NBER WORKING PAPER SERIES

PATENT TEXT AND LONG-RUN INNOVATION DYNAMICS: THE CRITICAL ROLE OF MODEL SELECTION

Ina Ganguli
Jeffrey Lin
Vitaly Meursault
Nicholas F. Reynolds

Working Paper 32934
http://www.nber.org/papers/w32934

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2024

The views expressed in this paper are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia, the Federal Reserve System, or the National Bureau of Economic Research. No statements here should be treated as legal advice. Any errors or omissions are the responsibility of the authors. Acknowledgements: We gratefully acknowledge support from an NBER Innovation Policy Grant. We also received excellent RA support from Aaron Rosenbaum, Joseph Huang, Cameron Fen, Annette Gailliot, Jake Moore, and Isaac Rand. Finally, we received useful feedback from Matt Clancy, Darya Davydova, Gaétan de Rassenfosse, Luise Eisfeld, Deanna James, Semyon Malamud, Roxana Mihet, participants of the seminar at EPFL, and participants of the NBER Innovation Information Initiative Technical Working Group Meeting and TADA 2023. First version: December 21, 2023.

Patent Text and Long-Run Innovation Dynamics: The Critical Role of Model Selection
Ina Ganguli, Jeffrey Lin, Vitaly Meursault, and Nicholas F. Reynolds
NBER Working Paper No. 32934
September 2024
JEL No. C81, L19, O31

ABSTRACT

As distorted maps may mislead, Natural Language Processing (NLP) models may misrepresent. How do we know which NLP model to trust? We provide comprehensive guidance for selecting and applying NLP representations of patent text. We develop novel validation tasks to evaluate several leading NLP models. These tasks assess how well candidate models align with both expert and non-expert judgments of patent similarity. State-of-the-art language models significantly outperform traditional approaches such as TF-IDF. Using our validated representations, we measure a secular decline in contemporaneous patent similarity: inventors are “spreading out” over an expanding knowledge frontier. This finding is corroborated by declining rates of multiple invention from newly-digitized historical patent interference records. In contrast, selecting another single representation without validating alternatives yields an ambiguous or even opposing trend. Thus, our framework addresses a fundamental challenge of selecting among different black-box NLP models that produce varying economic measurements. To facilitate future research, we plan to provide our validation task data and embeddings for all US patents from 1836–2023.

Ina Ganguli
Department of Economics
Crotty Hall 304
412 N. Pleasant Street
University of Massachusetts Amherst
Amherst, MA 01002
and NBER
iganguli@econs.umass.edu

Jeffrey Lin
Federal Reserve Bank of Philadelphia
Research Department
Ten Independence Mall
Philadelphia, PA 19106
jeffr.lin@gmail.com

Vitaly Meursault
Federal Reserve Bank of Philadelphia
vitaly.meursault@gmail.com

Nicholas F. Reynolds
University of Essex
Wivenhoe Park
Colchester
United Kingdom
nicholas.reynolds@essex.ac.uk

1. Introduction

Measuring invention similarity is important for many decisions. Patent examiners must assess the similarity of new claims against prior art to determine novelty. Inventors, particularly those in competitive fields, need to understand how their ideas overlap with those of rivals. Policymakers may balance encouraging specific innovations versus minimizing redundant efforts. Each of these contexts demands a reliable measure of invention similarity.

Beyond these practical applications, invention similarity has long been a core concern for economists. Similarity influences patent value, the intensity of knowledge spillovers, the direction of technological change, and the efficiency of R&D investments (Griliches 1979; Jaffe 1986). Accurate similarity measurement is also fundamental to deriving other economically significant metrics. For example, Kelly et al. (2021) identify "breakthrough" patents as those dissimilar to prior art but highly similar to subsequent patents.

Many methods have been used to measure patent similarity. How should researchers choose among them? And how should readers evaluate these choices? Our main contribution is to develop and implement a pipeline for the construction, validation, and selection of measures of economic interest derived from patent text. This pipeline also serves as a step-by-step guide for researchers constructing other measures and for readers assessing the reliability of these measures. Our pipeline emphasizes domain-specific validation and model selection as essential steps. Crucially, we demonstrate that the choice of representation is not innocuous and can dramatically affect economic measures of interest.

