Measuring the Commercial Potential of Science*

Roger Masclans† Sharique Hasan† Wesley Cohen‡

January 8, 2025

Abstract

We develop an *ex-ante* measure of commercial potential of science, an otherwise unobservable variable driving the performance of innovation-intensive firms. To do so, we rely on LLMs and neural networks to predict whether scientific articles will influence firms' use of science. Incorporating time-varying models and the quantification of uncertainty, the measure is validated through both traditional methods and out-of-sample exercises, leveraging a major university's technology transfer data. To illustrate the methodological contributions of our measure, we apply it to examining the impacts of university reputation and university privatization of science, finding that firms' reliance on reputation may lead to foregone opportunities, and privatization (i.e., patenting) appears to increase firms' use of the science of one university. We make our measure and method available to researchers.

Keywords *— Innovation, Technological Opportunity, Science Commercialization, Machine Learning, Large Language Models*

*Authorship in reverse alphabetical order. Sharique Hasan would like to thank the Kauffman Foundation, which funded this study through their knowledge challenge grant. The authors would also like to thank seminar participants at Duke, Bocconi, University of Maryland, NBER i3 Conference, Entrepreneurship and Innovation Policy Research Seminar, the REER Conference at Georgia Tech, as well as helpful feedback from Ashish Arora, Lee Fleming, Dan Gross, Dominique Guellec, Bronwyn Hall, Matt Marx, Robin Rasor, Frank Rothaermel, Yoko Shibuya, and John Walsh.

†Duke University, The Fuqua School of Business.

‡Duke University, The Fuqua School of Business and National Bureau of Economic Research

1 Introduction

Understanding how scientific discoveries contribute to firms' development of new products and services has long concerned managers, policymakers, and academics (Stokes, 1997; Teece, 2020). Managers in innovation-intensive industries are often challenged to identify and use science that can drive their firms' commercial success (e.g., Klevorick et al., 1995; Hounshell and Smith, 1989). Policymakers, too, have long sought to realize the economic value of scientific research, leading to numerous initiatives such as the funding of land-grant colleges in the 19th century to the Bayh-Dole Act of 1980. Yet, despite these efforts, fundamental questions about the contribution of science to commercial outcomes remain an active area of research (Arora et al., 2018).

In this paper, we argue that a methodological barrier impedes our understanding of how science contributes to commercial outcomes: the unobservability of firms' beliefs about the commercial potential of scientific discoveries (Marx and Hsu, 2022). To address this challenge, we develop a novel measure that quantitatively approximates these beliefs by using large language models and neural networks.[1] This measure enables us to distinguish between the *ex-ante* commercial potential of a scientific finding and its eventual realization reflected in firms' R&D investments and related decisions and outcomes.

Understanding the distinction between commercial potential and its realization is crucial because only a fraction of science, as reflected in natural and applied sciences and engineering publications, contributes to commercial applications (Klevorick et al., 1995). Assessing the contribution of science using only *ex-post* measures, such as backward citations to the scientific literature, poses significant challenges. Such measures are problematic due to selection issues, unobserved heterogeneity, and the potential conflation of *ex-ante* commercial potential with other determinants of commercial outcomes. Prior research has emphasized the limitations of relying solely on *ex-post* measures (Azoulay et al., 2007; Marx and Hsu, 2022; Lane and Bertuzzi, 2011). While econometric approaches like twin discoveries (Bikard, 2020) provide less biased estimates of factors influencing commercialization outcomes, they do not address the core issue: the inability to

[1]Both the code and the data are publicly available. The code can be accessed at github.com/CommercialPotentialScience, and the data is available at zenodo.org/records/10815144. We provide the measure for 5.2 million U.S.-affiliated articles published since 2000. Commercial potential scores for all English-language scientific articles published since 2000 are also available at scientifiq.ai

observe commercial potential directly.

This limitation restricts our ability to: (1) properly identify the factors affecting firms' utilization of scientific research due to omitted variable bias associated with unobserved heterogeneity (Bikard and Marx, 2020; Marx and Hsu, 2022); (2) examine the characteristics of scientists, institutions, and firms that shape the commercialization of science (i.e., explore when specific factors are most effective through heterogeneous treatment effects); and (3) evaluate the extent to which commercial opportunities arising from science are unrealized. We suggest that a measure of firms' beliefs in commercial potential could address these limitations, providing a valuable tool for analyzing firms' decisions in industries where science significantly influences innovation. For instance, do drug firms' choices to commercialize a new drug hinge on the drug candidate's inherent commercial potential, or are they driven by the firms' superior commercialization capabilities, such as access to distribution channels? With a measure of commercial potential, it becomes easier to determine whether the discovery of a drug candidate or other moderating factors primarily drive behavior and performance. By quantifying the commercial potential of the science underlying an invention, we can better distinguish the roles of science versus other related factors in shaping firms' innovative performance.[2]

To develop our *ex-ante* commercial potential measure, we employ an ensemble of machine learning algorithms, including large language models and neural networks. These models are trained on scientific article abstracts to produce *ex-ante*, out-of-sample, and out-of-training-time-period predictions of an article's commercial potential, independent of factors influencing its realization. In this context, commercial potential represents the likelihood that a firm perceives an article as contributing to its economic gain. We operationalize this as the *ex-ante* probability that a scientific article is later cited in a renewed patent. This approach assumes that such citations reflect a firm's belief in the value of the underlying scientific finding or idea (Kuhn et al., 2020; Marx and Fuegi, 2020).

Validating our measure using standard holdout samples, we achieve an average AUROC of 0.82,

[2]The extent of this omitted variable problem varies across industries, depending on the reliance of firms' inventions on science. R&D managers in industries such as pharmaceuticals, computers, semiconductors, communication equipment, medical devices, aerospace, and navigation equipment report substantial reliance on academic and government research in science and engineering (Cohen et al., 2002).

with a range of 0.80 to 0.84.[3] To further test the measure, we conduct two additional validation exercises. In the first, we analyze commercial outcomes, such as patent citations and renewals, for over 5.2 million academic papers from US-based research universities. The results strongly confirm the predictive power of our measure: articles with high commercial potential, as identified by our model, are significantly more likely to be commercialized, evidenced by their citation in renewed patents. For instance, an article in the top quartile of our measure is more than 20 times more likely to be cited by a renewed patent than an article in the bottom quartile.

We conduct a second exercise that validates our measure against outcomes that, arguably, more closely approximate firms' beliefs about commercializability than patent citations. Leveraging detailed data on the progression of science through the technology transfer process of a major research university, we validate our measure against the decisions of two key actors: (1) technology transfer office (TTO) personnel who are trying to anticipate firms' decisions and (2) the firms themselves, who are acting on their judgments of an invention's commercializability. The TTO dataset includes comprehensive information on faculty invention disclosures, TTO financial investments, patenting activity, and firms' agreements, licensing, and revenue—all outcomes our model was **not** trained on. Our findings confirm that our measure of the commercial potential of science robustly predicts these outcomes, further supporting its validity.

Finally, we present two applications focusing on firms' use of science to demonstrate the ability of the measure to address omitted variable bias associated with unobserved heterogeneity and heterogeneous treatment effects (e.g., the moderating role of commercial potential). The first application explores how universities' and researchers' reputations shape firms' use of scientific research. We find that firms disproportionately target high-potential research from prominent universities, neglecting comparably commercializable research from less prominent universities. Moreover, we find this effect of reputation is driven more by the reputation of individual researchers than that of institutions. The second application investigates whether the "privatization" of scientific knowledge (i.e., patenting) by research institutions limits its diffusion across firms, showing how patenting and commercial potential interact to influence the use of scientific knowledge. Our

[3]For AUROC comparisons with related studies, see Appendix A.3.

results suggest that university patenting decisions primarily reflect the science's commercial potential and that high-commercial-potential science diffuses more widely when patented.

In addition to the substantive findings noted above, our study can contribute to the study of innovation more generally. Most importantly, it develops a novel, scalable method to measure *ex-ante* commercial potential, addressing challenges that traditional outcome-based measures and natural experiments are not designed to overcome (Azoulay et al., 2007; Marx and Hsu, 2022). For example, it allows scholars to address selection biases by disentangling whether performance gains stem from access to commercially promising ideas versus complementary resources such as skilled teams, organizational structures, and commercialization capabilities. Additionally, using such a measure can facilitate the study of firm heterogeneity by examining how commercial potential may interact with other firm- and market-level variables to influence outcomes. By quantitatively approximating what is otherwise a key unobservable input into a range of decisions, the measure itself can potentially advance our understanding of how firms choose among options to realize value from scientific knowledge, choose across uncertain technological trajectories, source external innovation, or assess startup commercialization paths. Finally, the measure can allow us to address an otherwise unanswerable question (e..g, Christensen and Bower, 1996): To what degree do firms' innovation strategies forego potentially commercializable inventive opportunities?

In addition to its application to the study of innovation, our time-varying models that employ large language models and deep learning to train classifiers for *ex-ante* measures of commercial potential are suited to firms' decisions where conditions change over time. Specifically, our method of estimating a new model for every year, where model weights evolve, confers an ability to account for dynamic business environments, and it is just such environments in which firms typically operate. Another way in which our application of large language models is distinctive is our approach to the fact that such models offer predictions and are thus subject to varying degrees of uncertainty. Accordingly, to better ground subsequent use of our measure, we conduct Monte Carlo dropout simulations to quantify how the degree of uncertainty varies across the range of values, allowing us to identify where our predictions of commercial potential are most uncertain. These features make our methodology well-suited for evaluating a wide range of strategic choices,

such as acquisition potential, alliance opportunities, employee potential, market-entry, and even identifying risks like competitive threats.

2 Data and Methods

This section describes the data, training methodology, and performance of the models developed to measure the *commercial potential* of a scientific article. Based on the abstract text of a scientific article, the measure predicts the *likelihood that a renewed patent cites that article*. We interpret this quantity as reflecting a *firm's belief* that the *knowledge or information in the article* has *commercial potential*, meaning it contributes to the firm's economic advantage. We elaborate on this definition in Section 2.1.

We develop our measure using a transformer-based Large Language Model (LLM) and deep neural networks. Our approach involves training binary classifiers to categorize textual data into predefined classes—whether the scientific article has commercial potential. First, we label articles as having commercial potential if they are cited by at least one renewed patent within the training time frame. As detailed below, we train our models using over 420,000 scientific articles published between 2000 and 2020, *training one model per year* both to avoid data leakage as well as to reflect the changing technological and economic conditions affecting the commercialization of science.

Next, we fine-tune the language model with our labeled dataset, associating each abstract with its commercial potential class. This process identifies regions in a high-dimensional space that correlate with patent citation patterns, allowing us to categorize new abstracts as having high or low commercial potential based on learned patterns. We use SciBERT (Beltagy et al., 2019), a model trained on 1.14M scientific articles, derived from BERT, a foundational model developed by Google AI in 2018 (Devlin et al., 2018).

2.1 Renewed patents as a proxy for commercializability

Identifying scientific discoveries with commercial potential at the time of publication, especially at scale, is a complex task. It requires a clear definition of commercial potential and the selection of appropriate data to evaluate it. We define commercial potential as the probability that a firm anticipates a scientific article will contribute to its economic gain—often, though not exclusively, through developing a marketable product or process. We use patent citations to academic articles

to assess firms' beliefs about this potential, assuming such citations reflect firms' reliance on academic research for commercial purposes (Marx and Fuegi, 2020). Our goal of predicting firms' beliefs or expectations about the commercial potential of a scientific contribution rests on the assumption that these expectations drive firms' decisions to build upon that science. Consequently, our measure focuses on scientific findings with more immediate commercial impact on outcomes such as patents rather than the long-term potential of early-stage research that may take decades to commercialize. Unsurprisingly, among papers that are eventually cited, the average time to the first citation in a renewed patent is 3.5 years.

Although patent data has limitations, it is arguably the most suitable source for our purposes. Patent data are abundant, a critical attribute for training machine learning models, and they encompass a diverse range of firms, including private companies and smaller entities, offering a comprehensive sample across industries and firm sizes. Moreover, patent data provide detailed insights into firms' beliefs about the commercial potential of inventions across fields, industries, and periods. This richness is essential for our objective: training machine learning models to predict *ex-ante* firms' *beliefs* about the commercializability of inventions—i.e., *commercial potential*.

Patent citations to articles are, however, proxies for measuring firms' beliefs about an article's commercial potential. The question is how closely these citations approximate those beliefs. Firms typically patent to protect inventions they intend to incorporate into new products or processes. For instance, research indicates a 60% likelihood that a patented invention in the U.S. will be embodied in a new product or process (Chuang et al., 2024). However, firms also patent for defensive reasons, to deter litigation, block competitors, enhance bargaining power in cross-licensing negotiations, or as a signaling strategy (Cohen et al., 2000). We argue that all these patenting motives—whether or not the patent is ultimately embodied in a product—reflect beliefs about the invention's contribution to the firm's economic performance. For example, when firms patent numerous substitutes to safeguard a core product, as DuPont did with nylon in the 1930s (Hounshell and Smith, 1989), those patents enhance the firm's economic position even if unused in products. Supporting this view, Arora et al. (2008) estimates that, for those inventions that are patented, patenting itself adds nearly 50% more value than if those inventions were not patented, aligning

with findings from Schankerman (1998).

Firms may revise their expectations about the value of a patented invention as they gather new information. To account for this, we use patent renewals as an additional signal of expected economic value (Kuhn et al., 2020; Schankerman, 1998; Pakes, 1986). The rationale is that renewing a patent—an action requiring payment of fees every four years—signals a firm's confidence that the invention will yield future economic benefits. Thus, we operationalize a firm's belief about the commercial potential of a scientific article by whether it is cited in a patent that is renewed at least once.[4] Renewal decisions, like initial patent filings, are forward-looking indicators of firms' beliefs about future profits. While citations in renewed patents are a useful proxy for assessing the commercial potential of scientific discoveries, alternative patent-based measures exist. For example, Kogan et al. (2017) compute stock market reactions to patent grants as a measure of private patent value. However, this approach is limited to publicly traded firms, excluding private companies, whose patents account for only 22% of publications cited in renewed U.S. patents. Furthermore, stock market reactions may not directly reflect firms' beliefs about commercializability. Nonetheless, Section 3.4 shows a strong correlation between our renewal-based measure and Kogan et al. (2017)'s stock market-based measure of patent value.

2.2 Scientific articles and patent data

We use Dimensions.AI as our source of scientific article data.[5] The dataset includes over 139 million publications, including titles, abstracts, sources, author information, fields of research, and other metadata. We limit our analysis to peer-reviewed journal articles and conference proceedings to ensure data quality beyond standard cleaning (e.g., removing duplicates or missing data). Furthermore, we focus on eleven scientific fields covering most natural and applied sciences and engineering, excluding social sciences. The fields are: Agricultural, veterinary, and food sciences; Biological sciences; Biomedical and clinical sciences; Chemical sciences; Earth sciences; Engineering; Environmental sciences; Health sciences; Information and computing sciences; Mathematical

[4]Due to data limitations, we focus on first renewals at four years, though model performance remains consistent when using renewals at eight or twelve years.

[5]Dimensions is a research and innovation database that contains detailed information on publications, patents, grants, clinical trials, and policy documents.

sciences; and Physical sciences. The resulting sample comprises 50,362,042 academic papers.

We source patent citations for scientific papers from the Reliance on Science dataset (Marx and Fuegi, 2020, 2022). The dataset contains 22,660,003 linkages between 3,017,441 patents and 4,017,152 papers. We merged the Reliance on Science dataset with Dimensions using the DOI (Digital Object Identifier—a unique, universal identifier). This procedure matched all 4,017,152 papers in the Reliance on Science dataset to those in our Dimensions subsample.

Next, we use data from the United States Patent and Trademark Office (USPTO). We collect assignee information and renewal status for each patent citing a paper. Using the patent number, we merge the Reliance on Science and USPTO patent datasets. After merging, we find that one or more renewed patents cite 4.93% of papers in our training sample. For this analysis, we assume that firms do not commercially apply papers not cited by a renewed patent.

2.3 Commercial potential model: Training

2.3.1 Year-based model training: Out of sample, out of time-period predictions

We train the commercial potential models by developing one model per year from 2000 to 2020, resulting in 21 models. Each model is trained independently using the same procedure with time-constrained data. For each focal model year (e.g., t), the articles and their labels (i.e., whether they are cited in a renewed patent) are truncated one year prior ($t-1$) to maintain temporal integrity.

This approach of using a moving temporal window offers two benefits. First, it includes only information up to the year before the focal year, preventing data leakage and avoiding upward bias in our predictions.[6] This ensures our commercial potential measure is predictive of out-of-sample and out-of-time observations, mirroring the expectations driving firms' decision-making processes. Second, this approach captures the dynamic nature of knowledge, enabling our models to learn from recent data and assess research's commercial potential based on up-to-date information.[7]

For each focal year t, allowing for the four-year interval before a first renewal, we use articles

[6] Our predictions are also subject to potential data leakage introduced by the LLM (SciBERT). See Appendix B.4 for a discussion of this concern.

[7] Alternatively, using all past data to train one model would entail learning from over two decades of research for the most recent periods, biasing the model toward older years and dampening model performance.

published from $t-14$ to $t-5$ and citation and renewal data from $t-14$ to $t-1$. For instance, for the 2020 model, which measures the commercial potential of articles published in 2020, we train on articles published from 2006 to 2015 and label them with patent and renewal data up to 2019.

We impose a five-year gap between the article data and the focal year because patents can only be renewed four years after their grant. Thus, no article published after 2015 will be labeled as having commercial potential, as there is insufficient time for a citing patent to be renewed by the end of 2019.[8] For articles in the sample, we allow patent citations and renewals up to 2019. For example, we label an article published in 2012 as "high commercial potential" if it is cited by a patent in 2015 that is renewed in 2019. Figure A.1 in Online Appendix A schematically represents the process.

2.3.2 Training sample and process

We take the following steps to construct the training set for each yearly model. First, from the sample of articles defined in Section 2.2, applying the relevant temporal constraints, we randomly select a balanced set of 20,000 articles along with their patent citation data. For example, to train the 2020 model, we randomly sample 10,000 articles published between 2006 and 2015 cited in a renewed patent by 2019 (commercial potential) and 10,000 articles not cited in a renewed patent by 2019 (no commercial potential).

Note two key points. First, we use a fraction of the available data to train each model (20,000 observations each). In machine learning, using a subset of data to minimize computational costs is common practice, provided that model performance remains high, and the sample is representative. Second, we "artificially" balance the training sample by undersampling, meaning we sample fewer observations from the majority class (no commercial potential), so both classes are equally represented during training. This step is necessary because classification tasks—particularly neural networks tend to perform better on balanced datasets (Miric et al., 2022). In Online Appendix A.7, we provide details on both points and show that performance converges with a training size

[8] We could omit this step and still truncate the data, but, by omitting it, we would be including numerous articles in the training set for these four years, all classified as lacking commercial potential, resulting in biased estimates.

of 20,000 articles per year and that the best performance is reached with a balanced sample.

Next, once the training sample is created, we follow best practices by splitting it into three subsets: training, validation, and test sets. Typically, 75% of the data is used for training, 12.5% for validation, and 12.5% for testing. In a neural network classifier, the model is trained through multiple iterations (epochs) to adjust internal weights and minimize prediction errors. The training set optimizes these weights, while the validation set evaluates the model's performance on unseen data during each epoch. This configuration means the validation set is involved in training and guiding adjustments in subsequent epochs. Consequently, the test set, kept entirely separate from training, is used only after the model is fully trained to provide an unbiased evaluation of its generalization performance on new data. This split ensures the model's performance is evaluated using a strict hold-out sample.

Finally, we feed the abstracts and their labels to the classifier for training (using only the training and validation samples). Details of the training process, including the language model, tokenization, and transformer architecture for computing text embeddings, are provided in Online Appendix A.2. Two points are worth noting. First, the text is preprocessed using standard practices. SciBERT processes up to 512 tokens (approx. 512 words) per abstract. Only 1% of abstracts exceed this limit and are truncated. These longer abstracts are evenly distributed across classes, indicating no bias in our results. Second, we framed the task as binary classification and used a sigmoid activation function in the network's final layer. All models were trained for five epochs, with peak performance typically achieved within the first three epochs.

2.4 Commercial potential model: Out of sample performance

Using the estimated models, we make predictions for the hold-out test sample and evaluate performance.[9] The 21 models achieved an average AUROC of 0.82, ranging from 0.80 to 0.84 across yearly models. Similarly, we analyze model performance by scientific field, finding that Health Sciences and Physical Sciences achieve the highest AUROC scores, while Chemical Sciences and

[9]Note that we do not report the out-of-time-period performance in this section. We report the crude performance metrics of the models using a hold-out sample per standard procedures. In section 3, we conduct a deeper evaluation of our models with out-of-sample and out-of-time period predictions. Most importantly, in Section 4, we externally evaluate our models' performance on commercial outcomes that the models were never trained on.

Information and Communication Sciences score lower.[10]

We tested various hyper-parameters to identify the training configuration with the highest out-of-sample performance. The optimal combination was a batch size of 32, a learning rate of 1e-5, and a dropout rate of 0.3.[11] Additionally, we evaluated our classifier using different language models—SciBERT, BERT, and Specter 2.0 (Cohan et al., 2020; Singh et al., 2022)—another state-of-the-art model for scientific document representation. SciBERT consistently outperformed the other models, improving AUROC by 2.8% over Specter 2.0 and 7.4% over BERT. Online Appendix A.6 provides a detailed analysis and performance metrics based on these hyper-parameters and language models. Furthermore, to ensure the robustness of the results and confirm that they are not due to a random draw, we employ cross-validation by training the model five times and computing performance metrics on the hold-out sample for each run. The results are consistent with the reported performance (see Online Appendix A.7).

2.4.1 Model limitations and robustness

While comparable to prior research, our classifier's performance faces three limitations. First, predicting commercial potential from text is inherently complex, comparable to the challenging tasks facing other efforts using machine learning, including, for instance, Liang et al. (2022)'s models predicting invention success (AUROC 0.71-0.76) and Guzman and Li (2023)'s startup success predictions (AUROC 0.60-0.65). Second, although partially addressed by our training approach using a moving temporal window, the model's effectiveness may still be constrained by the evolution of commercial language patterns. Moreover, it may be subject to bias introduced by authors' intentional use of language to overstate the commercial potential of their discoveries. Finally, to the degree that our model predictions do not hold, it is unclear whether such errors are due to model or human error.[12] While there is little we can do to address the first limitation, we

[10] See Online Appendices A.3 for AUROC interpretation details and detailed performance metrics for each yearly model; A.4 for examples of articles from the top and bottom 25 percentiles of commercial potential; and A.5 for field-level performance metrics.

[11] Batch size refers to the number of training subsamples processed by the neural network at each epoch; learning rate controls how much the model adjusts weights in response to error; dropout rate helps prevent overfitting by randomly removing connections between neurons during training.

[12] For instance, consider a scenario where the model forecasts that a renewed patent should cite an article, but it does not. This discrepancy could arise from two possibilities. First, the model's prediction is incorrect, indicating

address the other two.

Commercially framed abstracts and model bias: To validate that our model classifies commercial potential based on substantive scientific content rather than superficial language differences across articles, we conducted two analyses. A log-odds ratio analysis revealed that words most associated with commercial potential primarily reflected scientific and technical concepts rather than generic commercial terms (see Online Appendix B.1). In an experimental exercise, we also prompted ChatGPT to rewrite 50,000 abstracts to provide them with more "commercial flavor". We found no significant change in the model's classifications or predictive ability (see Online Appendix B.2), suggesting our model robustly identifies commercial potential tied to the scientific finding or result rather than linguistic style or framing.

Model uncertainty: We use Monte Carlo dropout estimation to assess our model's uncertainty and robustness against other potential biases. Through repeated simulations with dropout at prediction time, we found our model's predictions were particularly stable for high commercial potential cases (i.e., scores at the 80th percentile) showing low uncertainty in this critical region (see Online Appendix B.3 for details).

2.5 Secondary model: Scientific potential

In addition to commercial potential, scientific potential can influence commercialization decisions because the scientific promise of an idea or discovery may also correspond to commercial applicability (e.g., CRISPR). This is especially true for science in what Stokes calls Pasteur's Quadrant (Stokes, 1997). Thus, we control for an article's scientific potential to isolate the independent effect of commercial potential, as the two may be correlated.

To measure scientific potential, we adapted the methodology developed to measure commercial potential. The main differences follow. First, we use academic citations as an indicator of the eventual realization of a paper's scientific potential. We develop a classification variable using the number of academic citations a paper receives, and we define high scientific realization if the number of scientific citations for a given paper is above the median in our sample: 16 citations.

that decision-makers were justified in not utilizing the scientific knowledge from the article (indicative of a model error), or second, the model's prediction is accurate, suggesting that the decision-makers overlooked or misjudged the commercial value of the information in the article (suggesting human error).

That is, papers cited 16 times or fewer are categorized as having *low scientific potential*, while those cited more than 16 times are classified as having *high scientific potential*. Second,our moving time windows are simplified because, with scientific citations, we do not face the truncation issues that we faced with patent renewals. We train 21 models and truncate academic citations to the focal year, but in this case, the window used for training is $t-10$ to $t-1$. These models perform satisfactorily, with an average accuracy and AUROC of 0.71. We provide more details on the methodology and performance of these models in Online Appendix A.3.2.

