracking assignee positioning over time

regarding technological frontiers.

- We combine domain-specific embeddings (PatentsSBERTa) with unsupervised topic modeling (BERTopic), and we can map the evolving dynamics of ML innovation.
- We empirically show that US assignees tend to lead in early-phase developments, but that international assignees increasingly converge toward the frontier in subsequent years.

These results provide insight into the evolving geography of AI innovation and demonstrate the value of combining topic modeling with lifecycle theory in patent analysis.

2 Related Work

2.1 Patent-based Indicators of Innovation

Patents have long underpinned measures of innovation output. Traditional metrics, such as patent counts, forward citations, and the size of patent family, are frequently used to assess technological leadership and firm performance (Jaffe and Trajtenberg, 2002; Griliches, 1990). These indicators are often interpreted as proxies for innovation intensity or technological impact, but are limited in their ability to capture the semantic content of inventions.

To address this, scholars have explored more nuanced textual features embedded in patents. For instance, Cockburn et al. (2018) argues that AI not only drives innovation but transforms the innovation process itself, necessitating new approaches for assessing invention quality and direction. Zhu et al. (2022) extends this by showing how technological diversity, derived from patent classification and abstract analysis, can predict examination delays and influence patent value. These studies underscore a growing recognition that the content of innovation matters, not just its volume or citation footprint.

2.2 Textual Analysis and Topic Modeling in Patent Research

Building on this shift, researchers have increasingly applied natural language processing (NLP) and topic modeling techniques to patent corpora. Tseng et al. (2007) demonstrates how clustering and text mining can reveal the structure of innovation systems. Hongshu Chen (2016) explores topic transitions within patent claims to detect emerging shifts in technology direction. Others have employed Latent Dirichlet Allocation (LDA) to detect sectoral convergence and diversification. However, traditional topic models often rely on bag-of-words representations and struggle to capture semantic nuance or time-evolving trends.

To overcome these limitations, transformer-based methods like BERTopic have gained prominence. Grootendorst (2022) introduced BERTopic as a neural topic modeling approach that leverages pre-trained language models. Angelov (2020) further validates the efficacy of semantic topic embeddings in noisy corpora. These frameworks enable more coherent topic discovery and the detection of subtle shifts in thematic focus over time.

2.3 Patent Similarity and Organizational Positioning

Another line of work uses semantic similarity measures to track innovation proximity. Younge and Kuhn (2016) developed a vector space model of patent similarity that improves upon manual classification and provides a foundation for pairwise comparison at scale. While this model advances semantic understanding, it does not classify organizational roles or model temporal evolution across the technology lifecycle.

Organizational positioning, particularly concerning leading or lagging technological trends, has received less attention in the patent literature. Lifecycle theories suggest that firms adopt distinct innovation roles across stages of technological maturity (Abernathy and Utterback, 1978). However, few studies link these theoretical perspectives to empirical classifications based on patent content. Moreover, cross-national comparisons remain rare, with most studies focusing on firm- or industry-level dynamics within a single country.

2.4 Gap and Positioning of Our Work

Our work bridges these strands by integrating neural topic modeling with lifecycle-informed classification of assignees. Using BERTopic, we analyze ML related patents and classify each organization per year as a leader, follower, or laggard based on their alignment with semantic trends in the global corpus. We then compare the trajectories of US-based and international assignees to assess whether national origin is associated with leadership in AI innovation.

3 Methodology

3.1 Data Collection and Filtering

We use two complementary patent data sources: the USPTO Artificial Intelligence Patent Dataset (AIPD;Pairolero et al., 2025) and PatentsView (PV). The AIPD provides an probability whether a patent is in the AI field, while PV offers comprehensive, structured patent data, including titles, abstracts, assignee information, grant years.

The two datasets are merged using patent IDs as unique keys. To narrow the scope to Machine Learning (ML), which is a core and rapidly evolving domain within AI, we apply the `predict93_ml` classifier in AIPD to filter patents most closely associated with ML. There are approximately 87 000 granted patents between 2015 and 2023.

For cross-assignee comparison, we retain only patents with valid assignee names and country information. Assignees are normalized to ensure consistency across records. No filtering is performed based on firm size or sector, preserving a heterogeneous mix of public and private actors contributing to ML innovation.

3.2 Patent Embedding and Topic Modeling

We employ BERTopic as the main topic modeling framework, combined with PatentsSBERTa as the base embedding model. PatentsSBERTa is a Sentence-BERT variant fine-tuned on patent filed It is designed to capture domain-specific semantic representations. We merged the title and abstract of a patent document, and converted into a dense 768-dimensional vector using PatentsSBERTa. These embeddings serve as the basic semantic unit for clustering.