The construction of patent similarity and other measures can be usefully separated into three distinct steps: (1) representation, (2) measurement, and (3) validation-based selection. The first step maps each patent to a location in idea space, representing it as a vector in $\mathbb{R}^n$. This mapping could be based on patent office classifications, traditional Natural Language Processing (NLP) methods that count words, or more modern NLP methods that produce distributed embeddings, where meaning is "distributed" over a whole vector. The second step measures a concept of economic interest using representations produced by each of several candidate models. Patent similarity is a classic quantity of interest; other concepts might be motivated from theory, derived from a structural model, or based on intuition.

Different representations lead to different measurements of the same concept. Therefore, the third step involves validating these representations using purpose-built, domain-specific tasks to select the optimal mapping among an ever-expanding list of NLP alternatives. This crucial step, often overlooked in economics, is a central focus of our paper. We emphasize that validating a single method without considering alternatives can lead to spurious conclusions. Instead, our approach aligns with the NLP literature's view that no method is universally superior, and each should be evaluated based on its task-specific performance (Ash and Hansen 2023). As Grimmer, Roberts, and Stewart (2022) note, "the best method depends on the task" and "validations are essential and depend on the theory and the task."

To illustrate the importance of model selection, we analyze the full text of US utility patent claims from 1836 to 2023 using different representations. Figure 1 shows the average pairwise similarity of patents by year, based on two different representations (see details in Section 4). According to the General Text Embedding (GTE) model, a top performer in our validation tasks, patent claim similarity has declined steadily over a century and a half. In contrast, patent claim similarity measured by the widely-used Term Frequency-Inverse Document Frequency (TF-IDF) representation increased steadily between 1850 and 1950, and has levelled off since. The divergence between GTE and TF-IDF measures of patent similarity underscores the critical role of representation choice in economic analysis of patents and technological change.

To implement our validation-based selection pipeline, we design novel, domain-specific validation tasks that compare the performance of seven leading, widely-used NLP models. Our validation tasks use (a) patent interference cases, (b) non-expert human annotations, and (c) patent office technology classifications to assess model performance. We evaluate (i) Term Frequency-Inverse Document Frequency (TF-IDF), a traditional workhorse model that counts words, and six modern models that produce distributed embeddings: (ii)