3 Out of sample, out of time-period validation

In this section, we expand our validation exercise to include the scientific contributions of U.S. research universities, examining publications from 2000 through 2020. For articles published after each of our 21 models' training periods, we investigate whether those predicted to have significant commercial potential are eventually commercialized (i.e., cited in at least one renewed patent).[13]

3.1 Commercial Potential at U.S. Research Institutions

Our dataset comprises 5,211,133 articles across the eleven academic fields described above.[14] We focus on articles authored by researchers affiliated with commercially active U.S. research institutions. We use the 'R1: Doctoral Universities – Very high research activity' designation from the 2021 Carnegie Classification of Institutions of Higher Education to identify these institutions. Additionally, we identify commercially active research institutions by their membership in the Association of University Technology Managers (AUTM), requiring at least 0.5 full-time equivalent (FTE) staff dedicated to technology transfer. These criteria yield 126 U.S. universities. Our results are robust to alternative delineations of our institutional sample, including all U.S. universities regardless of AUTM membership. In Table 1, we describe the variables used in this exercise and

[13]This exercise differs from the usual training-test validation split used to calculate the AUROC in Section 2.4, where we randomly divide the training sample into training, test, and validation groups without looking at whether the training data came after the test sample data. In contrast, the validation method we use in this section is both out-of-sample and out-of-time-period, providing temporal generalization for our models and predictions. Refer to Section 2.3.1 for a clarification on how the year-based models are constructed.

[14]For each article published in a given year (e.g., 2017), we compute a commercial potential score using a model trained using data only until the prior year, with training data (e.g., articles, citations, and renewals) restricted to articles published or citations and renewals generated before the focal year (e.g., 2016 and earlier). This approach ensures that our scores are less likely to be contaminated by information from future years, making our commercial potential predictions in this exercise both out-of-sample and out-of-time-period.

the rest of the paper. In Table 2, Panel A, and Table 3, Panel A, we present the relevant statistical characteristics and correlations between this sample's key variables of interest.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

We first conduct a descriptive analysis. As shown in Table C.1, Online Appendix C, articles in the top quartile of commercial potential are 21 times more likely to be cited by renewed patents (15.56%) compared to those in the bottom quartile (0.72%).

We test the predictive validity of our commercial potential measure using our sample of over five million articles. First, we estimate a linear probability model predicting the likelihood of a paper being cited in a renewed patent based on its commercial potential. We then assess the additional variance explained by our measure alongside institution and field-year fixed effects. Next, we test the incremental predictive validity of our measure by including other citation-based predictors for commercial potential, such as the authors', institutions', and journals' lagged h-indices. These tests assess the additional predictive value of our commercial potential measure above and beyond these *ex-post* citation-based measures. We estimate the following specification:

$$y_i = \beta_0 + \beta_1 \phi_{i,t-1} + \beta_2 \psi_{i,t-1} + \beta \alpha_{i,t-1}^{high} + \iota_{it} + \theta_{it} + \epsilon_i \tag{1}$$

where for paper i published in year t, y_i indicates whether it is cited by at least one renewed patent. Additionally, $\phi_{i,t-1}$ represents the commercial potential and $\psi_{i,t-1}$ the scientific potential, as determined by models trained with data up to year $t-1$. $\alpha^{high} i, t-1$ is a binary vector indicating if the focal paper is associated with institutions, authors, and journals of high commercial and scientific prominence. For institutions and authors, assessment of prominence is based on whether their commercial and scientific H-indices are in the top 20% at $t-1$. Similarly, journals are considered high-prominence if their scientific and commercial impact factors are in the top 20% (see Table 1 for details on how we compute H-indices at the paper level). The term ι_{it} denotes

an institution fixed effect, and θ_{it} represents a grouped field-year fixed effect at the paper level. These fixed effects account for technological shocks and trends across the paper's field in year t.

Table 4 shows that adding our commercial potential measure to a model with 126 institution and 210 field-year fixed effects increases R^2 by 42% (from 0.090 to 0.128), significantly enhancing explanatory power. Next, comparing models 2 and 3, we find that model 2—containing only our commercial potential measure—has 10% greater explanatory power than a model with six *ex-post* citation-based predictors. In model 4, testing the incremental predictive validity of our measure, we observe a 19.8% increase in R^2 compared to model 3. This indicates that our measure explains nearly 20% more variation beyond fixed effects and other *ex post* commercial impact measures, substantially adding new information. Finally, we introduce our full specification in Table 4, Model 5. Our measure remains significant and substantially contributes to explanatory power even when accounting for other commercial potential proxies. A one standard deviation increase in commercial potential from the mean increases the likelihood of a patent citation from 7.40% to 12.1%—a 63.5% increase.

Additionally, the coefficients for controls reflecting the commercial track records of researchers, institutions, and journals are substantially reduced upon including our measure (see Figure C.1, Online Appendix C). Specifically, a researcher's commercial prominence decreases by 33%, an institution's by 23%, and a journal's by 24%. In contrast, variables linked to scientific prominence do not display similar coefficient changes. Our results further validate our measure of science's commercial potential. It is worth noting that the model incorporates fixed effects at the institution level, effectively accounting for most of the variation across institutions.

[Table 4 about here.]

3.2 Time horizon of the commercial potential measure

Scientific commercialization typically involves significant time lags, often up to 20 years (Adams, 1990; IIT, 1968), and we expect shorter time horizons to be associated with higher commercial potential. Confirming our prior, the analysis in Table C.2 (Online Appendix C) reveals that articles classified with high commercial potential are indeed commercialized more quickly. Articles in the top quartile of commercial potential are twice as likely to receive patent citations within their

first year compared to those in the bottom quartile (36.66% vs. 18.26%). To enhance the rigor of our analysis, we also performed a Kaplan-Meier survival analysis (Figure C.2, Online Appendix C), which yields even more pronounced differences, further confirming our expectation that our measure of commercial potential increases with temporal proximity to realization.

3.3 Cross-field robustness

Given the limitations of renewed patents as commercialization proxies, we examine how field-level differences in patent propensity(Schankerman, 1998) or even innate commercializability may affect the predictive performance of our measure, we test field-specific predictive performance in a simple linear probability model of the likelihood of a paper being cited in a renewed patent as a function of commercial potential—which compares to Model 2 in Table 4. Detailed results are reported in Online Appendix C, Table C.3. We find that across all fields in our sample, the predictive performance of our measure remains strong. However, some fields show higher variance explained and larger coefficients. For instance, Biological Sciences has the highest variance explained at 13.5% and a coefficient of 0.187 for commercial potential, while Earth Sciences shows the lowest variance explained at 2.7% with a coefficient of 0.056.

3.4 Validating against Kogan et al. patent values

We also compare our measure with a patent-based measure of value, notably the patent value estimates developed by Kogan et al. (2017), which are patent values estimated based on stock market reactions to patent grants. This approach provides a useful proxy for commercialization, as market reactions reflect expected revenue. However, this method's limitation is that it only allows us to analyze patents from publicly traded firms. Results are reported in Online Appendix C, Table C.4. We show that our measure, generated using data only from before the patent was granted, strongly predicts the estimated patent values. Our commercial potential measure also predicts patent value controlling for factors such as patent class (CPC), the number of scientific papers a patent builds upon, and the number of future patent citations it receives.[15]

[15]Thus, though trained on a binary, categorical variable representing the existence of a citation, our measure predicts this intensity-based measure, as well as other intensity-based variables such as licensing agreements or revenue, as shown below. This phenomenon is frequently observed in machine learning classification tasks. It can arise from the strong correlation between the features that predict the occurrence of an event and those that influence the magnitude or intensity of the event, conditional on its realization. In particular, embeddings from

4 External Validity: Commercial Potential and Tech Transfer

A limitation of these validation exercises is that they test the measure against patent citations, proxies for firms' beliefs about commercializability. In this section, we test the measure against outcomes marking the progression of science through a major university's technology transfer process that reflect university and firm decisions about commercializability. Providing detailed information from a major university's technology transfer office (TTO), the dataset encompasses all invention disclosures and subsequent actions and outcomes, including patenting, licensing, agreements, revenue, TTO investments per invention, licensee type (startup or established firm), inventor identity, and history with the TTO. We exclude inventions disclosed before 2000 and those not linked to an active researcher at disclosure. The resulting dataset includes 5,219 invention disclosures from January 2000 to December 2020.

One crucial element missing from the TTO data is the linkage between scientific articles and invention disclosures. To match faculty articles to invention disclosures, we follow three steps. First, we match our two primary datasets: (a) Dimensions, containing academic publication information, and (b) the TTO dataset. We extract the names of all researchers affiliated with the TTO's university from Dimensions. Next, we apply a fuzzy matching algorithm to align researchers' names in Dimensions with those who disclosed inventions in the TTO data. The resulting matched dataset links publications and invention disclosures by author name. From 2000 to 2020, 4,367 researchers in Dimensions are matched to the TTO data, linking to 53,180 unique publications and 4,505 inventions. Shared authorship between a paper and an invention does not, however, guarantee a direct match. Therefore, we take two additional steps to establish a match. First, we assess the temporal proximity between a paper and an invention disclosure. Second, we evaluate the textual similarity between the article and the invention disclosure.

We introduce a measure called "time gap" to assess the temporal relationship between academic papers and invention disclosures. It calculates the years between a paper's publication and its corresponding invention disclosure. We use the invention disclosure year as the reference point, marked as time '0'. The time gap is between the paper's publication and the disclosure year.

large language models often encode latent signals that capture an event's likelihood and underlying characteristics related to its intensity, even though the model is explicitly trained on binary outcomes.

For example, a paper published in 2013 and disclosed in 2015 has a negative two-year time gap. Similarly, a 2020 paper linked to the same invention by the same author has a positive five-year time gap. We determine a paper's influence on an invention—matching a paper to an invention—based on whether the time gap falls within a time window of [-1, 3] years. This range is based on discussions with the TTO and its guidelines, which advise inventors to disclose inventions before public dissemination to maintain patenting options in jurisdictions without a one-year grace period post-publication. Research also indicates that scientific publications leading to patents are often temporally close to each other (Azoulay et al., 2007; Marx and Fuegi, 2020). This method identifies 3,173 researchers linked to 19,381 publications and 3,127 inventions.[16]

Our final step in matching publications with inventions relies on textual similarity. We use BERT to generate textual embeddings for both paper and invention titles. We calculate the cosine similarity between their title embeddings for each potential publication-invention pair (sharing a common author and within the [-1, 3] year window). Our analysis indicates that matches with similarity scores above 0.5 likely represent publications that have contributed to an invention.[17]

This three-step process matched 13,445 unique publications to 2,728 inventions, linked through 2,717 researchers. Each invention is associated with a median of 2 publications, consistent with studies on paper-patent pairs (e.g., Marx and Fuegi, 2020). We then prepare two datasets for our analyses. The first, article-level dataset includes information about each article, such as its link to a TTO-disclosed invention, commercial and scientific potential, citations by renewed patents, and other relevant characteristics. This dataset comprises 96,564 articles, with 13,445 (13.92%) linked to an invention disclosure. Table 2, Panel B, provides the summary statistics for these articles.

The second dataset is aggregated at the invention disclosure level. Here, we examine the relationship between a disclosure's commercial potential and outcomes such as TTO investment, patent filings, licensing agreements, and revenue generation. As inventions are often linked to multiple articles, we average each invention's relevant variables (commercial potential, scientific

[16] An invention may have multiple inventors. Thus, it can be matched to publications from multiple researchers. Similarly, a publication can be linked to multiple inventions if it has authors with multiple disclosures or multiple authors who have disclosed inventions within the time window.

[17] We also applied this procedure using the publications' abstracts and the inventions' descriptions. While we observed similar results, title-based matching proved less prone to errors.

potential, and patent citations). This dataset includes 2,728 inventions. Table 2, Panel C, provides the summary statistics for these inventions.[18]

4.1 TTO results

In the first analysis, we regress our commercial potential measure on four sets of variables: 1) faculty decisions to disclose inventions to the TTO; 2) TTO experts' evaluations of commercializability; 3) firms' decisions to contract with the university for intellectual property access; and 4) revenue realization. We examine the relationships between our measure and key commercialization stages: disclosure of article-linked inventions to the TTO, TTO investment and patenting decisions, licensing and agreements, and revenue generation. In addition to the actual commercial outcome of revenue, note that these stages reflect the decisions, and thus the expectations of commercializability, of three types of actors–faculty, TTO experts, and firm managers. The analysis also controls for the invention's scientific potential and inventors' prior experience with the TTO.

We assume scientists disclose inventions to the TTO partly based on their beliefs about their research's commercializability. Figure 1 shows a density plot of the commercial potential of scientific articles, our primary variable of interest. The figure compares university-associated papers not linked to TTO invention disclosures with those linked, clearly showing that the latter have much greater commercial potential, providing a foundation for our analyses. Similarly, Table D.1, Online Appendix D, show the relationship between an article's commercial potential and its likelihood of TTO disclosure. The findings show that articles in the lowest quartile have a 4.62% chance of disclosure, while those in the highest quartile have a 24.74% chance—a 5.35-fold increase.

[Figure 1 about here.]

To test the relationship between commercial potential and commercial outcomes, we estimate the following linear probability model:

$$y_{i,t} = \beta_0 + \beta_1 \phi_{i,t-1} + \beta_2 \psi_{i,t-1} + \beta_3 log(\alpha^{hs}_{i,t-1} + 1) + \theta_{it} + \epsilon_i, \tag{2}$$

where for a scientific discovery in article i published in year t, $y_{i,t}$ is a binary variable indicating

[18] In Table 3, Panels B and C, we present the correlations between the key variables of interest.

whether it is linked to an invention commercial outcome. $\phi_{i,t-1}$ denotes the article's commercial potential from the model trained with data up to year $t-1$, and $\psi_{i,t-1}$ its scientific potential. $\alpha^{hs}i, t-1$ denotes the log-transformed scientific H-index of the paper's author at time $t-1$, referred to as scientific prominence. For papers with multiple authors, $\alpha^{hs}i, t-1$ is the maximum H-index among all authors j, i.e., $\alpha^{hs}i, t-1 = \max j(\alpha^{hs}j, t-1)$.[19] θit is a grouped field-year fixed effect to account for technological shocks and trends in the field of paper i in year t.

Table 5 shows the analysis results with disclosure as the dependent variable. Model 1 examines the baseline impact of fixed effects on disclosure rates. Model 2 shows commercial potential strongly predicts disclosure, increasing explained variation beyond year-field fixed effects from 0.025 to 0.061. A one standard deviation increase (0.31) from the median commercial potential score (0.57) corresponds to a 7.38 percentage point rise in disclosure probability.

Models 3 to 5 include control variables for scientific potential and researcher H-index, indicating scientific prominence. Our primary model, Model 5, confirms the significant role of commercial potential. A one standard deviation increase in commercial potential correlates with a 6.9 percentage point rise in disclosure probability—a 46% difference. The coefficient for commercial potential in Model 5 remains similar to Model 2, confirming its link to researchers' disclosure decisions.

[Table 5 about here.]

Following specification 2, Table 6 analyzes the relationship between our commercial potential measure and later-stage technology transfer outcomes, including TTO decisions to patent and invest, firm agreements, licenses, and revenue, with the article as the unit of observation. Disclosure reflects scientists' decisions, while TTO investment and patenting reflect TTO decisions. Agreements and licenses reflect firms' decisions to build on the invention, while revenue reflects a firm's commercial outcome. Results are expressed as percentage point increases for a one standard deviation change in the commercial potential measure. The results show that higher commercial potential correlates with increased likelihood across all stages.

[Table 6 about here.]

[19] All specifications are robust to using the average or sum of the authors' H-indices. Table 1 details how the H-index is defined.

Consistent with Table 5, the probability of an invention being disclosed to the TTO increases by 6.9% (a 44% increase over baseline). The likelihood of receiving TTO investment increases by 5.6% (51% over baseline), and the chances of obtaining a patent rise by 4.5% (49% over baseline). The data shows a 4.2% increase in the likelihood of reaching an agreement (43% over baseline), a 1.8% increase in securing a license (36% over baseline), and a 0.7% increase in generating revenue (41% over baseline). These results indicate that higher commercial potential is linked to initial disclosure and subsequent commercialization stages. Notably, the scientific potential of articles linked to an invention is unrelated to outcomes except revenue realization. In contrast, faculty inventors' scientific prominence relates to TTO decisions and firm licensing. For the TTO and firms, faculty prominence may signal inventor credibility or serve as a search heuristic for promising science.

In Table 7, we condition our analysis on the presence of invention disclosure, using it as the unit of observation. First, we examine the relationship between our commercial potential measure, aggregated to the level of the invention disclosure[20], and two key TTO decisions: TTO investment in the invention and patenting. TTO investment covers legal costs and marketing expenses. The amount invested signals the TTO's belief in the invention's commercial promise. Therefore, we expect inventions with high commercial potential to receive greater investment. Second, we analyze the number of patents the TTO files for an invention as another proxy for its perceived value.

On the right-hand side, alongside our measure of commercial potential, we control for whether the faculty inventors have prior TTO experience and interact this experience with the invention's commercial potential to account for potential TTO preference for experienced teams to reduce investment risk. Additional controls include the scientific potential of the invention's associated science, the authors' scientific prominence (H-index), and field-year fixed effects. The econometric specification, which resembles specification 2, is as follows:

$$y_k = \beta_0 + \beta_1 \phi_{k,t-1} + \beta_2 \alpha^{tto}_{k,t-1} + \beta_3 \alpha^{tto}_{k,t-1} \phi_{k,t-1} + \beta_4 \psi_{k,t-1} + \beta_5 log(\alpha^{hs}_{k,t-1} + 1) + \Theta_{kt} + \epsilon_k, \quad (3)$$

[20]Recall that more than one article is typically linked to a disclosure; the median is two.

Models 1 to 4 (Table 7) show TTO investment and patenting results. The findings show that high commercial potential inventions are more likely to receive TTO investment and patent protection. These results hold in Models 2 and 4 after controlling for the inventor's prior TTO experience and the invention's scientific potential. In Model 2, a one standard deviation increase in commercial potential raises the probability of investment by 8% (from 50.0% to 54.0%) and patenting by 9% (from 52.2% to 56.7%). Compared to the prior analysis where articles with commercial potential were 36-51% more likely to see commercial outcomes, this analysis, conditioned on disclosure, shows smaller, yet notable, differences of 8-9%. This reduced discriminatory power is due to conditioning on a subset of articles tied to invention disclosure, which, as shown in Figure 1, are more homogeneous in commercial potential. This aligns with the fact that faculty disclosure decisions already reflect commercializability judgments. After controlling for commercial potential, prior TTO experience does not affect investment or patenting decisions. However, scientific prominence and potential are linked to a higher likelihood of TTO investment and patenting decisions. These results raise the question: Do scientific prominence and potential enhance or distract from the TTO's assessment of an invention's commercializability?

[Table 7 about here.]

While Models 1 to 4 findings validate our measure, when the analysis is conditioned on TTO investment in Models 5 to 9, the predictive power of commercial potential for licensing, startup formation, VC investment, and revenue generation diminishes significantly.[21] Only for agreements (Model 5) does our commercial potential measure remain predictive through its interaction with an author's prior TTO experience. These results suggest that TTO investment decisions capture much of the predictive value of the commercial potential measure, supporting consistency of our measure with the expectations of experts.

5 Applications: Reputation, privatization, and firm's use of science

We demonstrate the utility of our measure through two exercises. The first examines how universities' and researchers' reputations for producing commercializable science influence its use by

[21] Figures D.1, D.2, and D.3, in Online Appendix D, provide a visual interpretation of the results.

firms. The second investigates whether the "privatization" (i.e., patenting) of scientific knowledge by research institutions restricts its diffusion across firms. Beyond their substantive contributions, these examples highlight how our measure addresses two key methodological challenges: (a) as a mediator, it mitigates selection bias due to unobserved heterogeneity in commercial potential, thus ensuring that outcomes are attributed to reputation or privatization rather than the underlying potential of the science; and (b) as a moderator, it captures the heterogeneous treatment effects of reputation and privatization on firms' use of science, the effects depending on their interactions with the commercial potential of the underlying research.

5.1 Reputation and realization of commercial potential in the U.S.

In this section, we examine how an institution's or individual researcher's reputation for producing commercializable science influences firms' use of their research. If reputation drives firms' use of science, then comparably commercializable science from less prominent universities or individuals may be overlooked. The challenge lies in disentangling whether reputation serves as a reliable indicator of research quality or merely reflects past achievements, which may not always align with the current or future relevance of the science. To paraphrase Azoulay et al. (2014): Does reputation accurately guide firms to better outcomes, or does it mirror historical success?

At the article level, we examine how the reputation of an institution, researcher, and journal influences the commercial application of science.[22] Our empirical model uses the citation of a publication in a renewed patent as the dependent variable, indicating a firm's belief that a finding is commercializable. The model includes controls for field-year and university fixed effects. We assume a university's or researcher's past record of producing commercializable research contributes to its reputation. As shown in Table 4, however, relying solely on an institution's or individual's history of producing commercialized research to measure reputation does not clarify whether such histories signal a capacity to generate commercializable science or whether reputation itself leads the firm to use the science in question irrespective of any capabilities (Bikard and Marx, 2020). Table 4 further shows that when we include our article-level measure of commercial potential in

[22]We include the journal's reputation for publishing commercializable research, as firms searching for useful science may prioritize journals known for applied work aimed at practical problem-solving.

the model, we can better isolate past achievement's role as a reputation driver versus its function as a signal that allows firms to identify useful science. Accordingly, we include our measure of commercial potential in the the following estimating equation:

$$y_i = \beta_0 + \beta_1 \phi_{i,t-1} + \beta_2 \psi_{i,t-1} + \beta_3 \alpha_{i,t-1}^{high} + \beta_4 \alpha_{i,t-1}^{high} \times \phi_{i,t-1} + \iota_{it} + \theta_{it} + \epsilon_i \tag{4}$$

where y_i indicates if paper i, published in year t, is cited by at least one renewed patent. $\phi_{i,t-1}$ represents the commercial potential, and $\psi_{i,t-1}$ is the scientific potential, both determined by models trained on data up to year $t-1$. $\alpha_{i,t-1}^{high}$ is a vector of binary variables indicating if the paper is linked to commercially and scientifically prominent institutions, authors, and journals. Institution and author prominence is assessed by whether their commercial and scientific H-indices are in the top 20% at $t-1$. Journal prominence is assessed by whether its impact factor, including commercial impact, is in the top 20%. ιit denotes an institution fixed effect, and θ_{it} a grouped field-year fixed effect at the paper level. These account for technological shocks and trends in the paper's field in year t.

Model 1 in Table 8 highlights disparities in commercialization rates between institutions. The significant positive interaction term "Commercial potential x High commercial impact institution" shows that high commercial potential research from prominent institutions is significantly more likely to be cited in renewed patents than comparably commercializable research from less prominent institutions. Holding other variables at their means, top-quartile commercial potential articles have a 14.65% likelihood of being cited in a renewed patent when originating from a highly prominent institution, compared to a 12.26% likelihood for similar articles from less prominent institutions—a 19.49% relative difference. Model 2 demonstrates that institutional prominence remains significant even after controlling for the journal of publication, highlighting the importance of the journal's commercial impact factor. Top-quartile commercial potential articles published in high-impact journals have a 16.29% likelihood of citation in renewed patents, compared to 9.35% for those in lower-impact journals—a substantial 74% difference.

In Model 3, individual researcher prominence is introduced, rendering institutional coefficients insignificant, which suggests that researcher prominence largely explains institutional differences,

perhaps because prominent universities often employ prominent researchers. Articles in the top quartile authored by prominent researchers have a 17.19% likelihood of citation in renewed patents, compared to 10.14% for those by non-prominent authors—a 69.52% gap. Further analyses indicate that such "realization gaps" persist across a range of scientific fields (see Figure E.1, Online Appendix E).

[Table 8 about here.]

Table 8 highlights two points. First, by incorporating our measure, we control for unobserved heterogeneity in commercial potential, enabling us to isolate the independent effect of reputation. Second, the insignificant standalone coefficients for prominent universities, journals, and researchers, combined with the significant interaction effects, allows to explore heterogeneous treatment effects. We find that prominence matters primarily when associated with highly commercializable research, suggesting that valuable research from less prominent sources may be overlooked.

Perhaps most importantly, from both managerial and policy perspectives, our results reveal that research with commercial potential is more frequently overlooked when the author, journal, and, to a lesser extent, the institution lack commercial prominence. Why might this occur? One possibility is that firms and venture capitalists rely on formal and informal search strategies, prioritizing sources with established records of producing commercially relevant research. As a result, equally commercializable research from less prominent sources may be overlooked, potentially disadvantaging both firms and society. Our findings suggest that overlooked opportunities for commercialization are widespread, with much of the high-potential research not utilized by firms originating from less prominent sources. Specifically, among articles with high commercial potential that remain unrealized, 64% are produced by researchers and 79% by institutions lacking prominence (i.e., are in the bottom three quartiles). This pattern highlights the possibility of "lost ideas", where the commercial potential of discoveries fail to be realized due to biases in search strategies favoring more prominent sources.[23]

[23]We define high commercial potential articles as those in the top decile, with a score of 0.89 or higher. While the 0.89 threshold is somewhat arbitrary, we select the top decile because our uncertainty analysis indicates that articles with scores above 0.85 have minimal measurement error.

5.2 Privatization of science and the scientific commons

A central debate surrounding the Bayh-Dole Act is whether university patenting practices restrict the use of academic science by firms, potentially limiting its economic impact. (c.f., Dasgupta and David, 1994; NRC, 2011; Nelson, 2004). Most studies addressing this question focus on the relationship between university patenting and follow-on academic research (c.f., ?Murray and Stern, 2007; Thursby and Thursby, 2002; Jensen and Thursby, 2001). Another critical question is whether university patenting restricts the diffusion of academic science to firms, undermining firms' abilities to capitalize on science. While transferring scientific knowledge to industry is key to driving innovation and economic growth, university patenting may impose barriers that limit firms' access to this knowledge. However, relatively few studies have explored this issue.[24] As several scholars have noted (e.g., Henderson et al., 1998; Murray and Stern, 2007; Mowery et al., 2015), assessing the true impact of academic patenting on firms' access to knowledge remains challenging without observing the counterfactual scenario and controlling for the inherent commercial potential of the underlying science (Azoulay et al., 2007; Marx and Hsu, 2022).