Given that patents often encompass multiple technical aspects within a single abstract or claim, relying solely on document-level topic assignments may ignore important subtopics. Moreover, BERTopic typically assigns only one dominant topic per document. To better capture the multifaceted nature of patents and detect emerging trends at a finer granularity, we further preprocess the data by splitting each patent text into individual sentences and treating each sentence as an independent unit for topic modeling. This sentence-level decomposition allows for more precise attribution of topics and enables us to track subtle shifts in technical focus across the field over time.

The topic assignments are then updated to ensure consistency between downstream analyses. This refinement step improves coverage and semantic consistency, particularly for documents on the borderline that may otherwise be excluded from temporal trend analysis.

3.3 Topic Dynamics and Centroid Construction

To observe topic shifts over time, we compute yearly topic centroids by averaging the vectors of all patents granted in a given year. These centroids serve as semantic anchors that define the evolving technological frontier. We later use them to measure an organization's positioning in the broader innovation landscape.

3.4 Technology Lifecycle Estimation

To classify the role of each assignee within the technology lifecycle, we implement a semantic alignment approach based on a topic vector space. For each year, we compute a global topic centroid by averaging the topic vectors from all granted patents. This centroid represents the general and prevailing semantic focus of the ML field.

We then calculate a yearly topic centroid for each assignee type by averaging the topic vectors of all patents attributed to that assignee type in that year. The cosine similarity between a assignee type's centroid and the global centroid demonstrates the degree of alignment with the mainstream technological trajectory. Based on the distribution of cosine distances, we classify each assignee into one of three lifecycle roles:

Leaders are assignees with vectors positioned ahead of the global centroid in semantic space, indicating early alignment with emerging topics. Followers are assignees closely aligned with the centroid, reflecting engagement with mature or dominant themes. Laggards are positioned further behind the centroid, often concentrating on outdated or trailing areas.

The classification thresholds are defined by empirical quantiles within each year to control for shifts in the global innovation frontier. This operationalization is conceptually consistent with classical lifecycle theories of innovation diffusion, and extends them through a dynamic, language-based semantic framework (Abernathy and Utterback, 1978).

4 Descriptive Statistics

With the rapid improvement in hardware computational power in recent years, AI developed with

an extreme fast speed, driving a surge in related patent filings. Overall, the number of patents rised steadily from about 4,500 in 2015 to nearly 18,000 in 2023.

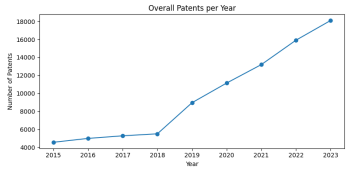

Figure 1: Total patent filings per year (2015–2023).

When broken down by assignee type, US companies and foreign companies have consistently dominated patent activity, growing from roughly 3,500 and 1,000 filings in 2015 to about 12,000 and 6,000 in 2023. In contrast, submissions by other assignee types remain negligible. This result matchs the common sense that companies are the main force that promote the progress of scientific and technological innovation.

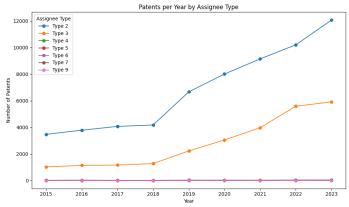

Figure 2: Patent filings by assignee type per year (2015–2023).

4.1 Topic Proportion Stability & Rapidly Growing Topics

Table 1 presents the ten hottest topics in machine learning over the past few years. Each topic's share is relatively small and its standard deviation is also low, indicating that comparing to other industrial fields, machine learning has a very wide range of applications, from computer science to engineering. And the field's development has been fairly stable.

Table 2 lists the ten patent topics with the highest compound annual growth rate from 2015 to 2023. Convolutional and Recurrent Networks has the highest CAGR, with an impressive 68.9%. Generative Adversarial Networks follows at 49.5%, and Transformer Attention Mapping at 44.7%. Each topic with CAGRs above 33%, reflects the rapidly accelerating research activity in these areas.