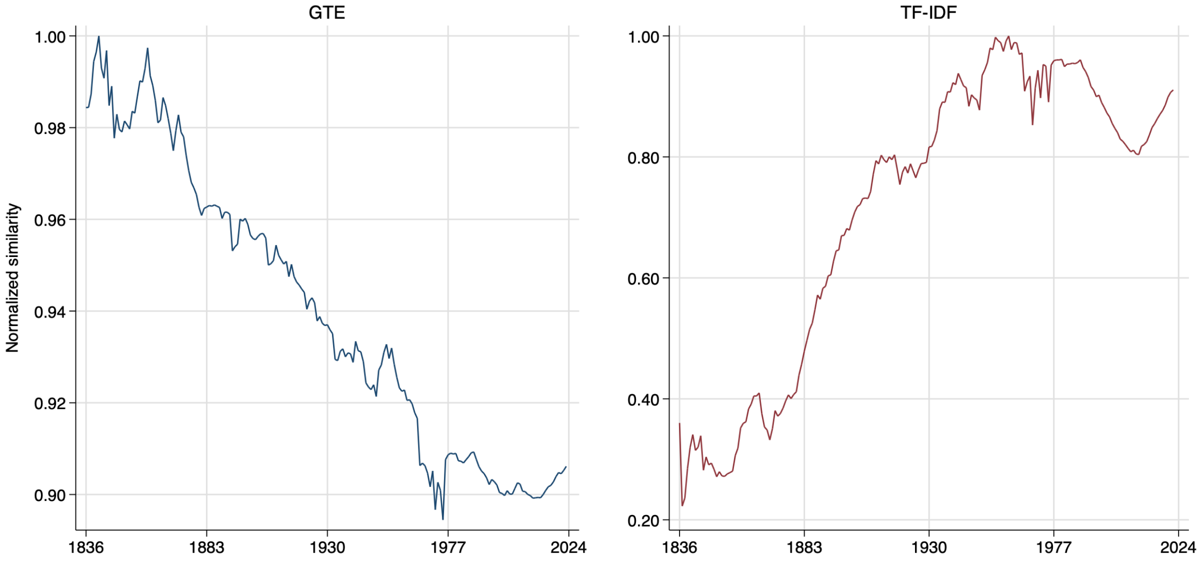


Figure 1: Average pairwise patent claims similarity by representation and year

doc2vec, (iii) Universal Sentence Encoder (USE), (iv) Sentence-BERT (S-BERT), (v) OpenAI's `text-embedding-3-large`[1], (vi) General Text Embedding (GTE), and (vii) Patent-level Representation Learning using Citation-informed Transformers (PaECTER).

In the interference task, PaECTER and GTE significantly outperform other models, including the proprietary OpenAI embeddings. Interferences were US patent office administrative proceedings that occurred when two or more independent inventors submitted applications containing identical claims of invention (Ganguli, Lin, and Reynolds 2020). Thus, this task's design combines modern patent application text and expert judgment of near-identical similarity. PaECTER achieves a precision-recall area under curve (PR AUC) of 0.65, closely followed by GTE at 0.64, substantially surpassing S-BERT (0.52) and TF-IDF (0.44). This difference is economically meaningful: at each model's F10-maximizing threshold (which heavily prioritizes finding true interference pairs over potential false positives), a patent examiner searching for interferences using GTE instead of TF-IDF could reduce false positives by a factor of 4.3 while identifying slightly fewer true positives.

[1] Released in January 2024, this model likely uses similar technology and training data to GPT-4.

In our human annotation task, GTE demonstrates superior alignment with non-expert human judgment about patent similarity, achieving an R^2 of 0.38, significantly outperforming PaECTER (0.27) and TF-IDF (0.12). GTE's robust performance across both tasks underscores its overall effectiveness in capturing patent similarity.

In our technology classification task, S-BERT representations outperform other models in correctly identifying common top-level technology sections. PaECTER and then GTE are competitive and far outperform TF-IDF. We weigh the performance on this task less, as by construction, it ignores the dynamics of between-class similarity.

Taken together, the results of our validation tasks suggest that GTE and PaECTER representations produce measures of patent similarity that best match non-expert human judgment and domain-specific, expert legal determinations. Using GTE representations, we find a long-term decline in contemporaneous patent similarity over the past century and a half, suggesting that inventors are "spreading out" over an expanding knowledge frontier (Figure 1). Combining information from both GTE and PaECTER—or combining GTE, PaECTER, and S-BERT together—yields similar results. However, as already shown, using TF-IDF representations suggests a strikingly different conclusion.

We also re-examine trends in "breakthrough" inventions following Kelly et al. (2021) using GTE instead of TF-IDF in the initial representation step. We are able to replicate some qualitative features of their study, with fewer discretionary researcher choices.

To corroborate the GTE-based finding of a long-run decline in contemporaneous invention similarity, we estimate the rate of interference over nearly 150 years. In theory, the rate of interference declines when inventors choose projects that are less similar. We combine newly-digitized 1864–1901 *Registers of Interferences* with summary statistics of interferences for 1950–1962 and 1981–1993 and our 1998–2014 database of interference decisions. Consistent with our GTE-based estimates, the rate of interference also exhibits a secular decline. Importantly, since interferences before 1998 were not used to validate the representations produced by GTE, this result represents independent, out-of-sample confirmation of the long-run decline in invention similarity.

In a companion paper (Ganguli et al. 2024), we develop a theory where the decline in contemporaneous invention similarity is related to recent findings on long-run invention dynamics, including the increasing burden of knowledge (Jones 2009), increasing R&D spending (Hirschey, Skiba, and Wintoki 2012), declining R&D productivity (Bloom et al. 2020), and constant R&D spillovers (Lucking, Bloom, and Van Reenen 2019). The increasing burden of knowledge raises the fixed costs of inventing over time. This restricts entry into invention as the space of inventions grows, leading inventors to spread out over an expanding knowledge frontier. Ideas get harder to find because there are weaker positive knowledge spillovers from "neighbors" that are now more distant in idea space. Inventors increase their own R&D inputs in response to weaker spillovers, thus reducing own-R&D productivity. (On net, total spillovers may be roughly constant as increasing idea distance is offset by increases in own-R&D investment.)