This section explores how our measure of can enhance understanding of the relationship between academic research and its utilization. Specifically, we examine the relationship between a university's patenting of scientific discoveries and their subsequent utilization while accounting for the ex-ante commercial potential of the underlying research. By controlling for commercial potential, we aim to clarify whether observed utilization patterns are shaped by the patenting process itself versus the intrinsic private economic value of the underlying science, offering a more precise test of how privatization affects firms' use of academic discoveries.

[Table 9 about here.]

In Table 9, we consider how academic patenting affects the breadth of university science use across firms. Estimated with a linear probability model, the dependent variable is a count of distinct firms using the science.[25] Commercial Potential is a binary variable indicating if an article

[24] One notable exception is Sampat and Williams (2019), which explores genome-related patents and their effects on both scientific and commercial follow-on innovation. Despite concerns about knowledge restriction, the analysis finds no significant evidence that gene patents restrict follow-on scientific research or commercial investments.

[25] We calculate this by compiling all patent assignees citing a paper and removing duplicates. Patent assignee

is in the top quartile of this university's distribution, marking it as high commercial potential. Patented is a binary variable indicating whether the TTO has patented the invention. Comparing model 3 to model 1, we observe that the Patenting coefficient drops 56% after controlling for commercial potential. Notably, high commercial potential articles linked to TTO-patented inventions are cited by more firms than those not patented. Model 4 shows that TTO-patented high commercial potential articles are cited by 55% more firms than comparably commercializable unpatented articles.[26]

Given the data are from only one university and the likelihood that unobservables other than the commercial potential of the underlying science may impact the value of the patented inventions, these results are tentative. Nonetheless, it appears that university patenting may enhance, rather than inhibit, firm utilization of science. This effect could stem from the increased visibility of patented articles or firms interpreting TTO patenting as a signal of valuable research. These explanations remain speculative, and further research on a broader sample is necessary to confirm these findings and uncover the underlying mechanisms.[27]

These results highlight two important points. First, accounting for commercial potential demonstrates the impact of otherwise unobserved heterogeneity. Notably, in Table 9, the coefficient on patenting shrinks substantially—from 0.041 to 0.028, a 32% decrease in magnitude—when commercial potential is included, while the coefficient for commercial potential remains stable.

Second, the effect of university patenting on firms' use of academic science depends on the commercial potential of the research. Rather than restricting diffusion, patenting appears to enhance utilization, especially for high commercial potential science. This is evidenced by the significant interaction between high commercial potential and patenting, with patented articles cited by a

data can be misleading due to naming inconsistencies (e.g., Apple Inc. vs. Apple Computer Inc.) and unaccounted subsidiaries. Despite these issues, such inconsistencies are likely orthogonal to TTO patenting, so errors should be evenly distributed across patents, regardless of TTO patenting.

[26]The results hold for articles associated with university-invested inventions, not just patented ones, showing nearly identical findings.

[27]Importantly, these results do not suggest that TTO patenting is the primary driver of academic science's impact on corporate innovation. At this university, most high-potential academic science is not disclosed to the TTO (see Table D.1), with only 15.36% of high-commercial potential articles being patented. While patented articles may receive citations from more firms, the far larger proportion of unpatented high-potential research highlights that public disclosure via publication and other channels remain the dominant channels through which academic research influences firm innovation.

broader range of firms—55% more, on average—than equally commercializable but unpatented research. Moreover, including the interaction term reduces the main effect of Patented substantially, from 0.028 to 0.018—a 56% decrease compared to when we did not include commercial potential in the model (e.g., Patented = 0.041). These findings suggest that failing to account for the underlying commercial potential of science may lead to incorrect inferences about the impact of university patenting on the diffusion of science to firms.

6 Discussion

Scientific research drives technological advance and economic growth, yet understanding how discoveries transition into commercial applications remains challenging. A key difficulty lies in distinguishing the commercial potential of scientific findings from their eventual commercialization. Our research addresses this by developing an *ex-ante* measure of the commercial potential of scientific discoveries, capturing firms' expectations about the likelihood that scientific articles will provide economic benefits. To create this measure, we use LLMs and neural networks to train a classifier that predicts whether academic articles will be incorporated into firms' renewed patents. Moreover, going beyond standard time-invariance classifiers, we develop time-varying models that are adapted to dynamic environments that tend to characterize the determinants of innovative activity and performance. Using Monte Carlo dropout simulations we also quantify the predictive uncertainty of our measure over the range of its values, enabling more informed use. Furthermore, beyond standard holdout and out-of-sample validations, we conduct an external, cross-domain validation by analyzing a scientific discovery's progress through a university's technology transfer process. Finally, in two empirical applications, we demonstrate the utility of our measure and method in permitting more accurate estimation of the treatment effects of different variables of interest affecting commercialization (i.e., reputation and patenting), and in underscoring the importance of controlling for the unobserved heterogeneity that can otherwise lead to omitted variable bias.

Our approach to measuring commercial potential allows strategy and innovation scholars to investigate research questions across diverse domains, addressing three key empirical challenges.

First, our measure of commercial potential addresses a key issue: omitted variable bias as-

sociated with unobserved differences in the commercial potential of firms' options. Ignoring this unobserved heterogeneity—across R&D projects, individuals, acquisition candidates, or market—can lead researchers to incorrectly attribute outcomes to firm attributes or decisions, rather than to unobserved differences in the underlying science's commercial potential.[28]

Second, understanding heterogeneous treatment effects is essential for addressing many key questions bearing on innovative activity and performance. By including commercial potential as a moderator in models, our measure enables researchers to explore how underlying differences in commercial potential moderate outcomes that may vary depending on whether the underlying ideas or technologies have high or low commercial potential. For instance, external sourcing of inventions may deliver better results for commercially promising technologies(e.g., Thursby and Thursby, 2002; Shane, 2002). Or, the impact of hiring star scientists on firms' innovative performance (e.g., Palomeras and Melero, 2010; Marx and Hsu, 2022) may produce varied outcomes based on the commercial potential of researchers' prior work. By quantifying these distinctions and using commercial potential as a moderator, our measure enables researchers to assess how the underlying science may shape the success of different types of decisions.

Third, our measure allows researchers to explore novel questions that would otherwise be difficult to address at scale and across diverse scientific domains. For example, researchers can use our measure to identify "foregone opportunities" within firms' R&D pipelines, such as promising projects with high commercial potential that are overlooked (Christensen and Bower, 1996; Cohen et al., 2025). Our measure could also highlight overlooked opportunities for startups and spin-outs, such as when innovations with high commercial potential fail to secure funding or scale effectively.

Beyond these examples, our measure has already been applied by researchers to address empirical challenges in various contexts. For instance, Mumtaz (2025) examines how press coverage and knowledge diffusion influence firms' use of science, while Rezaei and Yao (2025) investigate how government funding reduces technical uncertainty and attracts venture capital investments in neurotechnology, using our measure of commercial potential as a key control to account for

[28] As discussed in the paper, there are valid econometric approaches to address selection, such as identifying twin discoveries (Bikard, 2018; Marx and Hsu, 2022) or using an instrumental variable, but, among other issues, they are costly. They cannot always be applied at scale.

selection effects. Similarly, Yue (2025) uses both commercial and scientific potential as dependent variables. Focusing on highly applied research in artificial intelligence, the study finds that corporate support of AI projects increases their commercial potential but diminishes their scientific potential.

Our measure can also contribute to the study of what is considered to be one of the three key industry-level drivers of firms' incentives to invest in technical advance apart from appropriability and demand conditions—"technological opportunity"—the extent to which an industry's science base makes technical advance easier (i.e., less costly) (Cohen, 2010). Despite its importance as a determinant of R&D, technical advance, market structure, and entry (e.g., Nelson, 1982; Geroski, 1994; Sutton, 1998), there is no consensus on how to operationalize technological opportunity empirically. Existing measures remain limited to technology and industry dummies or survey-based approaches from the 1980s (Klevorick et al., 1995). Thus, we propose that our measure of commercial potential, aggregated and matched to industries or submarkets, offers a promising first step toward operationalizing this much discussed but empirically elusive determinant of innovative activity.[29]

Our work, of course, has limitations. Relying on patent data and assuming that citations from renewed patents reflect the commercial potential of scientific contributions can be questioned, though this assumption is widely accepted (Kuhn et al., 2020). Furthermore, many scientific contributions reach the market without an associated patent. Moreover, the model may only partially capture commercial potential due to variable, indirect paths to commercialization and long time horizons before contributions are embodied in new products (cf., Adams, 1990; IIT, 1968). Our NLP-based technique may miss non-textual factors influencing the decision to use a scientific contribution in technology development, and we do not analyze the nature of prediction errors. Such errors likely add noise to our measure of commercial potential. However, we see promise in the methodology as more data are incorporated into the models.

[29] An example of the construction of such a measure builds on Branstetter et al. (2022), who match scientific articles to the different therapeutic classes of the drug industry. By weighting each article with our measure of commercial potential, Cohen et al. (2025) use this measure of technological opportunity to analyze firms' investment choices across therapeutic classes.

References

Adams, J. D. (1990). Fundamental stocks of knowledge and productivity growth. *Journal of Political Economy*, 98(4):673–702.

Arora, A., Belenzon, S., and Patacconi, A. (2018). The decline of science in corporate r&d. *Strategic Management Journal*, 39(1):3–32.

Arora, A., Ceccagnoli, M., and Cohen, W. M. (2008). R&d and the patent premium. *International Journal of Industrial Organization*, 26(5):1153–1179.

Azoulay, P., Ding, W., and Stuart, T. (2007). The determinants of faculty patenting behavior: Demographics or opportunities? *Journal of Economic Behavior & Organization*, 63(4):599–623.

Azoulay, P., Stuart, T., and Wang, Y. (2014). Matthew: Effect or fable? *Management Science*, 60(1):92–109.

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Bikard, M. (2018). Made in academia: The effect of institutional origin on inventors' attention to science. *Organization Science*, 29(5):818–836.

Bikard, M. (2020). Idea twins: Simultaneous discoveries as a research tool. *Strategic Management Journal*, 41(8):1528–1543.

Bikard, M. and Marx, M. (2020). Bridging academia and industry: How geographic hubs connect university science and corporate technology. *Management Science*, 66(8):3425–3443.

Branstetter, L., Chatterjee, C., and Higgins, M. J. (2022). Generic competition and the incentives for early-stage pharmaceutical innovation. *Research Policy*, 51(10):104595.

Christensen, C. M. and Bower, J. L. (1996). Customer power, strategic investment, and the failure of leading firms. *Strategic Management Journal*, 17(3):197–218.

Chuang, H.-C., Hsu, P.-H., Lee, Y.-N., and Walsh, J. (2024). What share of patents is commercialized? *Georgia Tech Working paper*.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.

Cohen, W. M. (2010). Fifty years of empirical studies of innovative activity and performance. *Handbook of the Economics of Innovation*, 1:129–213.

Cohen, W. M., Higgins, M. J., Miles, W. D., and Shibuya, Y. (2025). Blockbusters, sequels, and the nature of innovation. *Duke University Working Paper*.

Cohen, W. M., Nelson, R., and Walsh, J. P. (2000). Protecting their intellectual assets: Appropriability conditions and why us manufacturing firms patent (or not).

Cohen, W. M., Nelson, R. R., and Walsh, J. P. (2002). Links and impacts: the influence of public research on industrial r&d. *Management Science*, 48(1):1–23.

Dasgupta, P. and David, P. A. (1994). Toward a new economics of science. *Research Policy*, 23(5):487–521.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

Geroski, P. A. (1994). *Market structure, corporate performance and innovative activity*. Oxford University Press.

Guzman, J. and Li, A. (2023). Measuring founding strategy. *Management Science*, 69(1):101–118.

Henderson, R., Jaffe, A. B., and Trajtenberg, M. (1998). Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. *Review of Economics and statistics*, 80(1):119–127.

Hounshell, D. A. and Smith, J. K. (1989). Science and corporate strategy. *Cambridge Books*.

Hsu, P.-H., Lee, D., Tambe, P., and Hsu, D. H. (2020). Deep learning, text, and patent valuation. *Text, and Patent Valuation (November 16, 2020)*.

IIT (1968). *Technology in retrospect and critical events in science*. Illinois Institute of Technology.

Jensen, R. and Thursby, M. (2001). Proofs and prototypes for sale: The licensing of university inventions. *American Economic Review*, 91(1):240–259.

Kapoor, S. and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9).

Klevorick, A. K., Levin, R. C., Nelson, R. R., and Winter, S. G. (1995). On the sources and significance of interindustry differences in technological opportunities. *Research Policy*, 24(2):185–205.

Koffi, M. and Marx, M. (2023). Cassatts in the attic. Technical report, National Bureau of Economic Research.

Kogan, L., Papanikolaou, D., Seru, A., and Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *The quarterly journal of economics*, 132(2):665–712.

Kuhn, J., Younge, K., and Marco, A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, 51(1):109–132.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Lane, J. and Bertuzzi, S. (2011). Measuring the results of science investments. *Science*, 331(6018):678–680.

Liang, W., Elrod, S., McFarland, D. A., and Zou, J. (2022). Systematic analysis of 50 years of stanford university technology transfer and commercialization. *Patterns*, 3(9):100584.

Manjunath, A., Li, H., Song, S., Zhang, Z., Liu, S., Kahrobai, N., Gowda, A., Seffens, A., Zou, J., and Kumar, I. (2021). Comprehensive analysis of 2.4 million patent-to-research citations maps the biomedical innovation and translation landscape. *Nature Biotechnology*, 39(6):678–683.

Marx, M. and Fuegi, A. (2020). Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9):1572–1594.

Marx, M. and Fuegi, A. (2022). Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy*, 31(2):369–392.

Marx, M. and Hsu, D. H. (2022). Revisiting the entrepreneurial commercialization of academic science: Evidence from "twin" discoveries. *Management Science*, 68(2):1330–1352.

Miric, M., Jia, N., and Huang, K. G. (2022). Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strategic Management Journal*, 44(2):491–519.

Mowery, D. C., Nelson, R. R., Sampat, B. N., and Ziedonis, A. A. (2015). *Ivory tower and industrial innovation: University-industry technology transfer before and after the Bayh-Dole Act*. Stanford University Press.

Mumtaz, S. (2025). Lost in translation? science communication and the commercial diffusion of ideas. *Berkeley, Haas, Working Paper*.

Murray, F. and Stern, S. (2007). Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior & Organization*, 63(4):648–687.

Nelson, R. R. (1982). *An evolutionary theory of economic change*. Harvard University Press.

Nelson, R. R. (2004). The market economy, and the scientific commons. *Research Policy*, 33(3):455–471.

NRC, N. R. C. O. T. N. A. (2011). Managing university intellectual property in the public interest.

Pakes, A. (1986). Patents as options: Some estimates of the value of holding european patent stocks. *Econometrica*, 54(4):755–784.

Palomeras, N. and Melero, E. (2010). Markets for inventors: Learning-by-hiring as a driver of mobility. *Management Science*, 56(5):881–895.

Pontikes, E. G. (2012). Two sides of the same coin: How ambiguous classification affects multiple audiences' evaluations. *Administrative Science Quarterly*, 57(1):81–118.

Rezaei, R. and Yao, Y. (2025). Venture capital response to government-funded basic science. *Available at SSRN 5044008*.

Sampat, B. and Williams, H. L. (2019). How do patents affect follow-on innovation? evidence from the human genome. *American Economic Review*, 109(1):203–236.

Schankerman, M. (1998). How valuable is patent protection? estimates by technology field. *The RAND Journal of Economics*, pages 77–107.

Shane, S. (2002). Selling university technology: Patterns from mit. *Management Science*, 48(1):122–137.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., and Feldman, S. (2022). Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*.

Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Brookings Institution Press.

Sutton, J. (1998). *Technology and market structure: theory and history*. MIT press.

Teece, D. J. (2020). Fundamental issues in strategy: Time to reassess. *Strategic Management Review*, 1(1):103–144.

Thursby, J. G. and Thursby, M. C. (2002). Who is selling the ivory tower? sources of growth in university licensing. *Management Science*, 48(1):90–104.

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

Yue, D. (2025). I, google: Estimating the impact of corporate involvement on ai research. *Georgia Tech Working Paper*.

Zunino, D., Suarez, F. F., and Grodal, S. (2019). Familiarity, creativity, and the adoption of category labels in technology industries. *Organization Science*, 30(1):169–190.

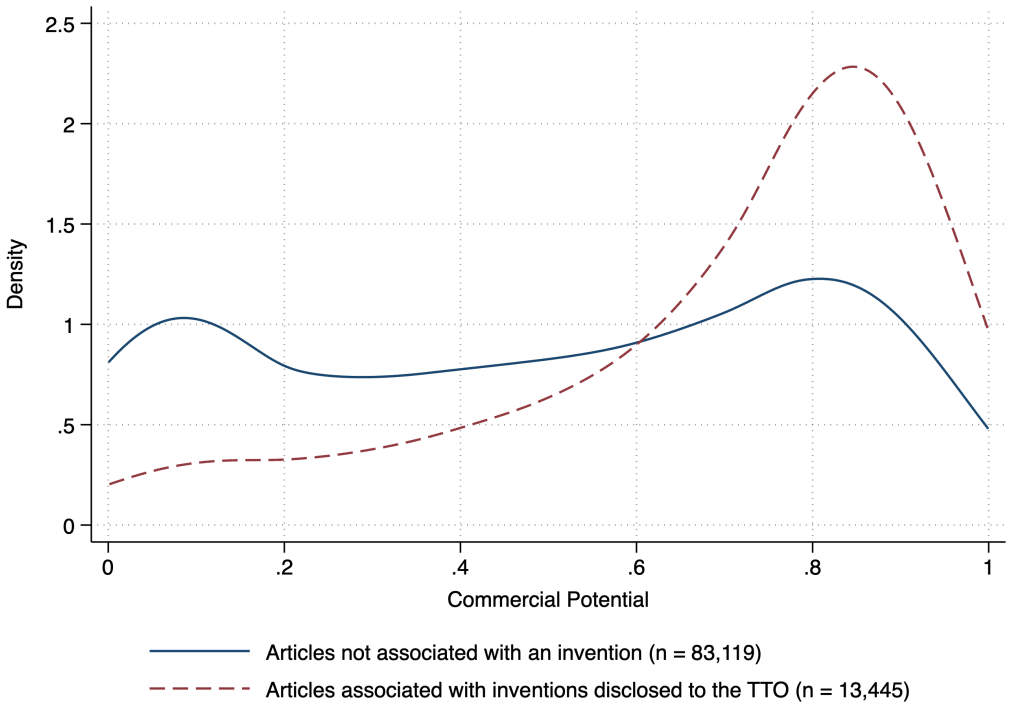


Figure 1: Bi-weight kernel density estimates of the distributions of the commercial potential of 1) articles published at this university not associated with an invention disclosure (solid line) and 2) articles associated with inventions disclosed to the Technology Transfer Office (dashed line). Articles tied to an invention are more likely to have high commercial potential.

Table 1: Variable descriptions.

Variable	**Measure**	**Description of measure**
Disclosed	Disclosed	Binary variable representing whether an article is tied to an invention disclosed to the TTO.
Investment	Investment	Amount invested ($) by the TTO to pursue the commercialization of an invention. Includes different natures of expenses, such as patenting and marketing expenses. The majority of the specifications use a binary variable, indicating whether an invention received any investment.
Patents	Patents	Number of patents the TTO filled to protect a given invention. The majority of the specifications use a binary variable, indicating whether at least one patent was filed.
Agreements	Agreements	Number of commercial agreements—of any nature—associated with an invention. The majority of the specifications use a binary variable, indicating whether at least one agreement was established.
Licenses	Licenses	For each invention, number of licensing agreements with third parties, such as firms or other institutions. The majority of the specifications use a binary variable, indicating whether at least one licensing agreement was established.
Revenue	Revenue	Amount of revenue ($) generated by the invention. The majority of the specifications use a binary variable, indicating whether the invention generated a positive revenue.
Startup	Startup	Binary variable indicating whether the invention has been commercialized via Startup.
VC Investment	VC Investment	Conditional on Startup, binary variable indicating whether the startup has raised venture capital financing.
Authors' TTO Experience	Authors' TTO Experience	Binary variable representing whether at least one of the authors/inventors associated with the invention, prior to the focal disclosure, has disclosed an invention to the TTO.
Commercial potential	ϕ = P(Patent renewal | patent cite > 0)	Probability that the focal article will be cited by at least one patent that, in turn, will be renewed. The probability is the output of our primary model, which uses the abstract text of the focal article to cast the prediction.
Scientific potential	ψ = P(Paper cite > 16)	Probability that more than 16 academic articles will cite the focal article. The probability is the output of our secondary model (Scientific potential), which uses the abstract text of the focal article to cast the prediction.
Author scientific prominence	Max of authors' scientific H-index	Author H-index at time $t-1$, excluding the focal article. If a paper is authored by more than one author, we use the maximum of the authors' scientific H-indices. The H-index captures the productivity and impact of an author and is calculated by counting the number of an author's publications that have been cited by other authors at least that same number of times. Formally, the H-index can be defined as $h_{\text{index}} = \max\{i \in N : g(i) \geq i\}$, where $g(i)$ represents the number of cites of the paper with index i.
Author commercial prominence	Max of author's commercial H-index	Author commercial H-index at time $t-1$, excluding the focal article. If a paper is authored by more than one author, we use the maximum of the authors' commercial H-indices. Similar to the scientific H-index, the commercial H-index is calculated by counting the number of publications cited by patents.
Institution scientific prominence	Max of institutions' scientific H-index	Institution H-index is computed as the author scientific H-index, but we use the institution as the focus of analysis and, thus, the papers affiliated with an institution. If a paper is authored by more than one institution, we use the maximum of the institutions' scientific H-indices.
Institution commercial prominence	Max of institutions' commercial H-index	Idem as institution H-index, but using patent citations to papers instead of academic citations.
Journal scientific impact factor	Journal impact factor	For every year, the average number of citations of articles published in the last two years in the focal journal (source: Marx and Fuegi (2020, 2022)).
Journal commercial impact factor	Journal commercial impact factor	For every year, the average number of patent citations to articles published in the last two years in the focal journal (source: Marx and Fuegi (2020, 2022)).

Table 2: Summary statistics. Panel A summarizes U.S. scientific research published between 2000 and 2020 in U.S. R1 Universities with an active TTO. Panel B summarizes the relevant features for articles whose authors were affiliated with the TTO's university at the time of publication, 2000-2020. Panel C summarizes the relevant features of the articles *associated* with disclosed inventions, 2000-2020. For confidentiality reasons, invention-level outcomes are removed (Investment, Patents, Agreements, Licensing, Revenue, Startup, and VC funding).

Panel A: Articles from R1 U.S. Universities with active TTOs (N = 5,211,133)

	Mean	SD
Commercial potential	0.49	0.33
Scientific potential	0.66	0.24
Cited by patent	0.10	0.30
Cited by renewed patent	0.07	0.26
Institution(s) commercial prominence	68.70	40.54
Institution(s) scientific prominence	411.40	227.77
Journal commercial impact factor	0.02	0.05
Journal scientific impact factor	3.08	3.05
Author(s) commercial prominence	4.23	4.76
Author(s) scientific prominence	32.39	25.59

Panel B: Articles from TTO's university (N = 96,564)

	Mean	SD
Commercial potential	0.52	0.31
Scientific potential	0.73	0.20
Academic cites	62.54	210.09
Patent cites	0.71	5.77
Cited by patent	0.11	0.31
Cited by renewed patent	0.08	0.27
Author(s) scientific prominence	45.26	31.04
Disclosed	0.14	0.35

Panel C: TTO inventions (N = 2,728)

	Mean	SD
Commercial potential	0.73	0.21
Scientific potential	0.76	0.15
Academic cites	74.95	140.32
Patent cites	2.41	9.38
Cited by patent	0.46	0.50
Cited by renewed patent	0.37	0.48
Author(s) scientific prominence	49.47	28.76
Author(s) TTO experience	0.68	0.46

Table 3: Correlations between the key variables of interest.

Panel A. Articles from R1 U.S. Universities with active TTOs

Variables	Cited by patent	Years patents renewed	Commercial potential	Scientific potential	Institution commercial prominence	Institution scientific prominence	Journal commercial impact	Journal scientific impact	Author commercial prominence	Author scientific prominence
Cited by patent	1.000									
Years patents renewed	0.786	1.000								
Commercial potential	0.253	0.206	1.000							
Scientific potential	0.045	0.041	0.224	1.000						
Institution commercial prominence	-0.035	-0.041	-0.011	0.104	1.000					
Institution scientific prominence	-0.045	-0.047	-0.029	0.100	0.990	1.000				
Journal commercial impact	0.338	0.341	0.299	0.056	-0.055	-0.065	1.000			
Journal scientific impact	0.195	0.153	0.244	0.191	0.078	0.059	0.415	1.000		
Author commercial prominence	0.110	0.083	0.258	0.137	0.585	0.566	0.103	0.197	1.000	
Author scientific prominence	-0.006	-0.006	-0.018	0.020	0.746	0.790	-0.010	0.022	0.437	1.000

Panel B. TTO: All articles

	Commercial potential	Scientific potential	Academic cites	Patent cites	Cited by at least one patent	Citing patent is renewed	Author scientific prominence	Disclosed	Author TTO experience
Commercial potential	1.000								
Scientific potential	0.212	1.000							
Academic cites	0.056	0.062	0.000	1.000					
Patent cites	0.106	0.021	0.339	1.000					
Cited by at least one patent	0.261	0.042	0.209	0.354	1.000				
Citing patent is renewed	0.224	0.041	0.201	0.389	0.843	1.000			
Author scientific prominence	0.181	0.240	0.073	0.002	0.009	-0.009	1.000		
Disclosed	0.219	0.049	0.030	0.070	0.129	0.119	0.072	1.000	
Author TTO experience	0.210	0.046	0.027	0.065	0.132	0.121	0.085	0.796	1.000

Panel C. TTO: Articles matched to inventions disclosed to the TTO

	Commercial potential	Scientific potential	Academic cites	Patent cites	Cited by at least one patent	Citing patent is renewed	Autho scientific prominence	Author TTO experience
Commercial potential	1.000							
Scientific potential	0.176	1.000						
Academic cites	0.043	0.053	1.000					
Patent cites	0.120	0.006	0.352	1.000				
Cited by at least one patent	0.206	-0.049	0.251	0.280	1.000			
Citing patent is renewed	0.166	-0.050	0.253	0.321	0.830	1.000		
Author scientific prominence	0.139	0.244	0.064	-0.053	-0.068	-0.104	1.000	
Author TTO experience	0.225	0.062	0.034	0.025	0.072	0.069	0.131	1.000

Table 4: Linear probability model estimating the probability of a paper being cited by at least one renewed patent. Model 1 shows the baseline effects of the fixed effects (publication field-year and university). Model 2 shows the effect of our commercial potential measure, $\phi_{i,t-1}$—a 42.22% increase in explained variation from Model 1. Model 3 contains potential correlates of commercialization outcomes: the commercial and scientific prominence of the originating universities and authors, with prominence measured using the H-index at time $t-1$ (log(H-index$_{t-1}$), as well as commercial and scientific impact journal. Model 4 presents the results with the commercial potential measure ($\phi_{i,t-1}$), and Model 5 adds our scientific potential measure ($\psi_{i,t-1}$) as an additional control. Fixed effects are incorporated at the field-year and university levels in all specifications.