Table 1: Descriptive Statistics of Topic Proportions (%) (2015–2023)

Topic Name	**Mean (%)**	**Std. Dev. (%)**	**Min (%)**	**Max (%)**
Image Segmentation / Point Cloud	6.41	0.36	6.00	6.93
Computer-Readable Storage Medium	6.13	0.44	5.62	6.78
Search Query Result Ranking	4.25	2.16	1.90	7.93
Autonomous Driving	3.12	0.74	2.16	4.15
Neural Network Layers	2.82	1.19	1.55	4.62
Social Networking Media	2.67	1.17	1.17	3.99
Performance Testing Metrics	2.53	0.23	2.05	2.88
Graph Edge-Node Analysis	2.25	0.20	2.06	2.73
Feature Vector Representation	2.23	0.24	1.80	2.63
Question-Answering (QA)	1.94	0.56	1.21	2.68

Table 2: Top Ten Topics by Compound Annual Growth Rate (2015–2023)

Topic Name	2015	2023	CAGR
Convolutional and Recurrent Networks	13	860	0.6888
Generative Adversarial Networks	10	250	0.4953
Transformer Attention Mapping	10	192	0.4468
Training Hyperparameter Optimization	30	535	0.4335
Variational Autoencoders	9	157	0.4296
Autonomous Drone Swarms	3	46	0.4067
Neural Network Layers	141	1733	0.3684
Online Game Churn Prediction	2	22	0.3495
Policy Rule Management	35	384	0.3491
Payment Transaction Processing	44	440	0.3335

5 Experiments

5.1 Experimental Setup

The BERTopic model is implemented using the previously defined components, including the `PatentsSBERTa` embedding model, `UMAP` for dimensionality reduction, `HDBSCAN` for clustering, `CountVectorizer` for term weighting, and a hybrid `representation model` combining KeyBERT-inspired ranking, part-of-speech filtering, and maximal marginal relevance. To reduce dimensionality while preserving local and global semantic structure, we apply UMAP with the following parameters:

1. `n_neighbors=10`; balances local versus global structure by defining the size of the local neighborhood.

2. `n_components=5`; projects embeddings into a 5-dimensional space, facilitating subsequent clustering.

3. `min_dist=0.02`; controls the tightness of the embedding space (smaller values preserve a more fine-grained separation).

4. `metric='cosine'`; better captures semantic similarity in high-dimensional language embeddings.

For scalability, UMAP is applied in batches of 100,000 documents to reduce memory usage, with all reduced embeddings subsequently concatenated. UMAP and later model takes`random_state=42` to ensure reproducibility.

For dimensionality reduction, we perform unsupervised clustering with HDBSCAN, a density-based algorithm that automatically identifies the number of clusters and handles noise. The parameter settings of HDBSCAN are as follows:

1. `min_cluster_size=20`; prevents formation of overly small, potentially noisy clusters.

2. `min_samples=5`; defines how conservative the cluster boundary should be (higher values result in tighter, more well-defined clusters).

3. `cluster_selection_method='leaf'`; selects fine-grained, leaf-level clusters for better topical separation.

4. `cluster_selection_epsilon=0.02`. controls the minimum inter-cluster separation and further reduces cluster fragmentation.

These settings prioritize cluster stability and coherence while allowing for the exclusion of noise points.

For topic representation, we adopt a **multi-component strategy** combining three representation models supported by BERTopic:

1. **KeyBERT–Inspired**; extracts salient phrases based on semantic similarity to cluster centroids.

2. **Part-of-Speech (POS) filtering**; restricts extracted terms to meaningful grammatical structures (e.g., nouns and noun phrases).

3. **Maximal Marginal Relevance (MMR)** with `diversity=0.3`; balances relevance and diversity in topic keywords to improve interpretability and reduce redundancy.

This multifaceted approach improves both the quality and the interpretability of topic representations.

Subsequently, we compute topic-term weights using class-based TF-IDF (c-TF-IDF). The accompanying `CountVectorizer` is configured as follows:

1. `ngram_range=(1,3)`; allows extraction of unigrams, bigrams, and trigrams to capture fine and coarse patterns.

2. `min_df=5`; excludes extremely rare phrases, improving stability.

3. `max_df=0.8`; removes overly common terms that lack discriminative power.

4. `max_features=10000`. limits vocabulary size to control sparsity and computational cost.

This setup ensures that the topic representations are constructed from statistically robust and topically distinctive terms.

To guide the topic discovery process and improve semantic coherence, we provide BERTopic with a manually curated list of seed topics. These seed topics were determined based on preliminary literature reviews. Each seed topic is defined by a group of representative keywords reflecting well-established subfields or methods in machine learning, such as "deep learning," "support vector machines," "reinforcement learning," "natural language processing," "graph neural networks," and "fairness and explainability," among others. This list consists of 24 seed topics, each capturing a distinct conceptual area within the ML patent landscape.