Our analysis of NLP model selection and validation yields several important takeaways for economists, particularly those studying innovation. These recommendations stem from (i) our finding of substantial differences in model performance on specific tasks and (ii) the diverging trends they measure in key economic concepts. First, we recommend testing several models, especially when some models are proprietary and expensive. (In our results, open-source GTE is competitive with proprietary OpenAI.) Second, researchers should design validation tasks specific to their domain and research questions, using these to motivate model selection. Third, for innovation studies, our validated embeddings can serve as a benchmark if no newer alternatives exist. To facilitate future research, we plan to provide both GTE and PaECTER representations for US patents from 1836–2023. Our results suggest these representations should be the current standard for patent text analysis in economics research. However, if new candidate models emerge or researchers develop their own, we encourage using our validation tasks alongside research question-specific ones.

Our paper builds on work measuring the similarity of inventions and ideas. Earlier work uses bibliometric features such as overlapping patent classifications (Akcigit, Kerr, and Nicholas 2017; Clancy 2018; Fleming 2001), keywords (Azoulay, Fons-Rosen, and Graff Zivin 2019), or citations (Berkes and Gaetani 2020; Wang, Veugelers, and Stephan 2017). Others use the workhorse model TF-IDF (Kelly et al. 2021) or newer NLP models including doc2vec (Feng 2020) and S-BERT (J. Lee and Hsiang 2019). Some recent work evaluates the performance of NLP models (Arts, Cassiman, and Gomez 2018; Arts, Hou, and Gomez 2021; Cheng, D. Lee, and Tambe 2022). A typical approach is to validate a single representation using expert judgment (e.g., patent office classifications) or choice behavior (e.g., citations). Compared with this work, our analysis contributes a comparative design that evaluates several leading NLP models against a common set of validation tasks. We also provide general guidelines for innovation researchers using NLP methods to measure economic quantities of interest, design novel validation tasks, and document new facts about invention similarity over time.

Our analysis focuses on pairwise, contemporaneous invention similarity. This measure is distinct from the related concepts of "novel" (Akcigit, Kerr, and Nicholas 2017), "disruptive" (Park, Leahey, and Funk 2023), "breakthrough" (Kelly et al. 2021), or "unconventional" (Berkes and Gaetani 2020) innovations. Those previous measures compare newly-issued patents against prior art, whereas contemporaneous invention similarity measures simultaneous decisions by inventors about where to locate in idea space. However, the choice of representation (and often, similarity itself) is a direct input into the measurement of these prior concepts. Our contribution is to highlight the importance of validation and model selection for constructing measures based on patent text.

Finally, our results have potential applications for innovation economics. For example, similarity measures may be used for constructing matched controls in studies of localized knowledge spillovers (Ganguli, Lin, and Reynolds 2020; Jaffe, Trajtenberg, and Henderson 1993; Murata et al. 2014; Thompson and Fox-Kean 2005). Similarity measures seem especially useful for empirical study of theories of idea space (Akcigit, Kerr, and Nicholas 2017;

Clancy 2018; Dasgupta and Maskin 1987; Olsson 2000).

The rest of the paper is structured as follows. Section 2 outlines a pipeline for the construction and validation of measures (including similarity) from patent text. Section 3 compares the performance of different representations in our validation tasks. Section 4 shows that the choice of representation affects the measurement of trends in patent similarity. Section 5 concludes.

2. Framework and Pipeline

For economists, NLP is an intriguing tool for measuring quantities of economic interest. However, the rapid pace of innovation in NLP, characterized by several qualitative breakthroughs in recent decades, has led to a proliferation of models. This diversity poses a challenge: different NLP models, despite appearing reasonable *ex ante*, can produce strikingly different representations of the same economic concept (Ash and Hansen 2023). Thus, traditional approaches of selecting a single model and validating it (Gentzkow, Kelly, and Taddy 2019) may no longer suffice. Instead, model selection should be an integral part of the NLP paper pipeline in economics, with validations designed to identify the best-performing models for specific tasks.