DV: Cited by renewed patent	(1)	(2)	(3)	(4)	(5)
Commercial potential		0.181		0.148	0.142
		(0.019)		(0.015)	(0.015)
High commercial impact institution			0.009	0.007	0.007
			(0.002)	(0.002)	(0.002)
High scientific impact institution			0.002	0.005	0.004
			(0.002)	(0.002)	(0.002)
High commercial impact journal			0.051	0.038	0.039
			(0.005)	(0.004)	(0.004)
High scientific impact journal			-0.011	-0.010	-0.011
			(0.006)	(0.005)	(0.005)
High commercial impact researcher			0.096	0.064	0.064
			(0.008)	(0.006)	(0.006)
High scientific impact researcher			-0.004	-0.001	-0.003
			(0.002)	(0.002)	(0.002)
Scientific potential					0.036
					(0.005)
Constant		-0.015	0.040	-0.023	-0.043
		(0.009)	(0.004)	(0.009)	(0.011)
Publication field - year FE	Yes	Yes	Yes	Yes	Yes
University-FE	Yes	Yes	Yes	Yes	Yes
Observations	5,211,133	5,211,133	5,211,133	5,211,133	5,211,133
R-squared	0.090	0.128	0.116	0.139	0.140

Standard errors clustered at the publication field-year level and the university level

Table 5: Linear probability models estimating, for a publication published at year t, the likelihood of disclosure as a function of commercial potential. The dependent variable is a binary variable indicating whether a paper is associated with an invention disclosed to the TTO. Model 1 presents the baseline impact of the fixed effects (field-year) on disclosure. Model 2 shows that the measure of commercial potential, $\phi_{i,t-1}$, trained with data up to $t-1$, predicts whether a scientific publication will be associated with a disclosure well above the fixed effects. Model 5 presents the full specification, controlling for the scientific potential ($\psi_{i,t-1}$) and the scientific prominence of a publication's authors at time $t-1$ ($log(\text{H-index}_{t-1}+1)$). Fixed effects are included at a publication field-year level in all models.

DV: Disclosed	(1)	(2)	(3)	(4)	(5)
Commercial Potential		0.238		0.232	0.221
		(0.016)		(0.016)	(0.016)
Scientific Potential			0.140	0.034	0.012
			(0.012)	(0.009)	(0.008)
Author Scientific Prominence					0.027
					(0.004)
Constant	0.139	0.015	0.037	-0.007	-0.083
	(0.000)	(0.008)	(0.009)	(0.008)	(0.014)
Publication field - Year FE	Yes	Yes	Yes	Yes	Yes
Observations	96,564	96,564	96,564	96,564	96,564
R-squared	0.025	0.061	0.029	0.061	0.064

Standard errors clustered at the Publication Category - Year level

Table 6: Linear probability model estimating the likelihood that a publication published at time t is associated with an invention that (1) is disclosed to the TTO, (2) receives TTO investment, (3) the TTO files patents for it, (4) leads to commercial agreements, (5) leads to licensing to firms, and (6) generates positive revenue. All dependent variables are binary. Commercial Potential—$\phi_{i,t-1}$, trained with data up to $t-1$—strongly predicts all the outcome variables. The models control for the scientific potential ($\psi_{i,t-1}$) and the scientific prominence of a publication's authors at time $t-1$ ($log(\text{H-index}_{t-1}+1)$). Fixed effects are included at a publication field-year level in all models.

	(1)	(2)	(3)	(4)	(5)	(6)
	Disclosed	Investment	Patent	Agreement	License	Revenue
Commercial Potential	0.221	0.180	0.146	0.137	0.057	0.023
	(0.016)	(0.013)	(0.011)	(0.011)	(0.006)	(0.003)
Scientific Potential	0.012	-0.002	0.002	0.013	0.010	0.013
	(0.008)	(0.007)	(0.006)	(0.006)	(0.005)	(0.003)
Author Scientific Experience	0.027	0.026	0.019	0.026	0.014	0.002
	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)
Constant	-0.083	-0.092	-0.066	-0.089	-0.045	-0.013
	(0.014)	(0.012)	(0.010)	(0.011)	(0.007)	(0.004)
Publication field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	96,564	96,564	96,564	96,564	96,564	96,564
R-squared	0.064	0.058	0.048	0.054	0.026	0.015

Standard errors clustered at the Publication field - Year level

Table 7: Linear probability models estimating the impact of Commercial Potential, $\phi_{i,t-1}$, on commercial outcomes at the invention level. An invention's Commercial Potential is computed as the average commercial potential of the articles associated with it. Models 1 and 2 use as a dependent variable whether an invention receives investment, and Models 3 and 4 whether the TTO files at least one patent. The average commercial potential of the articles tied to an invention strongly predicts both variables. Furthermore, we analyze whether an invention leads to commercial agreements (Model 5) and licensing deals (6); is commercialized via a Startup (7) and commercialized via a Startup with VC funds (8); and generates revenue (9). All dependent variables are binary indicators. We add controls for the author's previous experience disclosing inventions to the TTO (binary) as well as for the authors' scientific prominence at time $t-1$ (log(H-index$_{t-1}+1$)). We also control for the scientific potential of the articles associated with an invention, $\psi_{i,t-1}$. Fixed effects are included at an invention field-year level.

	(1) Invested	(2) Invested	(3) Patented	(4) Patented	(5) Agreement	(6) Licensed	(7) Startup	(8) VC Investment	(9) Any Revenue
Commercial Potential	0.296	0.250	0.315	0.259	-0.106	-0.051	-0.025	0.073	-0.048
	(0.047)	(0.075)	(0.050)	(0.079)	(0.160)	(0.150)	(0.104)	(0.069)	(0.102)
Author TTO Experience		0.119		0.107	-0.203	0.105	0.027	0.128	0.108
		(0.082)		(0.076)	(0.119)	(0.135)	(0.104)	(0.084)	(0.107)
Comm. Pot. x TTO Experience		-0.090		-0.063	0.352	0.008	0.033	-0.100	-0.113
		(0.108)		(0.100)	(0.161)	(0.161)	(0.129)	(0.108)	(0.139)
Author Scientific Prominence		0.042		0.065	0.071	0.020	-0.002	-0.021	-0.003
		(0.015)		(0.017)	(0.029)	(0.030)	(0.023)	(0.020)	(0.023)
Scientific Potential		0.315		0.185	-0.170	-0.211	0.054	0.070	-0.087
		(0.080)		(0.083)	(0.117)	(0.139)	(0.106)	(0.089)	(0.106)
Constant	0.277	-0.123	0.280	-0.104	0.615	0.416	0.102	0.041	0.263
	(0.034)	(0.084)	(0.037)	(0.092)	(0.169)	(0.183)	(0.129)	(0.093)	(0.149)
Invention field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,689	2,689	2,689	2,689	1,305	1,305	1,305	1,305	1,305
R-squared	0.115	0.126	0.125	0.136	0.186	0.132	0.161	0.160	0.173

Standard errors clustered at the Invention Category - Year level

Table 8: Linear probability models estimating the likelihood of a paper being cited by at least one renewed patent, focusing on the commercial potential ($\phi_{i,t-1}$) and its interactions with indicators of high prominence related to institutions, researchers, and journals. High prominence is defined by binary variables indicating if an article's affiliated institution, researcher, or journal ranks in the top 20 percentile of the H-index or journal impact factor. The model includes fixed effects for both field year and university. Interaction terms reveal that publications with high commercial potential are more likely to be cited in renewed patents when associated with high-impact institutions, researchers, or journals.

DV: Cited by renewed patent	(1)	(2)	(3)
Commercial potential	0.164	0.140	0.117
	(0.017)	(0.014)	(0.012)
Scientific potential	0.033	0.035	0.032
	(0.005)	(0.005)	(0.005)
High commercial prominence institution	-0.007	-0.008	-0.004
	(0.008)	(0.007)	(0.006)
Commercial potential x High commercial prominence institution	0.039	0.040	0.023
	(0.018)	(0.017)	(0.015)
High scientific prominence institution	-0.002	-0.003	-0.002
	(0.007)	(0.006)	(0.006)
Commercial potential x High scientific prominence institution	0.021	0.021	0.015
	(0.015)	(0.014)	(0.013)
High commercial impact journal		-0.032	-0.030
		(0.008)	(0.008)
Commercial potential x High commercial impact journal		0.127	0.120
		(0.014)	(0.014)
High scientific impact journal		0.039	0.039
		(0.011)	(0.012)
Commercial potential x High scientific impact journal		-0.081	-0.083
		(0.019)	(0.020)
High commercial prominence researcher			-0.007
			(0.012)
Commercial potential x High commercial prominence researcher			0.098
			(0.019)
High scientific prominence researcher			-0.003
			(0.003)
Commercial potential x High scientific prominence researcher			0.003
			(0.007)
Constant	-0.032	-0.036	-0.031
	(0.010)	(0.009)	(0.009)
Publication field - year FE	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes
Observations	5,211,133	5,211,133	5,211,133
R-squared	0.130	0.137	0.144

Standard errors clustered at the publication field-year level and the institution level

Table 9: Models 1 to 4 estimate the count of different firms citing an article in their patents. High Commercial Potential is a binary variable indicating whether the article is a the top quartile of commercial potential, and Patented is a binary variable indicating whether the article is associated with an invention patented by the TTO. Fixed effects are included at a publication field-year level for all models.

DV: Number of citing firms	(1)	(2)	(3)	(4)
High Commercial Potential	0.062		0.059	0.057
	(0.009)		(0.009)	(0.009)
Patented		0.041	0.028	0.018
		(0.006)	(0.005)	(0.004)
High Commercial Potential x Patented				0.022
				(0.007)
Constant	0.017	0.029	0.015	0.016
	(0.002)	(0.000)	(0.002)	(0.002)
Publication field - Year FE	Yes	Yes	Yes	Yes
Observations	96,564	96,564	96,564	96,564
R-squared	0.073	0.058	0.074	0.075

Standard errors clustered at the Publication field - Year level

Online Appendix to:

Measuring the Commercial Potential of Science

Appendix A Commercial potential model: Training and performance

In this section, we provide additional details regarding the training process and the models' performance.

A.1 Year-based models

Figure A.1 provides a schematic representation of how the training sample is created. As described in section 2.3.1, commercial potential models are developed for each year from 2000 to 2020, resulting in 21 distinct models. Each model is trained independently using data from prior years, avoiding data leakage. This method ensures that only information available up to the year preceding each focal year is used, preventing the models from learning from future data. For each focal year t, articles from the preceding ten years (from $t-14$ to $t-5$) are used for training. These articles are labeled using patent citation and renewal data up to $t-1$. A five-year gap is imposed between article publication and the focal year to allow for patent renewals, ensuring that only commercially relevant articles are included. This approach maintains out-of-sample and out-of-time validity, avoiding bias in estimating commercial potential.

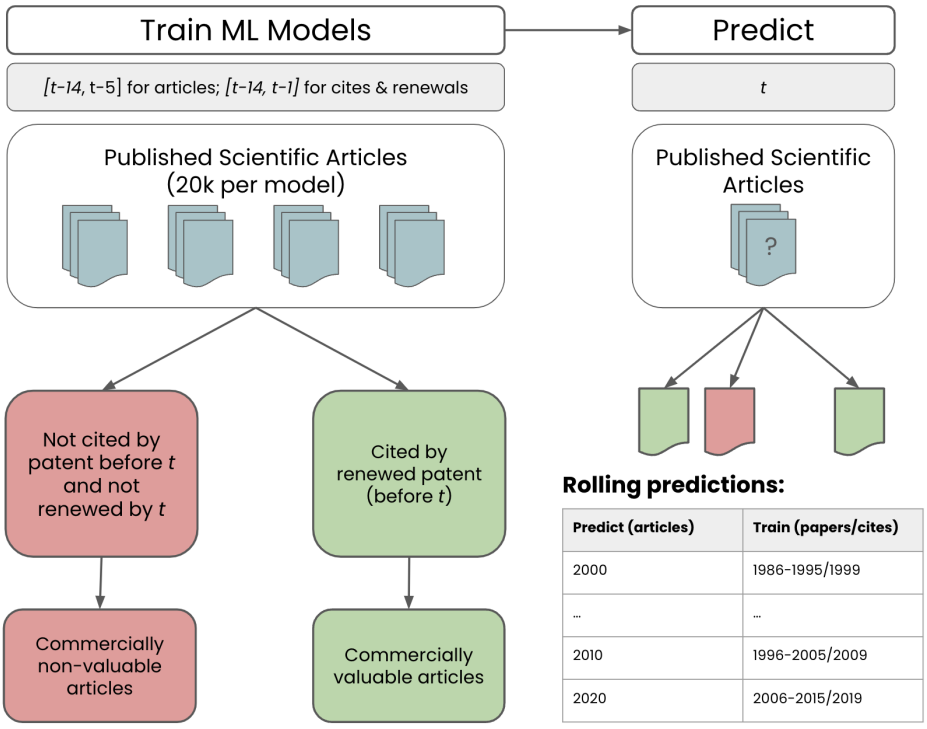


Figure A.1: Schematic representation of year-based training sample construction.

A.2 Processing the input text

Our methodology relies on large language models and natural language processing (NLP) techniques, which use text as input. Specifically, we use the abstracts of the articles in which findings are reported. The pre-trained language model we use is SciBERT (Beltagy et al., 2019), which in turns derives from BERT, a language model created by Google (Devlin et al., 2018). Pre-trained language models, such as BERT and SciBERT, create accurate representations of documents in a high-dimensional space. This is achieved through algorithms that convert text documents into embeddings—numeric vectors serving as representations of the document's content. This capability is highly valuable, as it enables various tasks based on these embeddings. Because its trained with scientific, domain-specific text, SciBERT provides state-of-the-art performance in a wide range of natural language processing tasks for scientific domains, improving BERT's performance. We tested whether this holds in our classification task and, indeed, our models' performance increases when using SciBERT instead of BERT.

SciBERT relies on transformers (Vaswani et al., 2017), a novel type of neural network architecture.[30] In short, as opposed to previous natural language processing techniques, transformers can model long-range dependencies and learn contextual representations, being able to "understand" complex semantic relationships within and across documents.[31] The first step we undertake consists of "tokenizing" the abstracts, i.e., converting each abstract into an array of discrete linguistic units—usually, units are words, parts of words, numbers, symbols, and stems. We tokenize using the version that SciBERT's authors recommend, *scibert-scivocab-uncased*, which is expected to yield the highest performance.[32]

The tokenizer maps each word into an integer based on the model's vocabulary and adds special tokens such as sentence separators, padding, and classification task-specific codes. For each token, the tokenizer looks for its pre-trained embeddings (Token Embeddings)—a vector representing each word in a high-dimensional space in relation to an extensive vocabulary. In addition, the tokenizer adds information regarding the position of each token in the text, both in the sentence (Segment Embeddings) and in absolute terms (Position Embeddings). Combining the

[30] At a high level, a transformer model consists of multiple layers of self-attention and feed-forward neural networks, enabling it to weigh the probabilities of different parts of the input sequence (i.e., sentences of the text) and process it in parallel. The attention mechanisms allow transformers to learn contextual representations of words and phrases.

[31] A possible limitation of our analysis is that the training sample for SciBERT (Beltagy et al., 2019) comprised 82 percent life science articles and 18 percent computer science articles. Although these two fields represent a large share of the entire corpus of published articles, this could represent a limitation given that we are also trying to evaluate the commercial potential of articles from fields other than life sciences and computer science.

[32] SciBERT's tokenizer uses its wordpiece vocabulary based on a subword segmentation algorithm created to match best the corpus of scientific papers used to train the model (*scivocab*) (Beltagy et al., 2019).

three embeddings produces a unique embedding for each token in the abstract, which serves as the input to the first layer of the neural network. This final embedding captures information about the token's relative position within a document, enabling the contextualization of its meaning when fine-tuning the models.

It is worth noting that, for computational reasons, SciBERT, like BERT, is limited to processing up to 512 tokens per document. There are various techniques to handle longer documents, but a simple analysis of the abstracts we use to train our model reveals that only 1% of them contain more than 512 tokens. Additionally, there are no differences in the average number of tokens between the classes (which could create bias in our findings). Therefore, we truncate the abstracts at 512 tokens. Figure A.2 shows the distribution of abstracts' length. Once the abstracts have been processed by the tokenizer, they are input to the neural network and the model is fine tuned based on the labels.

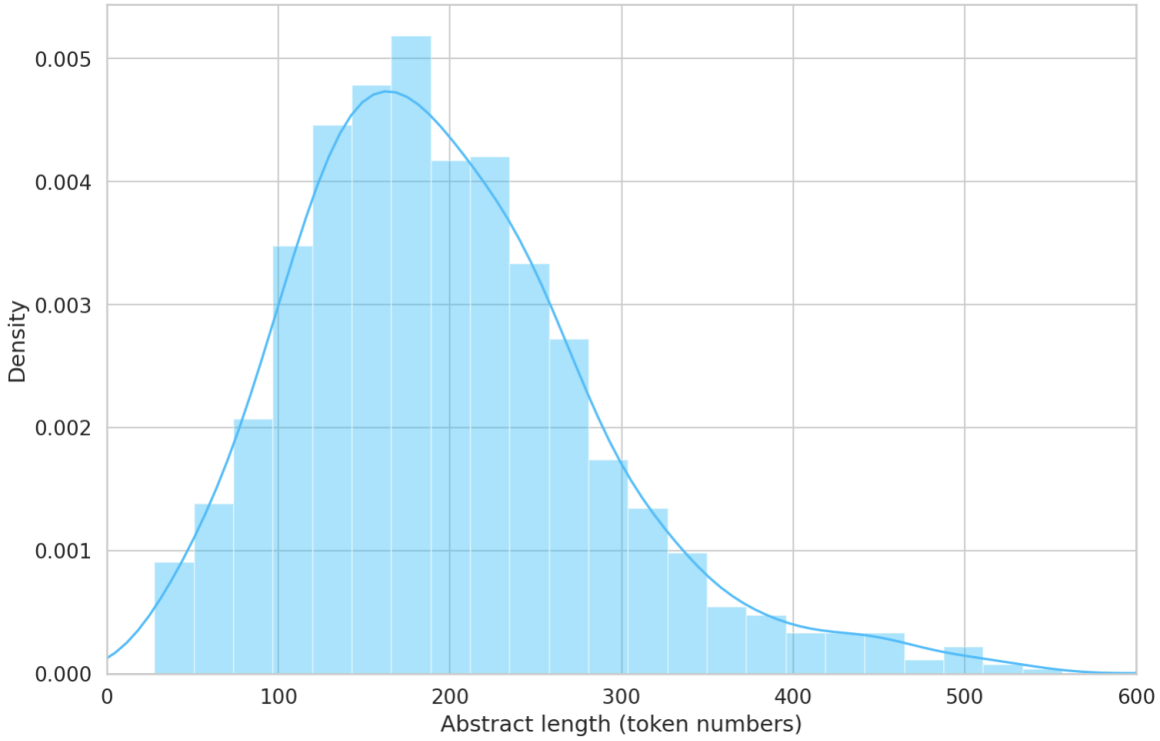


Figure A.2: Abstract's token length distribution

A.3 Models' performance

A.3.1 Commercial potential

Figures A.3 and A.4 detail performance metrics for each of the 21 commercial potential models we trained. We report the performance of our year-based models using several measures: precision, recall, F1-score, accuracy, and the Area Under the Receiver Operating Characteristic (AUROC). The average AUROC across the 21 models is 0.82, ranging from 0.84 to 0.80. The AUROC is a performance metric that measures a model's ability to distinguish between classes. It represents the area under the ROC curve, which plots the true positive rate against the false positive rate at various threshold levels. An AUROC of 1.0 indicates perfect classification, while 0.5 suggests performance equivalent to random guessing. As is common in many machine learning tasks, AUROC is preferred over metrics like accuracy because it does not depend on a fixed classification threshold (e.g., 0.5). In the analysis conducted along the paper, we similarly avoid using a fixed 0.5 threshold for defining commercial potential. Instead, we rely on the continuous variable or percentile indicators, making AUROC a more suitable metric as it better captures the model's overall ability to discriminate between classes across different thresholds.

For context, Manjunath et al. (2021) report an AUROC of 0.83 in their model predicting patent citations of articles. However, they only use PubMed abstracts in the life sciences and do not consider patent renewals. Similarly, Koffi and Marx (2023) employ a BERT-derived measure of science commercializability. However, they do not report sufficient detail to permit comparison with our methods or results. In contrast, Liang et al. (2022) created a model based on the text of inventions disclosed to Stanford's Technology Transfer Office (TTO), aiming to predict commercial value generation, achieving an AUROC of 0.76. While our use of natural language processing is distinct in using academic paper text to predict citations from renewed patents, other work has also implemented NLP models. These papers use patent text and other indicators (e.g., author, patent, and institution characteristics) to predict patent value, measured by forward citations, use in commercialized products, or market responses to patenting by public firms (Chuang et al., 2024; Hsu et al., 2020).