Considering that sentence-level analysis is being applied, we set `nr_topics=200`to restrict the maximum number of topics generated and to avoid overly fine-grained fragmentation. The model is configured to extract the top 10 keywords per topic (`top_n_words=10`), and a minimum topic size of 20 documents (`min_topic_size=20`) is enforced to filter out small, potentially noisy clusters.

Probabilistic topic assignment is disabled (`calculate_probabilities=False`) to reduce computational overhead and focus on hard clustering results. With these configurations, BERTopic identifies a fixed number of interpretable topics that are both data-driven and domain-informed, enabling a more targeted analysis of technical trends in ML-related patents.

To further improve topic assignment quality and address the presence of noise in unsupervised clustering, we apply **outlier reduction** after initial topic modeling. Specifically, documents initially assigned to the topic `-1` (indicating noise or low-density regions) are reassigned using the model's built-in `reduce_outliers` function with a `c-TF-IDF`-based strategy and a similarity threshold of 0.05. The threshold is based on the recommendation of the BERTopic documents. This method compares each outlier's c-TF-IDF representation with the centroids of existing topics, reassigning the document if a sufficiently close match is found.

5.2 Refinement and Evaluation

To improve the semantic clarity and practical usability of topic labels, we used the latest OpenAI's ChatGPT-O3 model (OpenAI, 2023). Instead of generating labels from scratch, we designed a custom prompt that compared each topic's raw keyword composition with its pre-cleaned name and refined it toward a more human-readable form. The prompt we used is 'Based on the following keywords:..., can you tell me whether it is a research topic in machine learning? If possible, please help me rewrite it into a more fluent and understandable phrase based on your knowledge tags, and categorize its confidence into three categories: low, medium, and high.'

Based on semantic consistency, technical clarity, and domain relevance, each label was assigned a confidence score (high/medium/low). For downstream analysis, we retained only high-confidence labels, resulting in a final set of 103

well-interpretable topics for reporting and visualization.

5.3 Centroid Construction and Similarity Measurement

To quantify how different assignee groups align with the overall evolution of machine learning topics, we construct two types of centroids for each year:

Global Topic Centroid The mean embedding of all high confidence patent topics published in a given year, representing the aggregate research frontier for that period.

We then compute the cosine similarity between each Assignee Type Centroid and the Global Topic Centroid. This similarity score serves as a continuous indicator of whether a group is *leading* (above the 75^{th} percentile), *following* (between the 25^{th} and 75^{th} percentiles), or *lagging* (below the 25^{th} percentile) relative to the global research trajectory. The quantiles are calculated on the basis of the empirical distribution of similarities of all categories in that year. By combining mean based aggregation for trend smoothing and median based aggregation for robustness against outliers, this centroid framework provides the foundation for our subsequent role classification and temporal analysis.

5.4 Results

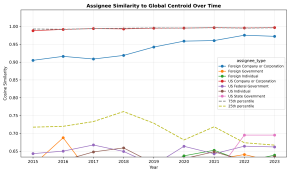

Figure 3: Assignee Similarity to Global Centroid Over Time

Figure 3 shows the evolution of cosine similarity scores for each assignee type's centroid relative to the global centroid from 2015 to 2023. Three main patterns emerge:

1. **US companies lead consistently.** The curve for the US Company or Corporation begins near 0.99 in 2015 and rises to almost 1.00 by 2023, always remaining above the 75th percentile throughout. It indicates that US corporate patents could on behalf of the global development trend of machine learning topics.

2. **Foreign companies converge rapidly.** Foreign Company or Corporation starts at approximately 0.90 in 2015 and exhibits the steepest increase, reaching about 0.98 by 2022. This reflects a strong "catch up" dynamic, as foreign industry progressively aligns with the cutting edge.

3. **Governments and individuals lag but show gradual gains.** Assignee types such as US Federal Government, US State Government, US Individual, and Foreign Individual have lower similarity values (around 0.64–0.70) and greater fluctuation. Although they lie below the 25th percentile in early years, all these groups exhibit modest upward trends toward 2023, showing slow convergence toward mainstream machine learning research.

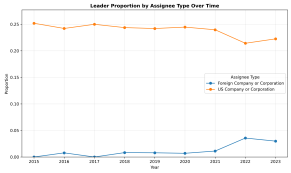

Figure 4: Leader Proportion Over Time

Figure 4 illustrates that the Leader cohort is overwhelmingly dominated by US companies. From 2015 through 2023, "US Company or Corporation" consistently accounts for about 24-26% of all assignee type by year entries in the Leader category. Foreign companies emerge as Leaders only after 2020, rising from near 0% to around 3–4% by 2022–2023. All other assignee types (individuals, government bodies, etc.) contribute negligibly to the Leader segment.

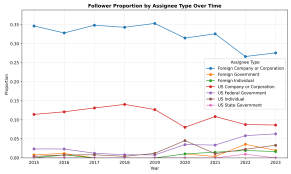

Figure 5: Follower Proportion Over Time

As shown in Figure 5, the Follower role is chiefly comprised of Foreign companies, which represent roughly 32 to 36% of the Follower group each year. US companies form the second largest slice (around 8–14%), followed by US Federal Government and US Individual assignees (each under 6%). Other categories, such as Foreign Government, For-

eign Individual, US Government, together make up the remaining small proportions.

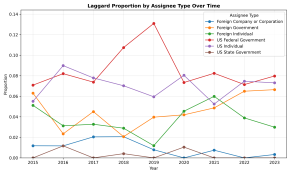

Figure 6: Laggard Proportion Over Time

Figure 6 reveals that the Laggard cohort is more heterogeneous. The largest contributors are the US Federal Government and US Individual assignees, each holding approximately 6–9% of the Laggard group annually. The foreign individual and foreign government assignees account for 2–6%, while the corporate assignees (both US and foreign) occupy minimal shares (generally below 2%). US State and County Government assignees appear only sporadically, with very low proportions.

Overall, corporate assignees, especially US companies maintain the closest alignment with the global patent topic centroid, while other types demonstrate more variability and only gradual movement toward the global research center over time.

6 Conclusion

Based on the above results, it's clear that American companies will lead the major developments in machine learning between 2015 and 2023. While foreign companies' research directions closely mirror those of mainstream companies, their share of leaders is relatively low, though the gap has gradually narrowed in recent years. This phenomenon is primarily due to the fact that the US boasts the world's top research institutions and internet companies.

In comparison, the research of other institutions and organizations accounts for a relatively small proportion, which also shows that patents, as the most important incentive for innovation, are mainly driven by companies and enterprises. Other institutions have relatively weak innovation incentives. Future work should integrate citation networks and market level outcomes to validate these patterns, and policymakers may consider strengthening academia–industry partnerships and refining incentive structures to foster broader innovation leadership.

Data and Code Availability

The complete code and supplementary materials for this project are available at

https://github.com/Noddlezip/Final-Project

References

William J. Abernathy and James M. Utterback. 1978. Patterns of industrial innovation. *Technology Review*, 80(7):40.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.

Iain M. Cockburn, Rebecca Henderson, and Scott Stern. 2018. The impact of artificial intelligence on innovation. National Bureau of Economic Research, Cambridge, MA. NBER Working Paper No. w24449.

Zvi Griliches. 1990. Patent statistics as economic indicators: A survey. Technical report, Harvard Institute of Economic Research, Harvard University.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794.

Donghua Zhu Hongshu Chen, Yi Zhang. 2016. Identifying technological topic changes in patent claims using topic modeling. In *Anticipating Future Innovation Pathways Through Large Data Analysis*, pages 187–209. Springer International Publishing AG.

Adam B. Jaffe and Manuel Trajtenberg. 2002. *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. MIT Press, Cambridge, MA.

OpenAI. 2023. Chatgpt [large language model]. Online service. Retrieved March 2023, from https://openai.com/.

Nicholas A. Pairolero, Adam V. Giczy, Gerardo Torres, Tasnuva Islam Erana, Mark A. Finlayson, and Andrew A. Toole. 2025. The artificial intelligence patent dataset (aipd) 2023 update. *The Journal of Technology Transfer*.

Yuen-Hsien Tseng, Ching-Jiuan Lin, and Yu-I Lin. 2007. Text mining techniques for patent analysis. *Information Processing & Management*, 43(5):1216–1247.

Kenneth A. Younge and Jeffrey M. Kuhn. 2016. Patent-to-patent similarity: A vector space model. SSRN Working Paper. Available at SSRN: https://ssrn.com/abstract=2709238.

Kejia Zhu, Shavin Malhotra, and Yaohan Li. 2022. Technological diversity of patent applications and decision pendency. *Research Policy*, 51(1):104364.