2000

	precision	recall	f1-score	support
Not cited by ren. patent	0.786	0.680	0.729	1224
Cited by ren. patent	0.728	0.823	0.773	1276
Macro avg	0.757	0.751	0.751	2500
Weighted avg	0.757	0.753	0.751	2500
Accuracy	0.753			
AUROC	0.836			

2001

	precision	recall	f1-score	support
Not cited by ren. patent	0.754	0.766	0.760	1267
Cited by ren. patent	0.755	0.744	0.749	1233
Macro avg	0.755	0.755	0.755	2500
Weighted avg	0.755	0.755	0.755	2500
Accuracy	0.755			
AUROC	0.834			

2002

	precision	recall	f1-score	support
Not cited by ren. patent	0.763	0.743	0.753	1263
Cited by ren. patent	0.744	0.764	0.754	1237
Macro avg	0.753	0.753	0.753	2500
Weighted avg	0.753	0.753	0.753	2500
Accuracy	0.753			
AUROC	0.831			

2003

	precision	recall	f1-score	support
Not cited by ren. patent	0.797	0.690	0.739	1308
Cited by ren. patent	0.703	0.807	0.752	1192
Macro avg	0.750	0.748	0.745	2500
Weighted avg	0.752	0.746	0.745	2500
Accuracy	0.746			
AUROC	0.820			

2004

	precision	recall	f1-score	support
Not cited by ren. patent	0.767	0.740	0.753	1245
Cited by ren. patent	0.751	0.777	0.764	1255
Macro avg	0.759	0.758	0.758	2500
Weighted avg	0.759	0.758	0.758	2500
Accuracy	0.758			
AUROC	0.837			

2005

	precision	recall	f1-score	support
Not cited by ren. patent	0.740	0.721	0.731	1241
Cited by ren. patent	0.732	0.751	0.741	1259
Macro avg	0.736	0.736	0.736	2500
Weighted avg	0.736	0.736	0.736	2500
Accuracy	0.736			
AUROC	0.811			

2006

	precision	recall	f1-score	support
Not cited by ren. patent	0.758	0.678	0.716	1207
Cited by ren. patent	0.726	0.798	0.761	1293
Macro avg	0.742	0.738	0.738	2500
Weighted avg	0.742	0.740	0.739	2500
Accuracy	0.740			
AUROC	0.810			

2007

	precision	recall	f1-score	support
Not cited by ren. patent	0.792	0.684	0.734	1280
Cited by ren. patent	0.710	0.811	0.757	1220
Macro avg	0.751	0.748	0.746	2500
Weighted avg	0.752	0.746	0.746	2500
Accuracy	0.746			
AUROC	0.823			

2008

	precision	recall	f1-score	support
Not cited by ren. patent	0.753	0.714	0.733	1248
Cited by ren. patent	0.729	0.766	0.747	1252
Macro avg	0.741	0.740	0.740	2500
Weighted avg	0.741	0.740	0.740	2500
Accuracy	0.740			
AUROC	0.818			

2009

	precision	recall	f1-score	support
Not cited by ren. patent	0.741	0.727	0.734	1252
Cited by ren. patent	0.731	0.745	0.738	1248
Macro avg	0.736	0.736	0.736	2500
Weighted avg	0.736	0.736	0.736	2500
Accuracy	0.736			
AUROC	0.811			

Figure A.3: Commercial potential models' performance (1/2)

2010				
	precision	**recall**	**f1-score**	**support**
Not cited by ren. patent	0.762	0.710	0.735	1262
Cited by ren. patent	0.724	0.774	0.748	1238
Macro avg	0.743	0.742	0.741	2500
Weighted avg	0.743	0.742	0.741	2500
Accuracy	0.742			
AUROC	0.834			

2011				
	precision	**recall**	**f1-score**	**support**
Not cited by ren. patent	0.781	0.656	0.713	1226
Cited by ren. patent	0.713	0.823	0.764	1274
Macro avg	0.747	0.740	0.739	2500
Weighted avg	0.747	0.741	0.739	2500
Accuracy	0.741			
AUROC	0.823			

2012				
	precision	**recall**	**f1-score**	**support**
Not cited by ren. patent	0.744	0.707	0.725	1218
Cited by ren. patent	0.734	0.769	0.751	1282
Macro avg	0.739	0.738	0.738	2500
Weighted avg	0.739	0.739	0.738	2500
Accuracy	0.739			
AUROC	0.812			

2013				
	precision	**recall**	**f1-score**	**support**
Not cited by ren. patent	0.786	0.639	0.705	1200
Cited by ren. patent	0.716	0.839	0.773	1300
Macro avg	0.751	0.739	0.739	2500
Weighted avg	0.749	0.743	0.740	2500
Accuracy	0.743			
AUROC	0.813			

2014				
	precision	**recall**	**f1-score**	**support**
Not cited by ren. patent	0.814	0.622	0.705	1285
Cited by ren. patent	0.680	0.849	0.755	1215
Macro avg	0.747	0.736	0.730	2500
Weighted avg	0.749	0.732	0.729	2500
Accuracy	0.732			
AUROC	0.806			

2015				
	precision	**recall**	**f1-score**	**support**
Not cited by ren. patent	0.788	0.654	0.714	1242
Cited by ren. patent	0.707	0.826	0.762	1258
Macro avg	0.747	0.740	0.738	2500
Weighted avg	0.747	0.740	0.738	2500
Accuracy	0.740			
AUROC	0.816			

2016				
	precision	**recall**	**f1-score**	**support**
Not cited by ren. patent	0.815	0.613	0.700	1248
Cited by ren. patent	0.691	0.861	0.766	1252
Macro avg	0.753	0.737	0.733	2500
Weighted avg	0.753	0.737	0.733	2500
Accuracy	0.737			
AUROC	0.810			

2017				
	precision	**recall**	**f1-score**	**support**
Not cited by ren. patent	0.800	0.607	0.690	1262
Cited by ren. patent	0.678	0.845	0.753	1238
Macro avg	0.739	0.726	0.721	2500
Weighted avg	0.740	0.725	0.721	2500
Accuracy	0.725			
AUROC	0.811			

2018				
	precision	**recall**	**f1-score**	**support**
Not cited by ren. patent	0.789	0.664	0.721	1221
Cited by ren. patent	0.721	0.830	0.772	1279
Macro avg	0.755	0.747	0.747	2500
Weighted avg	0.754	0.749	0.747	2500
Accuracy	0.749			
AUROC	0.820			

2019				
	precision	**recall**	**f1-score**	**support**
Not cited by ren. patent	0.796	0.623	0.699	1225
Cited by ren. patent	0.700	0.847	0.767	1275
Macro avg	0.748	0.735	0.733	2500
Weighted avg	0.747	0.737	0.734	2500
Accuracy	0.737			
AUROC	0.804			

2020				
	precision	**recall**	**f1-score**	**support**
Not cited by ren. patent	0.751	0.708	0.729	1192
Cited by ren. patent	0.747	0.786	0.766	1308
Macro avg	0.749	0.747	0.747	2500
Weighted avg	0.749	0.749	0.748	2500
Accuracy	0.749			
AUROC	0.824			

Figure A.4: Commercial potential models' performance (2/2)

A.3.2 Scientific potential

To measure scientific potential, we adapted the methodology originally designed to assess commercial potential, with two adjustments to suit the scientific context. First, we used academic citations as a measure of a paper's realized scientific potential. To classify papers, we defined a threshold based on the sample median of 16 citations. Papers cited 16 times or fewer are categorized as having *low scientific potential*, while those with more than 16 citations are classified as having *high scientific potential*. Second, while we also trained models based on temporal windows, these were simplified for this exercise. Unlike commercial potential, scientific citations are not subject to truncation issues related to renewals. Consequently, we employed a straightforward moving time window for training, spanning from $t-10$ to $t-1$, and trained 21 models using academic citations truncated to the focal year.

The models demonstrated satisfactory performance, achieving an average accuracy and AUROC of 0.71. We also conducted experiments with varying thresholds and configurations. To maintain consistency with the primary model's training sample, we used a balanced dataset and experimented with different language models and hyper-parameters. Notably, the same configuration that yielded the highest performance for the commercial potential models also produced the best performance for the scientific potential models.

In Table A.1, we report the average performance metrics of the scientific potential models.

Table A.1: Scientific potential model performance (average of all models: 2000-2020)

	Precision	Recall	F1-score
$\leq$ 16 scientific citations	0.73	0.71	0.72
$>$ 16 scientific citations	0.70	0.72	0.71
Accuracy			0.71
Micro-averaged ROC AUC			0.71

A.4 Examples of scientific articles and their commercial potential

Table A.2: Selected scientific articles in the top 25 percentile of commercial potential.

Title	Field	Institution	Journal	Year	Patent cites	Citing patent renewed?
High-resolution mapping of protein sequence-function relationships	Biological Sciences	U. of Washington	Nature Methods	2010	26	Yes
Combination strategies to enhance anti-tumor ADCC	Biomedical and Clinical Sciences	Stanford	Immunotherapy	2012	9	Yes
Engineering Tumor-Targeting Nanoparticles as Vehicles for Precision Nanomedicine	Engineering	Rutgers	Med one	2019	0	No
Species-Specific and Inhibitor-Dependent Conformations of LpxC—Implications for Antibiotic Design	Chemical Sciences	Duke	Chemical Sciences & Biology	2011	6	Yes
Multi-Scale 2D Temporal Adjacency Networks for Moment Localization with Natural Language	Information and Computing Sciences	U. of Rochester	IEEE Transactions on Pattern Analysis and Machine Intelligence	2021	0	No
Nanophotonic projection system	Physical Sciences	California Institute of Technology	Optics Express	2015	8	Yes
Conserved and Divergent Features of Mesenchymal Progenitor Cell Types within the Cortical Nephrogenic Niche of the Human and Mouse Kidney	Biological Sciences	U. of Southern California	Journal of The American Society of Nephrology	2018	0	No
Self-Healing Polyurethanes with Shape Recovery	Engineering	U. of Florida	Advanced Functional Materials	2014	7	Yes
Exploring mechanisms of FGF signalling through the lens of structural biology.	Biological Sciences	New York U.	Nature Reviews Molecular Cell Biology	2013	8	Yes
A high-energy-density sugar biobattery based on a synthetic enzymatic pathway	Chemical Sciences	Virginia Tech	Nature Communications	2014	11	Yes

Table A.3: Selected scientific articles in the bottom 25 percentile of commercial potential.

Title	**Field**	**Institution**	**Journal**	**Year**	**Patent Cites**	**Citing Patent Renewed?**
Extinction and Nebular Line Properties of a Herschel-selected Lensed Dusty Starburst at z = 1.027	Physical Sciences	Cornell University	International Journal of Mass Spectrometry	2015	0	No
An exotic invasive shrub has greater recruitment than native shrub species within a large undisturbed wetland	Biological Sciences	University of Wisconsin	Plant Ecology	2012	0	No
Dynamic programming solutions for decentralized state-feedback LQG problems with communication delays	Information and Computing Sciences	California Institute of Technology	Advances in computing and communications	2012	1	Yes
Effects of natural weathering on microstructure and mineral composition of cementitious roofing tiles reinforced with fique fibre	Engineering	Pennsylvania State University	Cement and Concrete Composites	2011	0	No
Thermodynamic database for the Co-Pr system	Chemical Sciences	Iowa State University	Data in Brief	2016	0	No
Hydrostatic equilibrium profiles for gas in elliptical galaxies	Physical Sciences	Yale University	Monthly Notices of the Royal Astronomical Society	2010	0	No
A Multilevel Quasi-Static Kinetics Method for Pin-Resolved Transport Transient Reactor Analysis	Engineering	U. Michigan	Nuclear Science and Engineering	2016	0	No
Turbulent cross-helicity in the mean-field solar dynamo problem	Physical Sciences	Stanford	The Astrophysical Journal	2011	0	No
A 4-year study of invasive and native spider populations in Maine	Biological Sciences	U. Massachusetts	Canadian Journal of Zoology	2011	0	No
Intrusion of a Liquid Droplet into a Powder under Gravity	Biomedical and Clinical Sciences	Princeton University	Langmuir	2016	0	No

A.5 Performance metrics by scientific field

Figure A.5 presents the average performance metrics (AUROC and F1-Score) for the 21 year-based models across different scientific fields, revealing significant variation. Health sciences, Physical sciences, and Biological sciences all show strong performance with an average AUROC exceeding 0.85. While the F1-scores for Health sciences and Physical sciences are around the overall mean, the F1-score for Biological sciences is notably high, indicating a strong balance between precision (the accuracy of positive predictions) and recall (the ability to identify all relevant positive cases) in predicting commercial potential within this field.

Conversely, Information and Communication sciences exhibit lower performance, with an AUROC of 0.73 and an F1-score of 0.69. This may be due to the inherent nature of the field, where innovations in computer science and software are less frequently patented compared to other disciplines. Additionally, the fast-paced evolution and lower patent propensity in these areas can make it more challenging to capture commercial potential using traditional patent-based indicators.

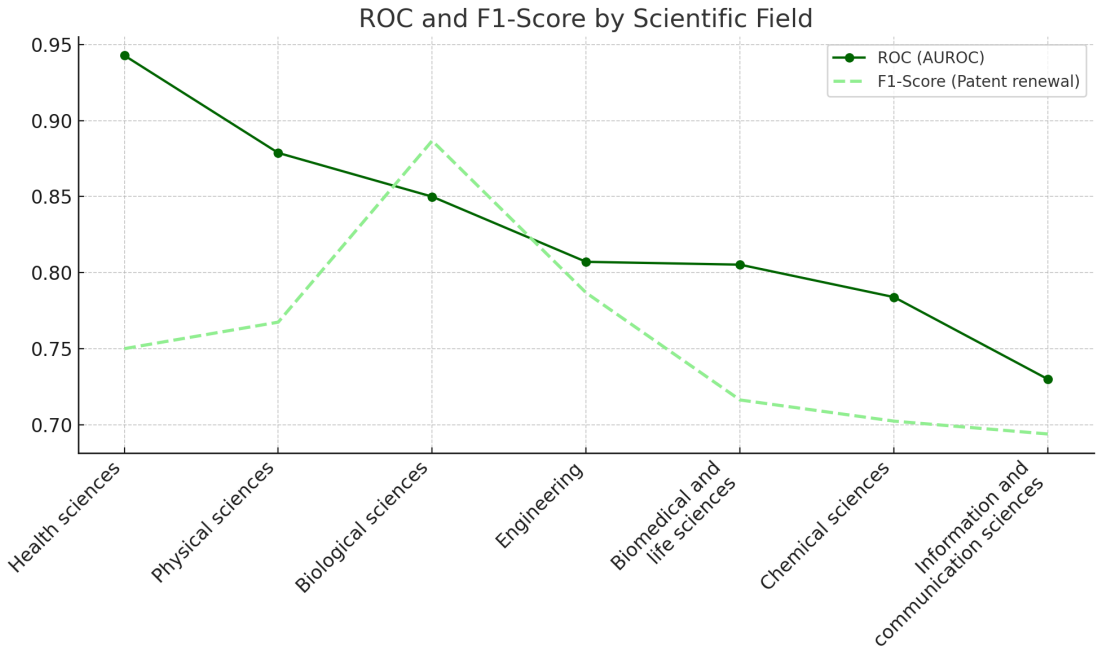


Figure A.5: Performance by field

A.6 Model performance by hyper-parameter and language model

In this section, we analyze the impact of various hyper-parameters and language models on the performance of our classifiers. Specifically, we examine how the models perform with different dropout rates, learning rates, and batch sizes, as well as with three distinct language models. The performance metrics under consideration include precision, recall, and F1-score, as our goal is to optimize the models in terms of achieving a strong balance between precision and recall. In all instances, the area under the receiver operating characteristic curve (AUROC) and accuracy improve when the key metrics—precision, recall, and F1-score—are optimized.

Our models are influenced by multiple variables. In addition to hyper-parameters and the choice of a language model, factors such as training size, year, number of epochs, and the use of a balanced sample may impact performance. Exploring all these variables simultaneously is challenging and would result in thousands of comparisons, with prohibitively high computational costs. Therefore, we adopt an iterative approach, where we explore variables incrementally and assume that certain choices remain constant across other models or variables. This approach allows us to manage complexity and resource constraints while still improving model performance.

We conduct this analysis by focusing exclusively on the year 2000, which exhibits average model performance as shown in Section A.3 and, thus, we consider to be representative. Models are trained following the methodology detailed in the paper and evaluated using a hold-out test sample. We train models based on the following combinations of hyper-parameters: dropout rates of 0.1, 0.2, 0.3; learning rates of 1e-5, 2e-5, 3e-5; batch sizes of 8, 16, and 32; and training sizes of 500, 1,000, 2,000, 5,000, and 10,000. This results in the training of 131 distinct models.[33] Once we obtain the performance metrics for these experiments, we average them based on each hyper-parameter.

We find that a batch size of 32 results in substantially better performance across all evaluated metrics. This may be due to the fact that a larger batch size allows the model to compute more stable gradient estimates during training, leading to more reliable weight updates. Likewise, larger batches can better capture patterns and correlations in the data, which can help improve the model's generalization ability (see Figure A.6a).

We find that a dropout rate of 0.3 yields the best results across all performance metrics, suggesting that a moderate amount of regularization through dropout prevents over-fitting without discarding too much information. As the dropout rate decreases, there is a consistent decline in

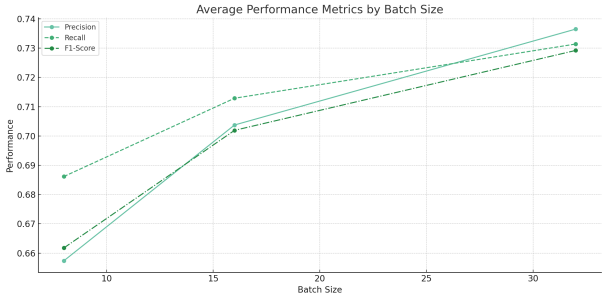
[33]Given the exponential increase in computation time and costs, we did not train models with a sample size of 20,000 (which would require an additional 27 models). As the performance gains when doubling the sample size from 10,000 to 20,000 are minimal (as shown in the next sections) we deemed the additional computation unnecessary.

precision, recall, and AUROC, which suggests that excessive dropout leads to under-fitting (see Figure A.6b).

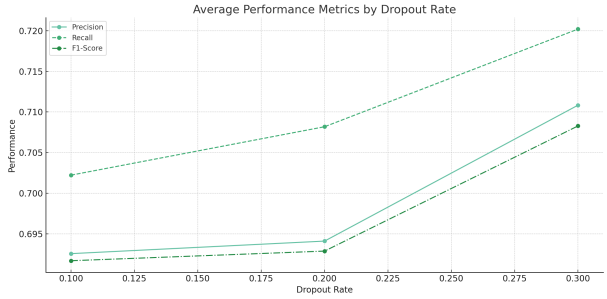
We find that a learning rate of 1e-5 outperforms higher learning rates across all metrics (though the difference is small between 1e-5 and 2e-5). This learning rate allows for more gradual updates to the model weights, leading to improved convergence and generalization. Conversely, a higher learning rate of 3e-05 causes fluctuations in the model's performance, particularly in precision and recall, indicating that the model is making overly aggressive updates during training (see Figure A.6c).

It is worth noting that the analysis uses the average performance metrics across all experiments. These results are consistent also with the maximum performance achieved. Specifically, the model with the highest performance across all considered metrics—precision, recall, F1-score, AUROC, and accuracy—has the following configuration: a dropout rate of 0.3, a learning rate of 1e-05, a batch size of 32, and a training size of 10,000.

(a) Average performance by batch size



(b) Average performance by drop out rate



(c) Average performance by learning rate

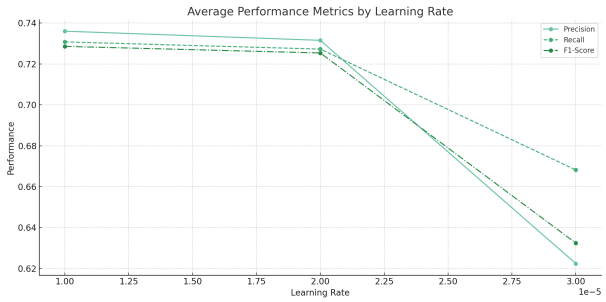


Figure A.6: Comparison of average performance by hyper-parameters (batch size, dropout, and learning rate.

Next, employing this optimal configuration, we analyze the performance across different language models, comparing BERT (Devlin et al., 2018), SPECTER 2.0 (Cohan et al., 2020; Singh et al., 2022), and SciBERT (Beltagy et al., 2019). These three are state-of-the-art models in natural language processing.

BERT (Devlin et al., 2018) is a general-purpose language model pre-trained on a large corpus of diverse text, making it suitable for a wide range of tasks. While its broad focus might limit performance in specialized domains, it could also mitigate overfitting by capturing general language patterns, thus making it more robust when applied to various types of data.

SPECTER 2.0 (Cohan et al., 2020; Singh et al., 2022) is designed specifically for scientific document embeddings, optimizing performance for citation-based tasks. This specialization makes it particularly suited for analyzing research articles and predicting commercial potential from patent citations, as it has been trained to capture relationships between scientific publications more effectively.

Finally, SciBERT (Beltagy et al., 2019) is a variant of BERT pre-trained specifically on scientific text, which enhances its ability to understand the specialized scientific language. However, SciBERT's fine-tuning on a subset of fields—primarily biological and computer sciences—might limit its generalizability across other scientific domains.

The comparison of BERT, SPECTER 2.0, and SciBERT reveals significant differences in performance across various metrics. SciBERT is superior in all metrics except precision, where SPECTER 2.0 performs better. When considering the area under the receiver operating characteristic curve (AUROC), SciBERT outperforms SPECTER by a 2.78% increase and BERT by a 7.46% increase. Similarly, SciBERT shows the highest accuracy, with a 1.45% gain compared to SPECTER and a 7.29% increase over BERT. In terms of the F1-score, SciBERT improves by 4.44% over SPECTER 2.0 and 7.51% over BERT (see Figure A.7).

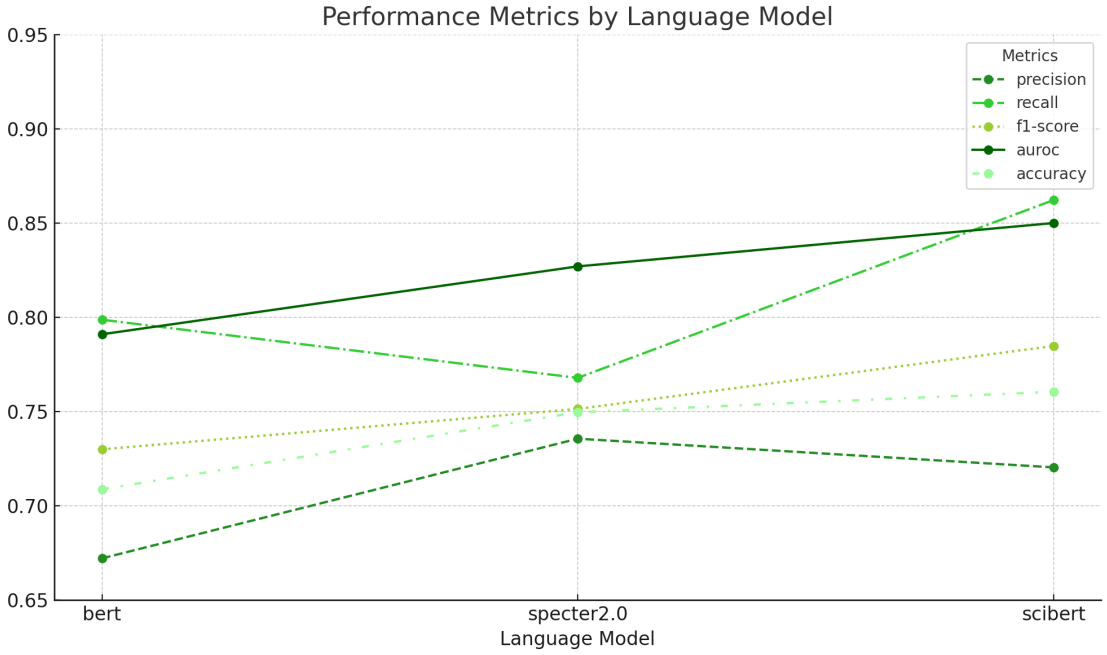


Figure A.7: Language model performance

A.7 Model performance by training size, cross-validation, and balance

Following the same approach as in the previous section, we now analyze the model's performance using different training sizes: 500, 1,000, 2,000, 5,000, 10,000, and 20,000. For this analysis, we fix the other hyper-parameters at their optimal values as per the previous analyses: a dropout rate of 0.3, a learning rate of 1e-05, a batch size of 32, and we use SciBERT as the language model.

Furthermore, to account for potential variations in performance metrics due to the randomly created training samples, we apply cross-validation (CV), a technique used to assess the generalizability of a model by dividing the dataset into multiple subsets—folds. Recall that in our neural network architecture, the model is trained on both the training and validation sets, and then evaluated on a hold-out test set. Under CV, this process is repeated five times.

We partition the training data into five equally sized folds. In each iteration, one fold is treated as the validation set, while the remaining four folds are used as the training and validation set. This process is repeated so that each fold is used as the validation set once, allowing the model to be trained and evaluated on different subsets of the same sample size.

In summary, for each defined sample size, we train five models, each with a different composition of training, validation, and test sets. The results from these five models are then averaged to yield a more robust estimate of the model's performance.

As expected, larger training sizes result in improved performance across all metrics. The model's ability to generalize increases as more training data is provided, leading to more reliable and consistent predictions. While small sample sizes may occasionally show high performance, the cross-validation (CV) exercise reveals that this is primarily due to the limited sample size on which the model is evaluated. Smaller training sizes exhibit high variability in the performance metrics, making it difficult to draw any conclusive insights from them. Finally, diminishing returns are observed as the training size becomes sufficiently large, indicating that there is little need to increase the sample size beyond 20,000 observations.

In Figure A.8, we plot both the ROC and F1-score metrics for each fold in the cross-validation process, alongside the overall average. This visualization helps to illustrate not only the individual performance of each fold but also the consistency of the model across different subsets of the data. By plotting the averages, we can clearly observe trends in model performance as training sizes increase, allowing us to assess how well the model generalizes and how the variability in performance decreases with larger training sets.

Finally, it is worth considering the effect of balancing the training sample across classes on model performance. To investigate this, we once again use the optimal conditions identified earlier and examine the model's performance across different sample balances. By doing so, we aim to

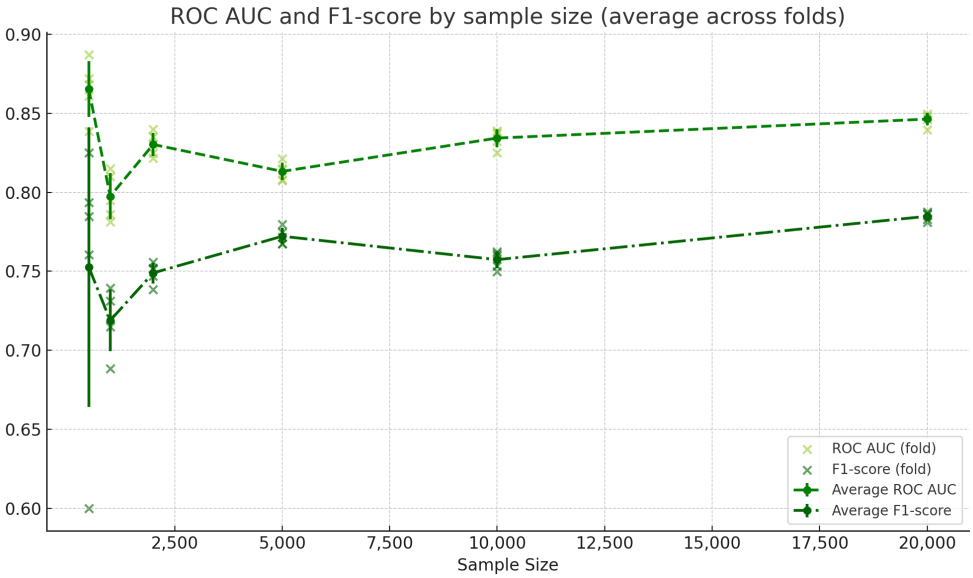


Figure A.8: Training size

understand how varying the balance between classes impacts key metrics such as precision, recall, and F1-score.

As Figure A.9 shows, increasing the balance of the training sample leads to more reliable and consistent performance across all metrics. Notably, the model achieves its most balanced performance in terms of F1-score when the sample is well-balanced, indicating that a balanced dataset enables the model to capture relevant patterns more effectively. This observation aligns with established research in machine learning and neural networks, confirming the well-known importance of data balance in achieving optimal model performance (Miric et al., 2022).

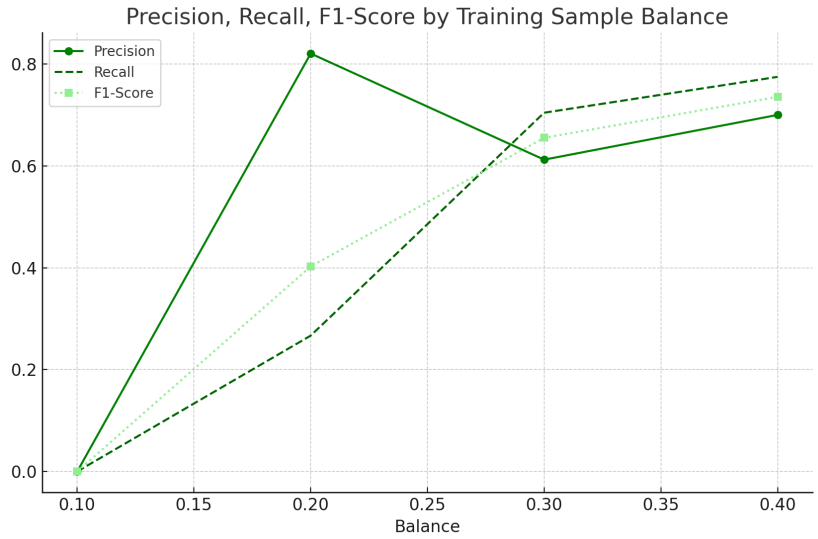


Figure A.9: Balance

Appendix B Model limitations and robustness

While our classifier demonstrates reasonable performance, there is potential for further enhancement. Three primary factors may be influencing its performance:

1. **Diverse Academic Fields**: Our models are trained to classify articles across various fields, from Biology to Engineering to Computer Science. Textual features indicative of commercial potential may significantly vary between these disciplines. This diversity necessitates compromises in parameter settings, consequently limiting the model's overall performance. For comparison, Manjunath et al. (2021) focused their model exclusively on the life sciences and biomedical fields, utilizing over 20 million articles from PubMed. They achieved an AUROC of 0.83, highlighting the benefits of field-specific models.
2. **Complexity of Task**: Predicting commercial potential from textual data is inherently complex and uncertain, making it challenging even for expert human analysis. While most natural language processing (NLP) classification tasks, such as identifying specific emotions in text, report accuracies above 95%, these tasks typically involve more straightforward information within the text. For more complex tasks, lower performance is expected. For instance, Liang et al. (2022) trained two NLP models to predict the financial success of inventions disclosed to Stanford's Technology Transfer Office. Their BERT-based model achieved an AUROC of 0.76, while the simpler TF-IDF-based model reached 0.71. Similarly, Guzman and Li (2023) used doc2vec to predict the early-stage success of startups and reported AUROCs between 0.60 and 0.65.
3. **Changing language**: The language signaling commercial potential may change over time, and our model is confined, per above, to a circumscribed sample period. This focus narrows our model's capacity to capture the nuanced dynamics of token emergence, usage, and interconnections and the detailed content in full texts, tables, and figures of articles that may affect the accuracy of our model predictions.

B.1 Commercial potential articles and most representative words

A concern is whether the classification task is influenced by the use of certain words that are not fundamentally related to the scientific content, but may superficially suggest greater commercial potential in a scientific contribution. For instance, the model could disproportionately classify abstracts with a "commercial flavor" as having higher commercial potential. In this scenario, the primary determinant of the results would be the language employed rather than the intrinsic scientific research and its potential commercial applications. That is, the model's predictions could be biased.

Indeed, previous studies suggest that the way ideas are written, beyond the inherent qualities being conveyed, influences the adoption of novel concepts and products. For example, Lakshminarayanan et al. (2017) and Koffi and Marx (2023) identify gender differences in how researchers describe their findings, with men tending to use more positive or boastful language—although the mechanisms behind these differences remain unclear. In a related domain, Zunino et al. (2019) finds that technological products described with overly familiar or creative language are less likely to be adopted, while Pontikes (2012) show that ambiguous wording and concepts can make new products less appealing to consumers, hindering adoption. However, these studies focus on text from firms' press releases in industries like smartphones and software, where claims are not formally evaluated, leaving the incentives to emphasize commercialization unclear. In contrast, scientific authors should have little motivation to exaggerate the commercial aspects of their work, as the peer-review process would likely moderate such claims.

Whether our commercial potential measure is subject to such bias is ultimately an empirical question. To investigate this, we conduct two additional analyses to assess whether specific language usage affects the commercial potential measures generated by our algorithm.

We begin with an exploratory analysis to identify the words most likely to be associated with each class in our classification task. This analysis is interesting in and of itself, as it reveals which words are most indicative of commercial potential or the lack thereof. Furthermore, it helps detect whether our measure is biased toward commercially oriented language. If words with commercial connotations appear more frequently in abstracts classified as having commercial potential, it may suggest a potential bias, indicating that the mere presence of these words could be driving the algorithm's classification results. Additionally, since the terminology used to convey commercial applications may vary across fields, we conduct this analysis at the field level

To analyze the distinctive vocabulary associated with different classes in our dataset, we employed two different, complementary methods. It is important to recognize, however, that the use of transformer models and neural networks in calculating the commercial potential measure means that the ranked words resulting of the following two exercises are merely indicative of how the presence or absence of specific words might influence the score for a given abstract. In practice, the impact of these words can vary significantly depending on their combinations, semantic similarities, and contexts within the abstracts—which is precisely the advantage of relying on neural networks over bags of words—, leading to different results. This is further explored in the next section, B.2.

B.1.1 Words more likely in one class *vs.* the other: Log-odds ratio

First, we run a log-odds ratio calculation. This method enables us to quantify the relative importance of words within each class compared to their overall distribution across the entire corpus. That is, with this method we are able to detect words that are disproportionately likely to be present in one class versus the other. We proceed as follows.[34]

First, we randomly sample 100,000 abstracts from our sample of articles to avoid large computational costs, and we classify abstracts as having commercial potential if they are in the top 20% of commercial potential score, normalized by year and field. Next, we process the abstracts. We remove stop words, punctuation and numbers, and sparse terms to reduce noise (those terms occurring less than 120 times across the entire corpus, representing a frequency below 0.1%). Next, we compute the probability of each word occurring within a given class by dividing its count in a class by the total word count for that class.[35] Next, we calculate the overall probability of each word across all classes. Finally, we compute the log-odds ratio by taking the natural logarithm of the ratio between the word's class-specific probability and its overall probability.[36]

Tables B.1 and B.2 present the results. For each scientific field, we plot the top 15 words most likely to appear in one class versus the other in terms of the centered log-odds ratios ("Log-Odds"). For example, under Biological Sciences, "reprogramming" and "metastatic" are words over-represented in abstracts classified as "Commercial potential" while little present in articles classified as "No commercial potential". Conversely, "dispersal" and "nest" are words more likely to appear in articles with "No commercial potential" than otherwise.

[34] We make the code available in our public repository, under the file *compot_words.ipynb*. The code contains all the steps detailed here with further detail.

[35] We apply Laplace smoothing by adding a small constant ($\alpha = 1$) to each word count. This step avoids zero probabilities, which could otherwise lead to undefined logarithmic values during subsequent calculations.

[36] To further refine these values, we center the log-odds ratios by subtracting the mean log-odds across all classes for each word. This centering step allowed us to focus on the deviation of a word's association with a particular class relative to the other classes, providing a more intuitive understanding of class-specific vocabulary.

Table B.1: Most representative words per class, Log-Odds Ratio (1/2)

Commercial potential		**No commercial potential**	
Word	Log-Odds	Word	Log-Odds
Agricultural, Veterinary and Food Sciences			
vaccine	1.45	litter	1.82
proteins	0.92	annual	1.44
assay	0.90	cover	1.39
gene	0.87	fertilizer	1.35
viral	0.87	irrigation	1.33
vitro	0.85	forests	1.26
genome	0.84	plots	1.23
genes	0.83	trees	1.21
virus	0.82	season	1.14
cells	0.78	forest	1.12
molecular	0.76	soil	1.12
mice	0.75	tree	1.02
isolate	0.72	ecosystem	1.01
expression	0.71	year	0.99
sequence	0.71	survey	0.98
Biological Sciences			
reprogramming	1.09	dispersal	2.30
metastatic	1.08	nest	2.23
egfr	1.04	forests	2.21
antiviral	0.96	climate	2.16
metastasis	0.94	season	2.15
engineered	0.91	america	2.10
therapeutics	0.90	ocean	2.08
oncogenic	0.87	predators	2.07
abeta	0.85	lake	2.06
glycosylation	0.85	breeding	2.05
site-specific	0.83	north	2.05
pik	0.83	land	2.04
angiogenesis	0.82	spring	2.03
nucleic	0.82	summer	2.03
viral	0.81	characters	2.02
Biomedical and Clinical Sciences			
xenografts	1.72	residents	2.61
transgene	1.67	educational	2.33
neutralizing	1.54	youth	2.27
aav	1.50	interviews	2.26
plasmid	1.50	aor	2.25
ctl	1.50	attitudes	2.17
foxp	1.47	schools	2.10
antigen-specific	1.45	school	2.09
mscs	1.39	income	2.06
ligands	1.39	faculty	2.04
gag	1.38	attending	2.03
engineered	1.36	fistula	2.01
xenograft	1.35	mace	1.98
mirnas	1.32	psychological	1.98
kinases	1.32	lvad	1.97

Commercial potential		**No commercial potential**	
Word	Log-Odds	Word	Log-Odds
Chemical Sciences			
potent	1.22	orbital	1.80
inhibitors	1.21	relaxation	1.51
therapeutic	1.08	equation	1.35
targets	0.93	dft	1.32
receptor	0.92	ground	1.30
delivery	0.92	agreement	1.19
discovery	0.90	calculation	1.17
vivo	0.89	symmetry	1.10
inhibitor	0.89	neutron	1.02
probes	0.88	calculations	0.99
platform	0.87	parameter	0.98
cancer	0.84	bands	0.96
agents	0.81	calculated	0.94
peptides	0.79	energies	0.92
analogues	0.78	polarization	0.90
Earth Sciences			
algorithms	1.27	holocene	1.88
oil	1.15	rift	1.85
shale	1.01	arc	1.79
detection	0.99	lava	1.78
acid	0.99	pleistocene	1.72
permeability	0.99	antarctic	1.60
image	0.98	monsoon	1.46
classification	0.97	lithosphere	1.43
wells	0.89	younger	1.43
mission	0.84	belt	1.41
retrieval	0.84	century	1.34
particle	0.83	phytoplankton	1.34
fracture	0.83	warm	1.32
nucleation	0.81	warming	1.31
pore	0.80	subduction	1.30
Engineering			
hydrogel	1.74	soil	2.32
scaffold	1.45	asphalt	1.95
hydrogels	1.40	fire	1.91
scaffolds	1.34	flames	1.89
regeneration	1.24	weld	1.77
microfluidic	1.21	sand	1.70
culture	1.08	boiling	1.64
vitro	1.04	turbulence	1.58
video	1.01	equations	1.55
collagen	1.01	concrete	1.53
cmos	0.99	rotor	1.47
engineered	0.98	fatigue	1.42
differentiation	0.94	turbine	1.40
nanofibers	0.94	flame	1.37
drug	0.93	seismic	1.32

Table B.2: Most representative words per class, Log-Odds Ratio (2/2)

Commercial potential		**No commercial potential**	
Word	Log-Odds	Word	Log-Odds
Environmental Sciences			
toxicity	0.52	biodiversity	1.69
treatment	0.51	forest	1.27
exposure	0.50	local	0.98
concentration	0.44	coastal	0.98
soils	0.39	conservation	0.91
potential	0.38	climate	0.72
samples	0.35	national	0.63
fish	0.34	change	0.59
contaminated	0.33	areas	0.58
conditions	0.31	scenarios	0.57
systems	0.30	sediment	0.54
measured	0.29	ecological	0.54
food	0.29	management	0.53
including	0.27	spatial	0.52
air	0.26	biomass	0.51
Health Sciences			
gait	1.34	violence	1.99
alzheimers	1.32	girls	1.86
muscle	1.29	schools	1.86
muscles	1.20	stigma	1.76
devices	1.19	inequalities	1.71
speed	1.07	thematic	1.71
ankle	1.05	end-of-life	1.68
knee	1.01	readiness	1.68
device	0.97	household	1.56
peak	0.94	parental	1.48
driving	0.85	climate	1.44
nicotine	0.79	coping	1.42
subjects	0.79	low-income	1.37
parameters	0.78	semi-structured	1.36
detection	0.77	poverty	1.32
Information and Computing Sciences			
malware	0.84	students	2.36
malicious	0.75	librarians	1.78
query	0.69	education	1.49
queries	0.67	behaviour	1.46
enables	0.61	genetic	1.38
scalable	0.59	fuzzy	1.36
camera	0.58	influence	1.24
device	0.57	science	1.19
retrieval	0.57	workshop	1.18
overhead	0.54	academic	1.12
speech	0.53	engineering	1.04
video	0.49	theory	1.03
protocols	0.49	university	0.84
semantic	0.48	course	0.84
authentication	0.48	numerical	0.81

Commercial potential		**No commercial potential**	
Word	Log-Odds	Word	Log-Odds
Mathematical Sciences			
design	1.29	spin	2.02
algorithms	0.96	decays	1.70
efficient	0.87	algebra	1.45
bayesian	0.87	theories	1.38
regression	0.84	gauge	1.38
optimization	0.82	algebras	1.15
algorithm	0.81	decay	1.15
computational	0.76	quantum	1.04
estimation	0.74	formula	1.03
network	0.70	invariant	1.00
propose	0.69	symmetry	0.99
networks	0.69	conjecture	0.97
modeling	0.62	mass	0.92
markov	0.61	let	0.89
treatment	0.59	operators	0.89
Physical Sciences			
tissue	1.69	stellar	2.29
phantom	1.68	photometric	2.10
graphene	1.49	gev	2.09
imrt	1.46	halo	2.06
tumor	1.24	galaxies	2.01
delivery	1.22	planets	2.01
microscopy	1.22	planet	1.97
lung	1.18	cosmological	1.93
technology	1.18	higgs	1.92
device	1.17	photometry	1.91
carrier	1.15	flare	1.90
scanning	1.12	star	1.90
planning	1.07	redshifts	1.89
films	1.06	boson	1.89
patients	1.06	mag	1.88

Across all eleven scientific fields, we find that the most prominent words in articles classified as having "Commercial potential" are predominantly scientific or technical in nature, with few words related to generic terms or commercialization aspects. For example, a relevant set of words speaks to potential problems to address, such as "virus", "metastasis", "cancer", "detection", "regeneration", "contaminated", "alzheimers", "retrieval", "estimation", and "lung". A second predominant set of words refers to techniques, such as "vitro", "delivery", "markov", and "microscopy". Finally, another set of concepts refers to a potential solution, such as "vaccine", "therapeutics", "antiviral", "nanofibers", "graphene", and "networks".

Overall, we find little evidence that the classification is based on generic or commercially oriented words. Note, however, that this exercise relies only on simple counts and is not indicative of how these words drive the aggregate results. For example, words like "patients", "results", "study", "using", "data", "cells", and "model" appear in more than 25,000 abstracts, but most of them are not present on the words listed in tables B.1 and B.2. Conversely, the word "xenografts", present in table B.1 under Biomedical and Clinical Sciences with a high log-odds ratio (1.72), appears, in all its forms, in 425 documents, the majority classified as "Commercial potential" articles. However, this word is unlikely to drive any of our results, since it only appears 425 times out of 100,000. Note that we did impose a threshold on the word frequency to avoid sparse terms (over 110 times, or more than 0.1%). However, this threshold is rather arbitrary and is difficult to set it in a rigorous way. To better understand what words are most *predictive* of commercial potential, we additionally run the following exercise.

B.1.2 Words most predictive of commercial potential: Linear regression

To get a sense of which words are most predictive of commercial potential, taking into account their overall frequency, we conduct a text-based regression analysis—linear probability model (LPM). As in the previous exercise, we run the analysis that follows by field, using the same sample of 100,000 abstracts, randomly selected from the main dataset.[37]

We first pre-process the text by cleaning it and removing common English stopwords. The cleaned text is then transformed into a Document-Term Matrix (DTM), representing the frequency of words across documents. To reduce noise, sparse terms are removed, and documents with no words are excluded to ensure that only meaningful data is retained. The resulting matrix represents each document by its word frequencies, which serve as the independent variables (X) in the regression analysis. The dependent variable (y) represents commercial potential, classified

[37] We also make this code available in our repository, under the file 'bow_compot_words.R'.

as 1 if the document falls in the top 20th percentile of commercial potential, normalized by field and year.

The goal of the regression analysis is to identify the terms most predictive of commercial potential. To achieve this, we employ an iterative technique using a forward stepwise selection method (fast_forward with BIC crietrion).[38] This approach iteratively adds variables (words) to the model based on their explanatory power. During each iteration, the algorithm evaluates which word, if added, would provide the greatest improvement in explanatory power. The process continues until no further improvement is possible or the maximum number of variables is reached. For ease of interpretability and to be able to compare the results with those of the previous exercise, where we had 30 words per field, we set the maximum number of variables to 30.

Results are presented in Tables B.3 to B.6. For each scientific field, we display the words identified by the iterative algorithm as most predictive of commercial potential, along with their corresponding coefficients. We do not report standard errors as these are all negligible. It is important to note that this method selects the words most predictive of commercial potential, regardless of the direction of the prediction. Therefore, one needs to pay attention to the sign of the coefficient to determine whether the word predicts "Commercial potential" or "No commercial potential". Additionally, we provide the variance explained by each LPM at the bottom of the table.

In this case, the results differ somewhat from the previous exercise. While many words are scientific and technical in nature, such as "cancer", "viral", "cell", "virus", "enzyme", "clinical", "muscle", "quantum", and "crystal", the frequency of more generic words is notably higher in this analysis. Words like "capable", "demonstrated", "able", "critical", "different", "influence", and "significant" show some of the largest coefficients. Although these terms do not directly convey commercialization concepts, they do emphasize the importance and relevance of the results.

This suggests that words like these are indeed present in articles classified as having commercial potential. However, the question of whether this language drives the classification (and thus introduces bias) or if it reflects or is confounded with the underlying scientific ideas remains open. Specifically, we do not yet know whether articles using this language are genuinely more promising, either scientifically or commercially. Notably, the variance explained by these words in predicting commercial potential in prominent fields such as Chemical Sciences, Engineering, Computing Sciences, Biological Sciences, and Biomedical and Clinical Sciences ranges from 4.7%

[38] We set the criterion for the optimization process as the Bayesian Information Criterion (BIC), a statistical measure used to balance model fit and complexity. The BIC-based forward selection allows the algorithm to efficiently choose a subset of words that best predict the commercial potential while avoiding the risk of overfitting by including irrelevant or redundant terms.

to 10.5%.

It is important to recall that our measure is based solely on the text of the abstracts. This means that only words can explain the commercial potential in our model, leaving 90% to 95% of the variation in commercial potential scores explained by the remaining words in the abstracts and their interactions. To address this issue, we devised a novel experimental method, presented in the next section.

Table B.3: Most representative words per class, Bag-of-words regression (1/4)

Agri., Vet., and Food Sciences		**Biological Sciences**		**Biomedical and Clinical Sciences**	
word	**estimate**	**word**	**estimate**	**word**	**estimate**
probes	0.310	biochemical	0.088	including	0.052
capable	0.163	critical	0.049	multiple	0.046
cancer	0.124	family	0.043	developed	0.037
chain	0.121	via	0.038	however	0.021
ability	0.091	roles	0.030	analyzed	-0.005
able	0.082	activity	0.026	data	-0.006
demonstrated	0.080	cells	0.023	failure	-0.008
viral	0.070	different	-0.004	patients	-0.009
humans	0.068	four	-0.017	associated	-0.013
products	0.063	field	-0.018	introduction	-0.018
cell	0.061	influence	-0.019	determine	-0.019
specific	0.057	relative	-0.023	five	-0.020
virus	0.050	significant	-0.025	duration	-0.022
potential	0.048	conditions	-0.026	assessed	-0.023
analysis	0.036	affect	-0.027	highest	-0.025
detected	0.025	rates	-0.027	differences	-0.026
species	-0.021	overall	-0.029	methods	-0.029
study	-0.025	species	-0.029	incidence	-0.030
treatments	-0.028	order	-0.031	evaluate	-0.031
average	-0.038	determine	-0.033	center	-0.031
significant	-0.044	related	-0.034	patient	-0.034
management	-0.044	time	-0.037	groups	-0.035
conducted	-0.053	observed	-0.044	academic	-0.036
areas	-0.055	behavior	-0.044	age	-0.037
relationship	-0.070	photosynthetic	-0.047	admitted	-0.041
included	-0.071	total	-0.049	conclusions	-0.044
usa	-0.080	assessed	-0.049	medical	-0.044
southern	-0.084	conducted	-0.057	lowest	-0.048
sampled	-0.086	decreased	-0.063	appropriate	-0.049
abstract	-0.094	abstract	-0.082	choice	-0.051
R-squared: 0.145		**R-squared: 0.072**		**R-squared: 0.066**	

Table B.4: Most representative words per class, Bag-of-words regression (2/4)

Chemical Sciences		**Earth Sciences**		**Engineering**	
word	**estimate**	**word**	**estimate**	**word**	**estimate**
efforts	0.240	organization	0.394	flexible	0.101
describe	0.114	processing	0.184	overcome	0.085
toward	0.113	mission	0.129	composed	0.078
expression	0.086	become	0.115	efficient	0.036
simple	0.074	challenges	0.082	show	0.025
binding	0.067	existing	0.078	field	-0.013
containing	0.056	work	0.068	material	-0.013
can	0.032	address	0.067	stress	-0.026
reaction	-0.017	methods	0.051	presented	-0.027
characterized	-0.018	images	0.050	soil	-0.029
structure	-0.019	paper	0.046	results	-0.033
energy	-0.022	can	0.037	initial	-0.034
crystal	-0.027	information	0.037	analytical	-0.036
hydrogen	-0.028	different	0.029	behavior	-0.036
present	-0.032	new	0.023	function	-0.037
state	-0.036	data	0.016	boundary	-0.042
temperature	-0.038	observed	-0.022	laboratory	-0.042
found	-0.038	layer	-0.027	particular	-0.043
determined	-0.039	area	-0.028	formulation	-0.045
diffraction	-0.039	river	-0.030	analysis	-0.045
xray	-0.039	local	-0.030	distributions	-0.046
analysis	-0.041	warming	-0.034	criteria	-0.048
dynamics	-0.042	basin	-0.036	tests	-0.050
experimental	-0.045	recent	-0.042	taken	-0.053
spectroscopy	-0.045	regional	-0.050	examined	-0.060
measured	-0.052	period	-0.052	identified	-0.063
constant	-0.060	scales	-0.056	predict	-0.064
energies	-0.066	suggest	-0.059	criterion	-0.065
agreement	-0.070	strong	-0.059	investigate	-0.067
terms	-0.071	observatory	-0.176	numerically	-0.093
R-squared: 0.100		**R-squared: 0.134**		**R-squared: 0.047**	

Table B.5: Most representative words per class, Bag-of-words regression (3/4)

Environmental Sciences		**Health Sciences**		**Computing Sciences**	
word	**estimate**	**word**	**estimate**	**word**	**estimate**
polymers	0.457	faster	0.113	preserving	0.162
transferred	0.441	generate	0.112	commodity	0.160
cellular	0.417	muscle	0.104	appearance	0.117
volatile	0.348	commonly	0.082	efficiently	0.086
crucial	0.192	available	0.064	propose	0.080
physicochemical	0.186	performed	0.060	encryption	0.074
enzyme	0.165	common	0.048	highly	0.070
recovered	0.151	clinical	0.044	applications	0.065
ratio	0.141	body	0.043	challenges	0.059
median	0.130	measured	0.037	protocols	0.053
presence	0.129	test	0.033	devices	0.048
biosolids	0.129	can	0.029	techniques	0.047
chemicals	0.103	data	0.024	object	0.046
degradation	0.095	time	0.022	across	0.043
exposed	0.083	performance	0.018	services	0.036
biological	0.082	measures	0.017	image	0.035
including	0.074	physical	-0.021	using	0.033
using	0.063	nursing	-0.023	users	0.027
chemical	0.049	nurses	-0.026	multiple	0.026
potential	0.032	among	-0.028	existing	0.024
adsorption	0.029	implementation	-0.030	service	0.021
management	-0.016	education	-0.033	model	-0.018
change	-0.027	practice	-0.037	results	-0.032
data	-0.028	article	-0.038	research	-0.035
sediment	-0.032	national	-0.045	article	-0.044
ecological	-0.034	experience	-0.045	development	-0.048
quality	-0.034	findings	-0.046	give	-0.067
biodiversity	-0.047	completed	-0.048	theory	-0.069
forest	-0.048	participation	-0.051	solve	-0.091
waters	-0.095	explore	-0.055	considering	-0.107
R-squared: 0.281		**R-squared: 0.117**		**R-squared: 0.105**	

Table B.6: Most representative words per class, Bag-of-words regression (4/4)

Mathematical Sciences		**Physical Sciences**	
word	**estimate**	**word**	**estimate**
biological	0.314	photonic	0.215
collection	0.163	design	0.143
inference	0.144	grid	0.123
finding	0.122	reduce	0.117
experiments	0.119	lasers	0.109
design	0.103	demonstrate	0.103
network	0.103	imaging	0.095
bayesian	0.085	diameter	0.088
article	0.076	crystal	0.087
experimental	0.069	based	0.070
interest	0.066	images	0.066
may	0.060	various	0.065
problems	0.059	quantum	0.027
data	0.058	simulation	0.026
can	0.052	also	-0.017
properties	0.050	component	-0.018
method	0.046	density	-0.021
provide	0.044	lines	-0.021
different	0.043	model	-0.023
problem	0.041	star	-0.028
develop	0.027	stellar	-0.037
algebra	-0.024	limit	-0.040
action	-0.029	observed	-0.043
prove	-0.030	present	-0.047
certain	-0.051	turbulence	-0.056
theory	-0.053	find	-0.059
infinite	-0.060	mag	-0.065
invariant	-0.061	decay	-0.068
special	-0.082	functions	-0.070
characterization	-0.136	facility	-0.094
R-squared: 0.226		**R-squared: 0.152**	

B.2 Revamping abstracts' commercial orientation with an LLM

While in the previous section we show that words associated with the commercial potential class are mostly scientific or technical in nature, we do find that certain words are more generic. However, this analysis does not fully address whether commercially oriented words are influencing the results. One the one hand, to classify an abstract, our methodology employs transformers, which consider combinations of words and their semantic meaning within a document, rather than individual words alone. On the other, it is also possible that these words appear only in abstracts that are indeed commercially promising, particularly since these are abstracts of papers published in peer-reviewed journals and that, thus, are potentially tuned down.

To conduct a more comprehensive analysis, we use a Large Language Model—OpenAI's ChatGPT—to re-write a random sample of abstracts and make them more "commercially appealing". This process allows us to compare the measure and its predictive performance taking into account at the same time both the underlying scientific promise of the abstract and how the abstract is written in its entirety.

We proceed as follows. First, we randomly sample 50,000 abstracts from the scientific papers detailed in Section 2, for the years 2010 to 2020.[39] Next, we revamp these 50,000 abstracts with ChatGPT, via API. The abstracts were revamped in June 2023 using ChatGPT's version gpt-3.5-turbo-0613. The specific prompt provided to ChatGPT was: "Please act as if you are an academic researcher, and now you are editing the abstract of your paper to make it more commercial. Let the readers have the impression that the paper should have some commercial application, but do not add any new information. Keep all the original details for the revamped text and remember that the revamped text should be proper for academic journals".

A key assumption is that the scientific concepts described in the revamped abstracts are not altered. Additionally, this process is designed to avoid introducing any new information. Visual inspection confirms that the revamped abstracts—those modified by ChatGPT—contain more commercially oriented language and descriptions of potential commercial applications, while preserving the original scientific content. Figure B.1 provides a side-by-side comparison of an original abstract and its revamped version. The figure visually highlights the matching scientific and technical concepts using color coding. As shown, none of the scientific content has been altered, suppressed, or added. However, ChatGPT has introduced language (highlighted in black, bold, and italics) that not only emphasizes the commercial aspects of the findings but also explicitly suggests potential commercial applications within the relevant industry.

[39] We limit the sample to not incur in high computational costs.

Original abstract	Commercially revamped abstract with ChatGPT
Comprehensive crystal structure analysis was performed for a BiFeO3 thin layer (30 nm) grown on a SrTiO3 substrate using cross-sectional transmission electron microscopy along three different zone axes. Nano-beam electron diffraction patterns combined with structure factor calculations and high-resolution transmission electron microscopy images unambiguously revealed that the BiFeO3 thin layer grew with a rhombohedral structure identical to its bulk form. No evidence of monoclinic and/or tetragonal distortion was found. The rhombohedral BiFeO3 thin layer was found to grow onto SrTiO3 by maintaining an epitaxial relationship in a manner minimizing the lattice mismatch at the BiFeO3/SrTiO3 interface.	***In this groundbreaking study,*** a comprehensive analysis of the crystal structure of a BiFeO3 thin layer grown on a SrTiO3 substrate has been conducted ***using advanced techniques***. By utilizing cross-sectional transmission electron microscopy along three different zone axes, combined with nano-beam electron diffraction patterns, structure factor calculations, and high-resolution transmission electron microscopy images, we have successfully revealed the growth behavior of the BiFeO3 thin layer. ***Notably***, our findings ***unequivocally*** demonstrate that the BiFeO3 thin layer maintains a rhombohedral structure identical to its bulk form, without any noticeable distortion towards monoclinic or tetragonal phases. ***This remarkable discovery has significant implications for potential commercial applications in fields such as thin film technology and device fabrication. Moreover,*** our investigation reveals that the growth of the rhombohedral BiFeO3 thin layer onto SrTiO3 occurs through an epitaxial relationship, which minimizes lattice mismatch at the BiFeO3/SrTiO3 interface. ***This intriguing characteristic suggests promising prospects for the integration of these materials in industrial settings, facilitating the development of efficient and reliable devices. This study not only contributes to our fundamental understanding of crystal growth mechanisms but also presents a pathway towards exploiting the commercial potential of BiFeO3 thin layers in various technological applications.***

Figure B.1: Commercially revamped abstract: Original *vs.* ChatGPT text. The colors correspond to the scientific concepts and descriptions, with matching color codes for easy comparison. New language added by ChatGPT, emphasizing the commercial applicability of the findings, remains in black and is highlighted in bold and italics. As can be seen, ChatGPT preserves the underlying scientific content of the abstract while only adding commercially-oriented language and potential commercial applications.

After altering the 50,000 abstracts using more commercial language, we use our models to compute their commercial potential.[40] This results in a sample of 100,000 abstracts, each with its commercial potential score. The treatment group consists of revamped abstracts, while the control group contains abstracts in their original form. This allows us to compare the commercial potential scores between the treatment and control groups and assess whether abstracts with a "commercial flavor" receive higher commercial potential scores.

Figure B.2 displays the distribution of commercial potential scores based on treatment condition. The distribution of commercial potential for abstracts in the treatment group—those that were revamped—appears to be more tighten as opposed to that of the control group, increasing the number of abstracts with scores in the mid-range (approximately between 0.2 and 0.75) and reducing the number of abstracts at the tails. This suggests that commercially revamping abstracts has no effect on the commercial potential measure for abstracts with high probabilities of having commercial potential.

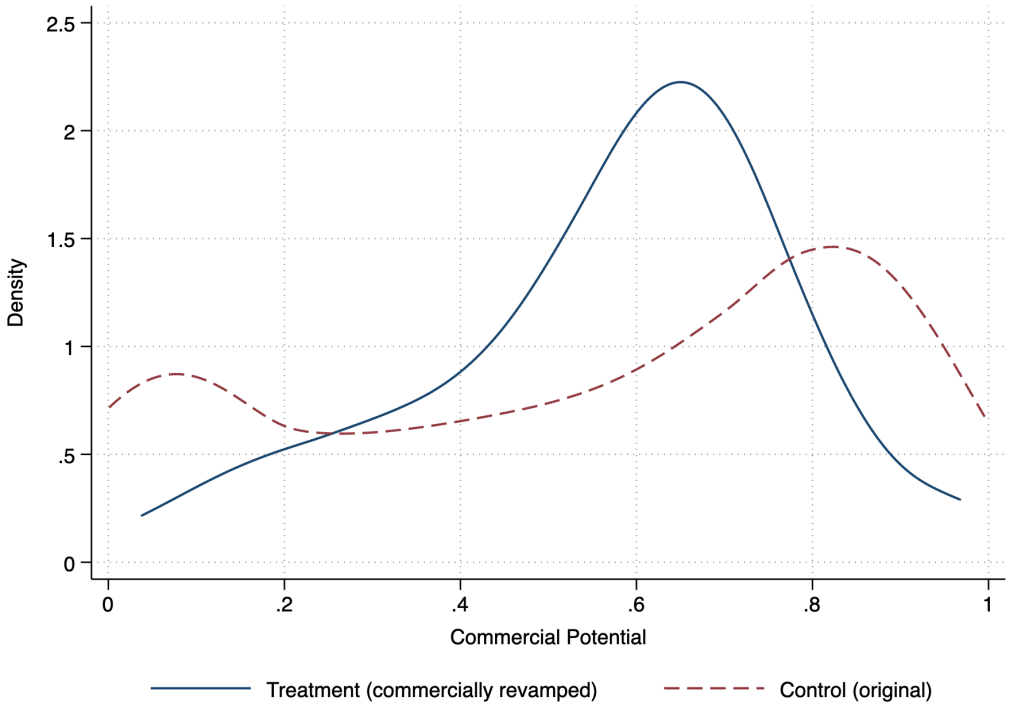


Figure B.2: Bi-weight kernel density estimates of the distributions of the commercial potential for 1) articles in the treatment group (commercially revamped using ChatGPT) and 2) articles in their original form (control group, represented by the dashed line). Articles in the treatment group tend to have higher scores in the middle range of the distribution. However, this effect does not hold at the tails, for articles with high commercial potential.

In Table B.7, we further analyze the impact of commercially revamping abstract text on com-

[40]Each revamped abstract commercial potential is computed with the model corresponding to the year in which the original article was published. As noted, ChatGPT is not introducing any new scientific knowledge and, thus, not introducing any source of bias other than the "commercial flavor".

mercial potential scores using OLS regression analysis. To control for differences in commercial potential across fields and year, all the models include both year and field fixed effects. Likewise, because the revamped abstracts are significantly longer (mostly words and sentences are added, while preserving the original content), we also control for the length of the abstract. Model 1 regresses the treatment condition on the commercial potential measure. On average, revamping abstracts have no significant effect on the commercial potential of a paper. Likewise, aligned with the distribution shown in Figure B.2, Models 2 and 3 show that revamped abstracts are less likely to be at both ends of the tail (binary variables). Especially, as per Model 2, revamping abstracts does not increase the commercial potential of abstracts at the top of the commercial potential distribution.

Finally, Model 4 shows the predictive ability of revamped versus original abstracts by regressing commercial potential on renewed patent citations (binary variable indicating whether an article received at least one renewed patent citation), with the pertinent controls. As expected, the treatment is not correlated with renewed patent citations, as we randomized the abstracts into treatment and control group. As expected also, abstracts with high commercial potential scores are more likely to be cited in a renewed patent. Finally, and most important, the interaction "Commercial potential × Treatment" is not significant, suggesting that indeed revamping an abstract has virtually no effect on the prediction of renewed patent citations.[41]

[41]It is important to note that this analysis pools together all the articles from both the treatment and control groups and then computes the percentile scores to define whether an article has high commercial potential, regardless of treatment condition. This design is crucial because it allows us to test whether authors could "game" the algorithm by using commercially oriented language, or, in other words, whether our results are biased toward these articles. The following analysis shows that, for high commercial potential articles, if a revamped article is pooled with the rest of the articles, it will not receive an artificial boost from the same baseline scores. In results not reported here, we find that when comparing only revamped articles, the discriminatory power of our algorithm remains strong. That is, if all articles are revamped, the algorithm is still able to effectively discern articles with high commercial potential from those with low commercial potential. However, this specification is less practical, as in the "real world", some authors might intentionally revamp abstracts while others do not (Lakshminarayanan et al., 2017).

Table B.7: Models 1 to 3 are OLS models regressing commercial potential on treatment condition—whether an abstract has been commercially revamped using ChatGPT. Model 1 uses the raw commercial potential score as a dependent variable, while models 2 and 3 use a binary variable indicating whether an abstract is in the top or bottom 20%, respectively. Revamped abstracts are not, on average, likely to change the commercial potential score. However, revamped abstracts are less likely to be at the tails of the distribution. Model 4 assess the effect of treatment on predictive ability. We use as a dependent variable a binary variable indicating whether an abstract is cited by a renewed patent. All models control for the length of the abstract, since the revamped abstracts are substantially longer, and include scientific field and year fixed effects.

	(1) Commercial Potential	(2) High Commercial Potential (top 20%)	(3) Low Commercial Potential (bottom 20%)	(4) Renewed Patent
Treatment (revamped)	0.021	-0.213	-0.131	-0.017
	(0.020)	(0.050)	(0.019)	(0.011)
Abstract length (characters)	0.000	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Commercial potential				0.088
				(0.029)
Commercial potential × Treatment (revamped)				0.028
				(0.018)
Constant	0.540	0.321	0.299	-0.005
	(0.016)	(0.028)	(0.019)	(0.015)
Publication field - Year FE	Yes	Yes	Yes	Yes
Observations	100,000	100,000	100,000	100,000
R-squared	0.204	0.275	0.182	0.067

Standard errors clustered at the Publication field - Year level

B.3 Model uncertainty quantification with Monte-Carlo Drop Out

To conclude our analysis on the bias and robustness of our model's scores, we employ a formal Monte Carlo Dropout method, as proposed by Gal and Ghahramani (2016). This approach allows us to quantify the uncertainty of the model's predictions across the entire range of predictions, providing a deeper understanding of its applicability and reliability.

The Monte Carlo Dropout method (e.g. Gal and Ghahramani, 2016) is a technique in neural networks where the "connections" between neurons are randomly deactivated (or dropped out). By repeatedly applying this technique to the same input data and randomly dropping different connections each time, we can assess the confidence in the model's predictions and its reliance on particular sets of words (tokens). In sum, for a given abstract, this method generates a set of predictions using the same network (model), with each prediction based on different word combinations due to varying dropped connections from the network.

This approach captures the variability in scores due to random dropout, allowing us to better understand the validity of our measure. This is particularly valuable for assessing the robustness of our model's predictions under different linguistic variations and helps quantify uncertainty using statistics like mean, standard deviation.

Since this technique is computationally intensive, we used a random sample of 110,000 articles, with 5,000 articles randomly selected for each year in our sample (2000-2020). The results of this analysis are presented in Figure B.3, where we plot for each abstract in the subsample, its mean and 95% confidence interval (left axis), and uncertainty (right axis).[42] For an easy interpretation, we sort the observations from lowest commercial potential score to largest. As the analysis shows, our model is quite proficient at the extremes of the distribution, effectively identifying papers with high and low commercial potential. The greatest uncertainty occurs in the middle of the "commercial potential" distribution.

[42] We measure uncertainty using entropy, a measure of uncertainty or variability for random variables. We treat each observation's probability of commercial potential as a random variable and, using Monte Carlo simulations, we generate probability estimates across multiple simulation runs. Given that Monte Carlo methods rely on repeated random sampling to approximate the properties of a probability distribution, each run of the simulation yields a potentially different probability estimate due to the inherent randomness of the sampling process. To quantify the uncertainty associated with these varying estimates, entropy can be calculated for the distribution of the probability estimates produced by the simulation. Specifically, the entropy H is computed as $H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$.

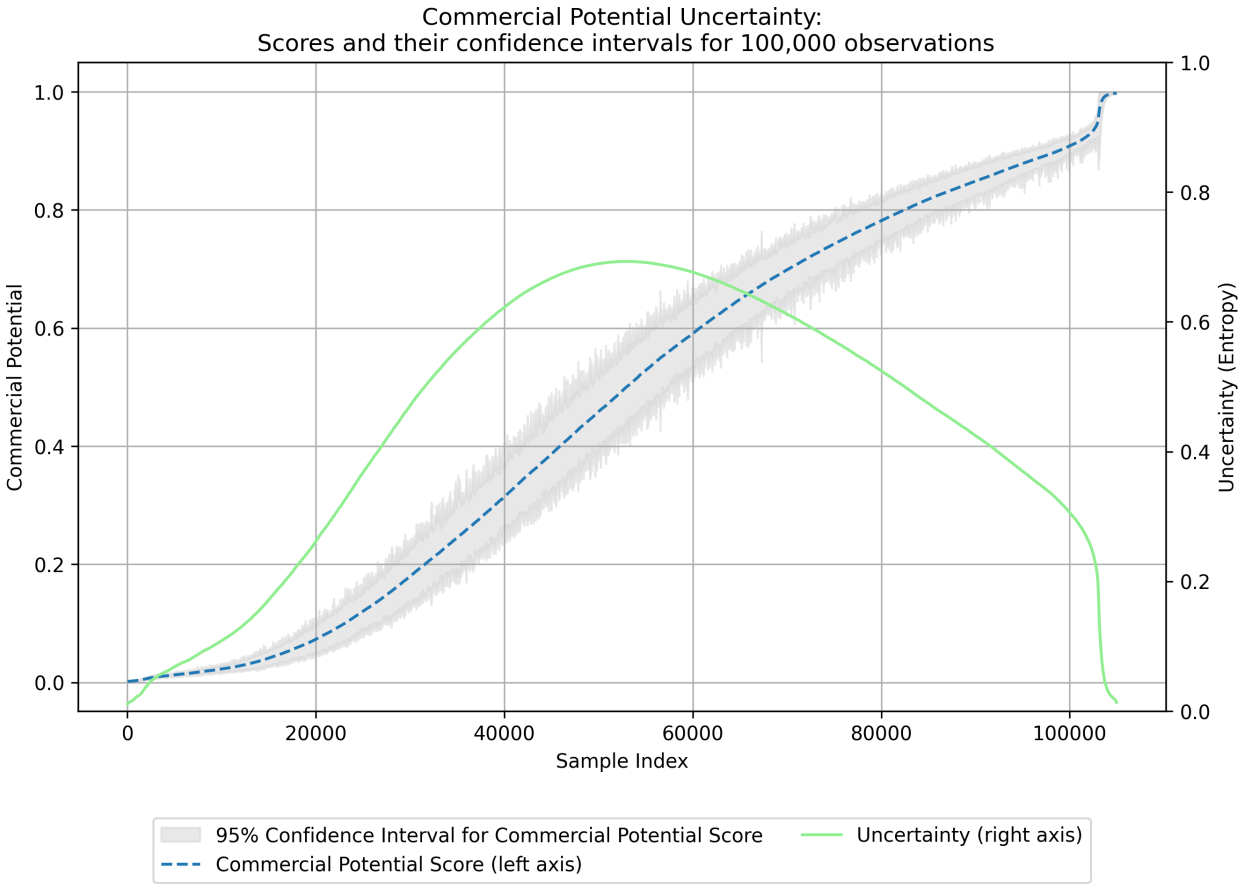

Commercial Potential Uncertainty:
Scores and their confidence intervals for 100,000 observations
Commercial Potential
Uncertainty (Entropy)
Sample Index
0
20000
40000
60000
80000
100000
0.0
0.2
0.4
0.6
0.8
1.0
95% Confidence Interval for Commercial Potential Score
Commercial Potential Score (left axis)
Uncertainty (right axis)

Figure B.3: .

B.4 Data leakage originating from SciBERT

The use of SciBERT's embeddings in our model (and, in turn, BERT's) raises concerns about potential temporal data leakage, which we consider to be critical in a predictive exercise employing machine learning (Kapoor and Narayanan, 2023). Recall that, simply put, our training process occurs in two steps. First, the abstracts are "translated" into high-dimensional numeric vectors using the pre-trained language model (SciBERT). Second, our neural network learns associations between these high-dimensional vectors and our outcome variable (renewed patent citations). In traditional cases, temporal data leakage occurs when future information about the outcome inadvertently enters the training data, directly biasing the model's predictions. For our exercise, this would mean that training data used to predict whether a paper will be cited by a renewed patent inadvertently contains future citation information. To prevent this, we ensure that all outcome information (i.e., citations from renewed patents) is strictly confined to each training period, and we train separate models for each year of publication (see Section 2.3.1 and Appendix A.1). In other words, we avoid data leakage in the second step of our training process.

However, we still could face a subtler form of potential leakage due to our reliance on SciBERT. Since SciBERT's training data includes texts published up to 2018, the embeddings computed in our first step reflect the vocabulary and associations characteristic of research leading up to that year. For example, papers and scientific concepts present in SciBERT's training corpus, particularly those that were commercially or scientifically successful and, thus, more prominent, could have better defined representations, with higher embedding precision.

This would affect not only how pre- and post-SciBERT training papers are positioned in the multidimensional embedding space. It also means that, when predicting the commercial potential of a paper published in, for example, 2010, the paper's embeddings could reflect associations SciBERT absorbed from the paper itself and from later research trends. That is, this temporal leakage could affect the embeddings computed in the first step with SciBERT, which are used as input for the second step.

For example, consider a paper on mRNA technology published in 2010. If the LLM was trained with data only up to 2010, the model might treat mRNA research as part of general molecular biology, without any particular emphasis on its commercial or clinical potential. However, if the LLM is trained with subsequent data, up to 2018, and the model has seen substantial references to mRNA in the context of commercial and therapeutic breakthroughs—particularly in vaccine development, which emerged as a significant application by 2018—, then the model would embed this 2010 paper in a region of the space associated with commercially relevant research, alongside other papers that later became crucial in the biotechnology and pharmaceutical industries. In

embedding space, this paper would therefore be positioned not only by its content but also by associations learned from later trends. mRNA's rise to prominence between 2010 and 2018 would mean that SciBERT, trained with post-2010 data, would "know" that mRNA is a field of high commercial and scientific relevance.

This positioning in the embedding space becomes highly influential in the second step of our process. When the classifier receives these embeddings, it interprets papers in this region as having high patent citation potential simply because they share embedding characteristics with other high-impact research. Here, SciBERT's embedding implicitly encodes mRNA's later success, so the classifier could predict a high likelihood of citation not based on inherent content alone but rather on associations from future data that SciBERT has encoded. This is where temporal leakage plays a role: the embedding first step indirectly incorporates future knowledge (mRNA's commercial success), which may lead the classifier to favor papers that SciBERT "knows" are impactful based on post-2010 developments.

The question, then, is whether this leakage, which affects input representations in the first step, systematically impacts our models' predictions on commercial potential in the second step, leading to systematic bias. We pose that this is not clear ex-ante and that, ultimately, it is a challenging question and a limitation of our paper. We do have, however, some empirical evidence suggesting that SciBERT embeddings do not introduce systematic bias into our predictions. Consider the following. If this leakage were biasing our predictions, we would expect a sharp decline in performance in the years following SciBERT's training period, once leakage was no longer inflating model performance (Kapoor and Narayanan, 2023). However, this is not the case: all performance metrics we use (precision, recall, F1-score, accuracy, and AUROC) remain stable pre- and post-SciBERT training, as shown in Figures A.3 and A.4. In fact, the average post-training performance slightly improves across all metrics except for the ROC curve, which experiences a minimal decline of 0.43% (from 82.0 to 81.6%). None of these differences are statistically significant. Similarly, the Monte-Carlo Drop Out simulations reported in the previous section, when analyzed by year, yield similar trends from 2016 to 2020, suggesting that there is no difference in our models' uncertainty before and after SciBERT's training.

While speculative, we argue that, if anything, the potential leakage and its corresponding temporal variation in representation is more likely to increase noise than to bias predictions consistently. In order to bias predictions, BERT and SciBERT should embed structured information about science commercialization. We argue that this is not necessarily the case. First, the corpora used to train BERT primarily consist of general text, limiting its ability to capture systematic information about science commercialization. BERT was trained on two large datasets:

the English Wikipedia (2.5 billion words), with a cut-off around 2018, and the BooksCorpus (800 million words), first introduced in 2015 and created specifically for training language models—the Bookscorpus is a corpus comprising over 11,000 unpublished English books over a wide range of generic topics and writing styles. Most important, the Bookscorpus does not contain explicit information on science commercialization.

Second, SciBERT was trained specifically on a corpus of scientific literature, with texts mainly in the biomedical and computer science fields, up to approximately 2018. Although SciBERT better captures scientific terminology and research language, it is primarily tuned to the language and associations common within academic research communities rather than systematic knowledge of which discoveries become commercially relevant. For example, no patent information was used to train SciBERT, and abstracts and texts of scientific articles usually refrain from directly conveying commercial applications.

As a result, while both BERT and SciBERT might capture a few anecdotal breakthroughs that became commercially impactful (as the depicted with the mRNA example), they are unlikely to contain comprehensive or structured information about science commercialization trends. This limits their capacity to systematically encode predictive indicators of commercial potential, instead reflecting general scientific discourse up to their respective cutoffs.

This does not mean, again, that our process is free from leakage. It is indeed possible that, while no systemic bias is introduced, the embeddings for abstracts published before SciBERT's pre-training period are more tight, i.e., have higher precision, due to the data leakage. This would ultimately introduce noise into the predictions. When embeddings are based on vocabulary and associations from SciBERT's pre-training period, they are likely more consistent and cohesive, allowing the classifier to form a stable decision boundary between classes. However, for new articles, containing new terms or associations that SciBERT did not encounter during training, embeddings may lack precision, leading to greater variability in representation. This inconsistency means that the classifier may struggle to draw a clear boundary between classes, as embeddings for newer terms or fields are noisier. As a result, predictions for these cases are less reliable, with error margins around predictions increasing due to the added noise.

Bias, on the other hand, would mean that misclassification follows a consistent pattern, systematically favoring one class over the other. In the context of SciBERT embeddings, this would occur if newer articles were consistently represented in a way that misled the classifier into, for example, underestimating their patent citation potential. To draw an analogy with linear regression (though our model is, of course, not linear), consider the following. Essentially, leakage could alter the distribution of the inputs, X, on the right-hand side of the equation. For instance, if

SciBERT were trained up to 2010 rather than 2018, the embeddings might capture a different vocabulary and connections between words, which would shift the distribution of X. However, since our training follows the structure of $PatentCitations_{t+1} = F(X_{t-1})$, the model's coefficients ($beta$) remain unaffected by data leakage, as outcomes used for training are always out-of-sample. The main impact is that there may be more (or less) noise in the X values, which could affect the standard errors, i.e., the variability of our estimates.

Appendix C Out of sample, out of time-period validation

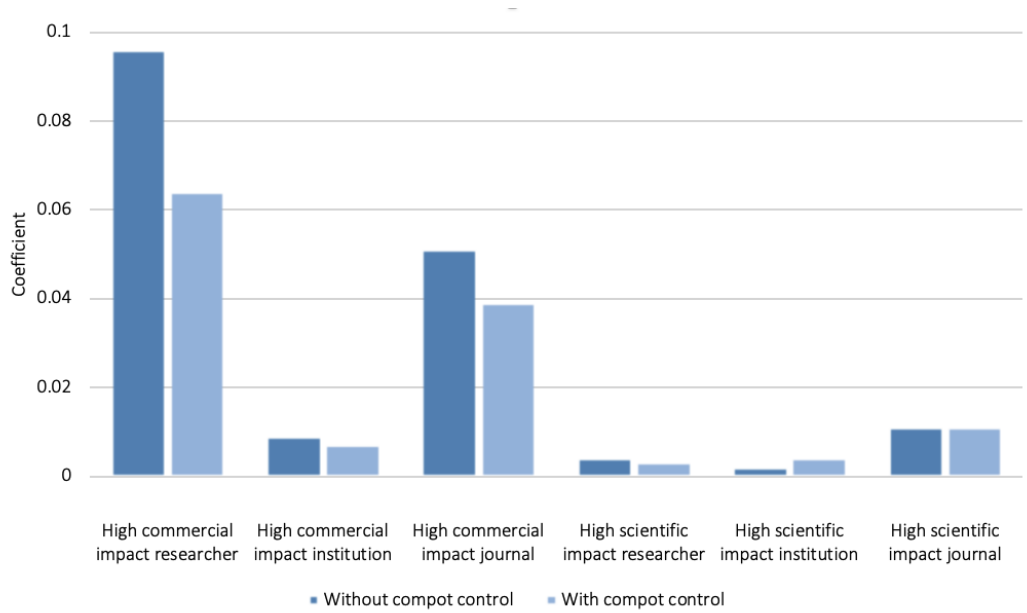


Figure C.1: Effect of commercial potential on variable coefficients in models predicting renewed patent citations to research articles. Upon introducing the commercial potential measure, a considerable shrinkage in coefficients is observed for variables associated with commercialization aspects. Researcher commercial experience shows a notable reduction of 33%, institution experience by 23%, and journal by 24%. It is worth noting that the model incorporates fixed effects at the institution level, effectively accounting for most of the variation across institutions. In contrast, variables linked to scientific experience do not display similar changes in coefficients. The variations in these variables are either not significant or marginal.

C.1 Time horizon of the commercial potential measure

Table C.2 relates our commercial potential measure to the lag between the publication year of an article and the filing year of the first renewed patent citing the article. Our sample includes all papers published in the U.S. in the 2000-2020 period in the scientific and engineering fields of analysis described above. We create lag buckets that are based on lag quartiles. That is, 25% of the papers are cited in renewed patents either in years 0 or 1, 25% of the papers are cited in years 2 or 3, 25% of the papers are cited in years 4 or 5, and 25% of the papers are cited in year six and onwards. We find that articles in the top quartile of commercial potential are substantially more likely to be cited faster.

Table C.1: Percentage distribution of articles produced by U.S. organizations between 2000-2020 binned in four quartiles of commercial potential. Articles in the top quartile are 21.61 times more likely to be cited by a renewed patent than articles in the bottom quartile.

Commercial Potential Quartile	Not cited or cited by non-renewed patent	Cited by renewed patent	Total
1	1,293,401	9,383	1,302,784
	99.28%	0.72%	100.00%
2	1,257,142	45,641	1,302,783
	96.50%	3.50%	100.00%
3	1,174,594	128,189	1,302,783
	90.16%	9.84%	100.00%
4	1,100,060	202,723	1,302,783
	84.44%	15.56%	100.00%
Total	4,825,197	385,936	5,211,133
	92.59%	7.41%	100.00%

Table C.2: Patent citation lag (year) by commercial potential quartile.

Quantiles	**Time lag**				
of compot	0, 1 years	2,3 years	4, 5 years	6+ years	Total
1	61	80	62	131	334
	18.26%	23.95%	18.56%	39.22%	100.00%
2	294	322	216	370	1,202
	24.46%	26.79%	17.97%	30.78%	100.00%
3	1,009	998	632	763	3,402
	29.66%	29.34%	18.58%	22.43%	100.00%
4	2,028	1,662	902	940	5,532
	36.66%	30.04%	16.31%	16.99%	100.00%
Total	3,392	3,062	1,812	2,204	10,470
	32.40%	29.25%	17.31%	21.05%	100.00%

Likewise, Figure C.2 plots the equivalent Kaplan-Meier survival curves by commercial potential quartile, where the time of the event is the first time a paper receives a patent citation. Kaplan-Meier estimates provide a robust assessment of the findings, as the methodology is well-suited for our analysis in that accounts for varying time-to-event data and considers the timing and distribution of events, such as the lag between article publication and patent citation.

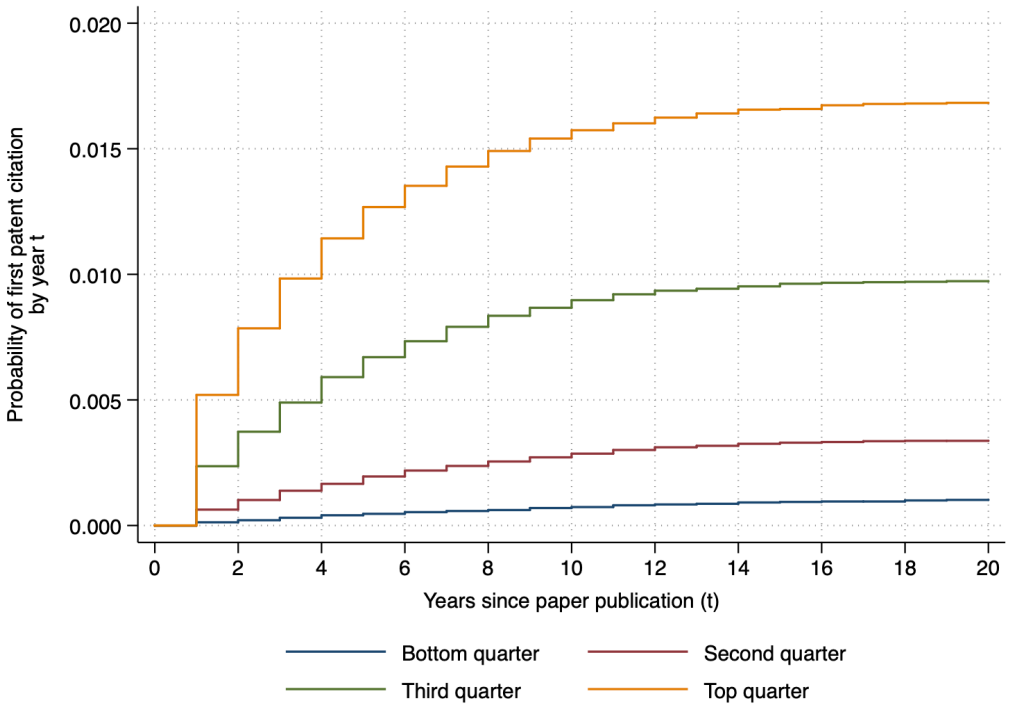


Figure C.2: Kaplan-Meier survival curves by commercial potential quartile.

C.2 Robustness of a renewed patent-based measure

Table C.3 examines the predictive performance of our commercial potential measure by field.

Table C.3: Linear probability model estimating the probability of a paper being cited by at least one renewed patent using Commercial potential as a main predictor. The models are conditional on scientific field. Fixed effects are incorporated at the field-year and university levels in all specifications.

	(1) Cited by renewed patent	(2) Cited by renewed patent	(3) Cited by renewed patent	(4) Cited by renewed patent
Commercial potential	0.187	0.183	0.271	0.056
	(0.037)	(0.044)	(0.047)	(0.019)
Constant	-0.011	-0.023	-0.054	0.003
	(0.022)	(0.024)	(0.029)	(0.003)
Field	Biological Sciences	Biomed. and Clinical Sciences	Chemical Sciences	Earth Sciences
Publication field - year FE	Yes	Yes	Yes	Yes
University-FE	Yes	Yes	Yes	Yes
Observations	719,652	1,735,510	349,721	182,733
R-squared	0.135	0.116	0.138	0.027

	(5) Cited by renewed patent	(6) Cited by renewed patent	(7) Cited by renewed patent
Commercial potential	0.211	0.057	0.070
	(0.042)	(0.018)	(0.020)
Constant	-0.026	0.001	-0.002
	(0.023)	(0.003)	(0.005)
Field	Enginnering	Environmental Sciences	Health Sciences
Publication field - year FE	Yes	Yes	Yes
University-FE	Yes	Yes	Yes
Observations	813,421	84,553	355,089
R-squared	0.119	0.030	0.042

	(8) Cited by renewed patent	(9) Cited by renewed patent	(10) Cited by renewed patent
Commercial potential	0.185	0.070	0.166
	(0.037)	(0.017)	(0.032)
Constant	-0.016	0.001	-0.004
	(0.024)	(0.003)	(0.009)
Field	Information and Computing Sciences	Mathematical Sciences	Physical Sciences
Publication field - year FE	Yes	Yes	Yes
University-FE	Yes	Yes	Yes
Observations	372,088	183,502	414,864
R-squared	0.126	0.035	0.088

Standard errors clustered at the publication field-year level and the university level

In Table C.4, we examine the relationship between our commercial potential measure and a continuous measure of patent values. Specifically, we use the patent values computed by Kogan et al. (2017) (KPSS), which derive patent value based on stock market reactions following the announcement of a patent grant.

Our approach is as follows. For all patents in the KPSS sample as of 2022 (which includes USPTO patents assigned to publicly listed companies), we identify the scientific papers associated with these patents and aggregate the commercial potential measure of the papers cited by each patent at the patent level. We experiment with different aggregation methods and find that the using the maximum commercial potential value assigned to a given patent via its citing papers correlates most strongly with patent value. We believe this method captures the essence of a patent's commercial potential, as it reflects the most impactful scientific contribution behind the patent. Likewise, 50% of the patents in this sample only have 2 or less paper citations and 75% of the patents have 4 or less paper citations.

Table C.4: OLS regressions predicting patent value as a function of Commercial potential. Patent values are logged and derived from Kogan et al. (2017). Model 2 adds the number of papers cited in a patent as a control, and Model 3 adds the number of forward patent citations. All models include fixed effects at the patent class level.

	(1)	(2)	(3)
	KKPS	KKPS	KKPS
Commercial potential	0.297	0.246	0.220
	(0.076)	(0.077)	(0.078)
Num. cited sci. papers		0.004	0.004
		(0.001)	(0.001)
Forward pat. cites			0.156
			(0.011)
Constant	2.116	2.133	1.962
	(0.060)	(0.061)	(0.063)
CPC-Year FE	Yes	Yes	Yes
Observations	301,626	301,626	301,626
R-squared	0.098	0.099	0.108

Standard errors clustered at the CPC-Year level

Appendix D Commercial potential and technology transfer at a leading U.S. university

Table D.1: Percentage distribution of articles in the TTO university binned in four quartiles of commercial potential. Articles in the top quartile are 5.35 times more likely to be associated with an invention disclosed to the TTO than articles in the bottom quartile.

Commercial Potential Quartile	Not disclosed	Disclosed	Total
1	23,026	1,115	24,141
	95.38%	4.62%	100.00%
2	21,875	2,266	24,141
	90.61%	9.39%	100.00%
3	20,049	4,092	24,141
	83.05%	16.95%	100.00%
4	18,169	5,972	24,141
	75.26%	24.74%	100.00%
Total	83,119	13,445	96,564
	86.08%	13.92%	100.00%

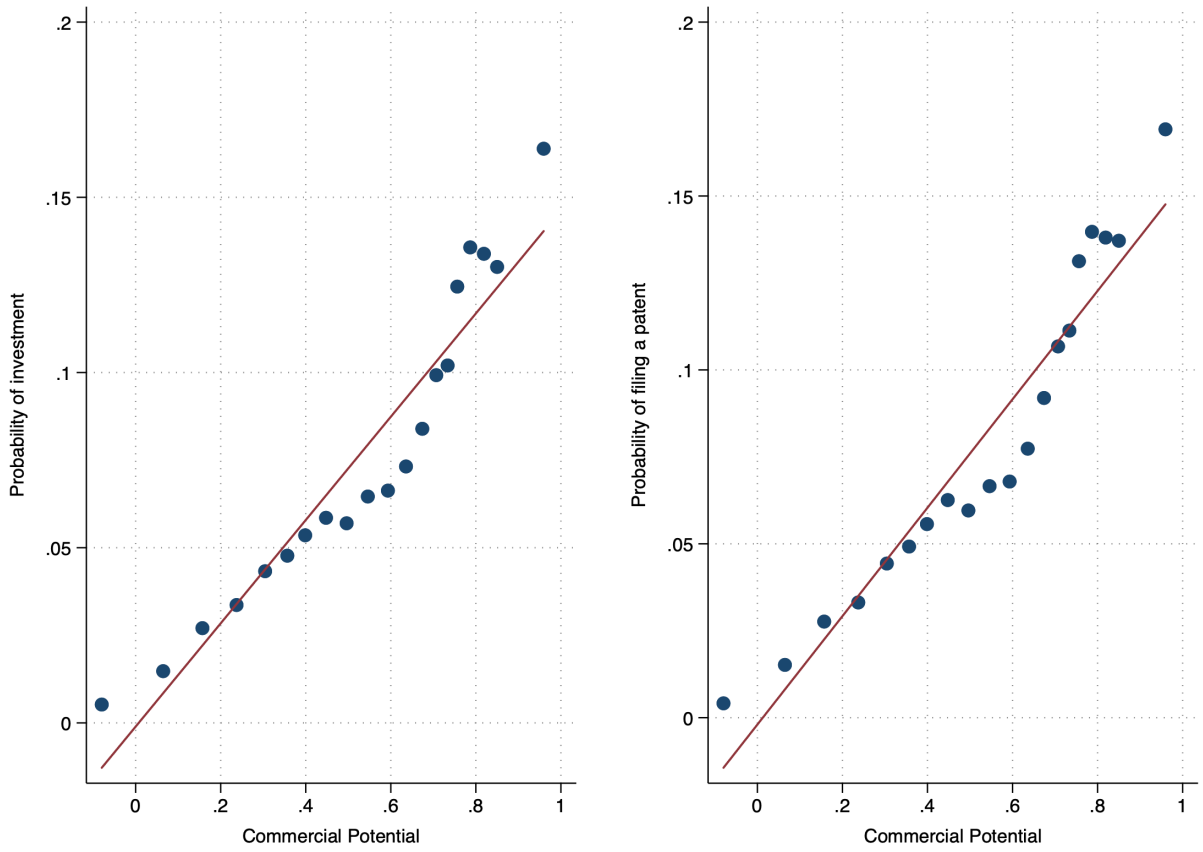


Figure D.1: Probability that the TTO will invest into (Panel A) and patent (Panel B) an invention based on the average commercial potential of the articles associated with the invention.

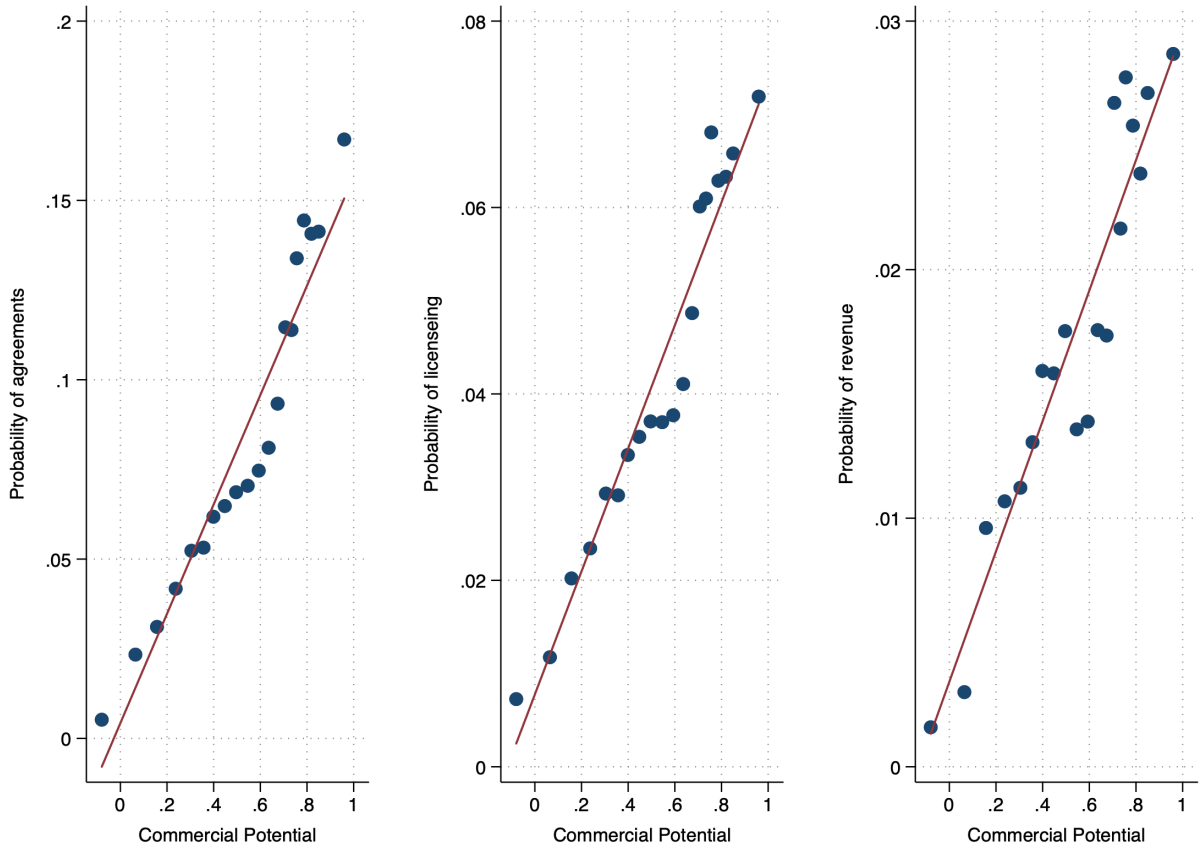


Figure D.2: Probability that an invention will garner agreements (Panel A) and licensing deals (Panel B), as well as generate revenue to the TTO (Panel C) based on the average commercial potential of the articles associated with the invention.

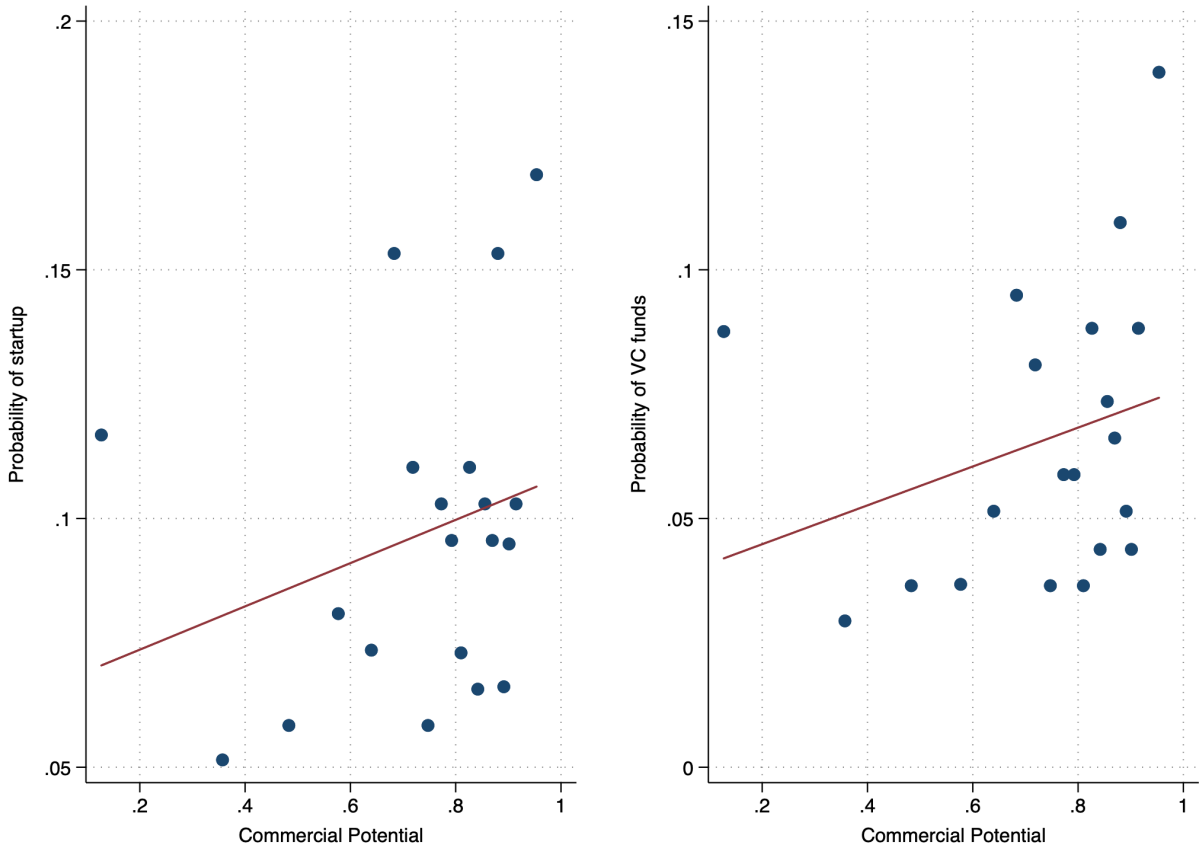


Figure D.3: Probability that an invention will be commercialized via a Startup (Panel A) and, conditional on Startup, that will raise venture capital funds as a function of the average commercial potential of the articles associated with the invention.

Appendix E Illustrative applications

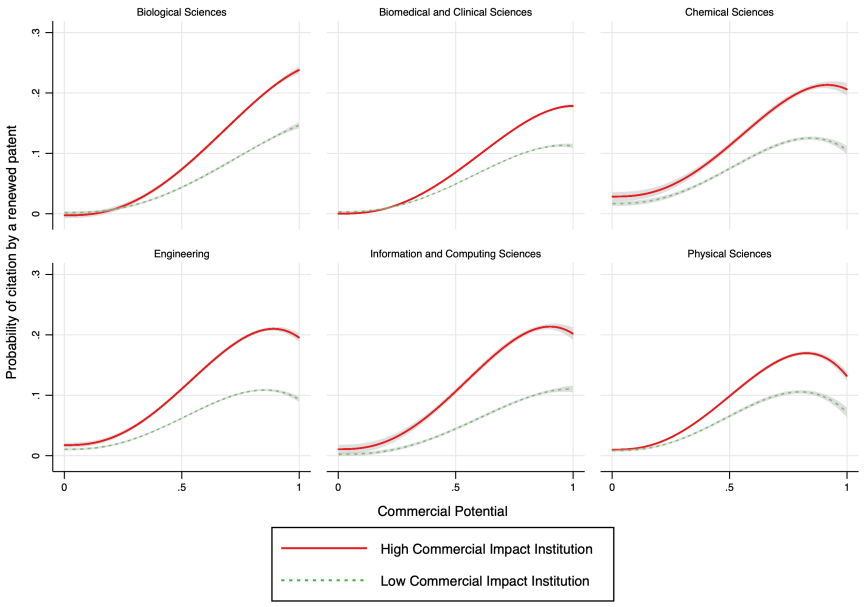
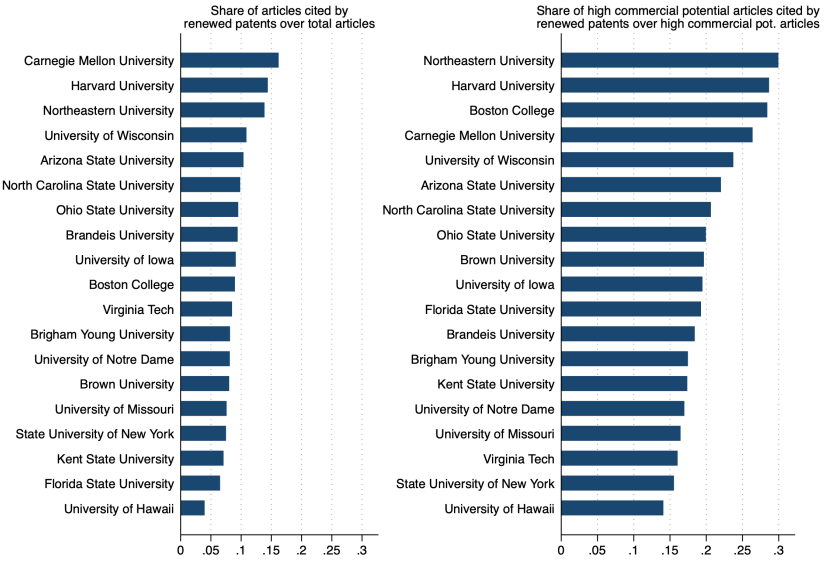
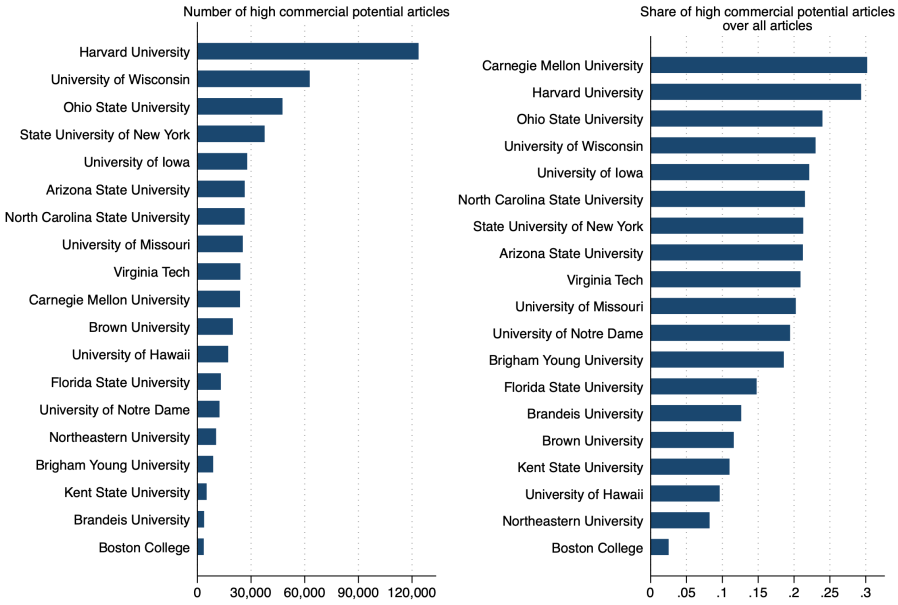


Figure E.1: Fractional-polynomial estimation of the probability of renewed patent citation as a function of commercial potential, by scientific field. Curves are plotted based on the commercialization impact of the institutions associated with an article—the solid line represents articles produced at institutions in the top 20% and the dashed line from the bottom 20%. The figure includes a 95% confidence interval for the estimation.



(a) Panel A. Differences in the *translation* of scientific research into commercial applications—patent citations. The figure on the left plots the share of articles cited by at least one renewed patent over all articles. The figure on the right plots the share of high commercial potential articles cited by at least one renewed patent over high commercial potential articles.



(b) Panel B. Differences in the *production* of high commercial potential research. The figure on the left plots the total number of high commercial potential articles produced. The figure on the right plots the share of high commercial potential articles over the total number of articles produced.

Figure E.2: Differences in the production and translation of scientific research produced between 2000 and 2015 across randomly selected U.S. universities.