We propose a comprehensive pipeline for creating robust measures of economic concepts using patent text. The initial step arises from an economist's aim to measure an economic quantity. This quantity could be based on intuition, informal theory, or derived from a structural economic model. Our pipeline emphasizes a critical distinction that the numerical representation of text is separate from the economic quantity of interest.

Figure 2 provides a schematic view of our proposed pipeline, which consists of four key steps. Each of these steps is discussed in more detail below.

1. Representation: Mapping each patent to a location in idea space.
2. Measurement: Quantifying the concept of interest using the chosen representation.

3. Validation-based selection: Evaluating multiple representations using purpose-built, domain-specific validation tasks.
4. Model selection: Choosing the representation that best aligns with human judgment or “ground truth.”

Our primary contribution lies in emphasizing the critical importance of steps 3 and 4, which go beyond traditional validation approaches in economics.

The usefulness of our pipeline extends beyond patent analysis, offering potential benefits to other subfields in economics. Our pipeline provides a systematic way for economists to evaluate and incorporate state-of-the-art NLP techniques into their research.

2.1. Data

We use the full text of claims in all US utility patents issued 1836–2023. For historical patents issued 1836–1975, we use the digitized patent text from the Patents Core database by ProQuest. For modern patents issued 1976–2023, we use the full text of patents from PatentsView (U.S. Patent and Trademark Office 2023). We also use patent metadata, the text of modern patent applications, historical and modern data on patent interferences, and human annotations. These data are described as they are used in later sections.

2.2. Representation: Mapping Patents to Idea Space

We denote a representation of patent text p_i to a location in idea space as:

$$m(p_i) = C_i^m, \tag{1}$$

where m refers to the particular method or model used to map the patent to a location in idea space and C_i^m refers to the coordinate vector for patent i based on method m.

This is Step 1 in Figure 2. Various methods have been employed to map patents to idea space, each with its own strengths and limitations. These methods broadly fall into two categories: classification-based and text-based approaches.

Text of patent 1
Text of patent 2
Text of patent 3

↓

Step 1: Numerical representations in three idea spaces $(C_i(m_k), k \in \{A, B, C\})$

Repr. A

$$\begin{bmatrix} 0.44 & 0.03 & 0.55 & 0.44 \\ 0.42 & 0.33 & 0.2 & 0.62 \\ 0.3 & 0.27 & 0.62 & 0.53 \end{bmatrix}$$

Repr. B

$$\begin{bmatrix} 0.13 & 0.51 & 0.18 \\ 0.85 & 0.49 & 0.85 \\ 0.51 & 0.07 & 0.43 \end{bmatrix}$$

Repr. C

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

↓

Step 2: Measurements of pairwise similarities $(Sim^{m_k}(p_i, p_j))$

Repr. A

Pat 1	Pat 2	Cos. Sim.
1	2	0.82
1	3	0.94
2	3	0.87

Repr. B

Pat 1	Pat 2	Cos. Sim.
1	2	0.71
1	3	0.48
2	3	0.96

Repr. C

Pat 1	Pat 2	Cos. Sim.
1	2	1
1	3	0
2	3	0

↓

Step 3: Validation-based selection $\left(V^l(m_k), l \in \{(i), (ii), (iii)\}\right)$

Task (i)

Repr.	Perf.	Rank
Repr. A	0.91	1
Repr. B	0.87	2
Repr. C	0.84	3
Baseline	0.51	4

Task (ii)

Repr.	Perf.	Rank
Repr. A	0.46	1
Repr. B	0.23	2
Repr. C	0.18	3
Baseline	0.03	4

Task (iii)

Repr.	Perf.	Rank
Repr. A	0.85	2
Repr. B	0.93	1
Repr. C	0.73	3
Baseline	0.05	4

↓

Step 4: Compute downstream measure based on the best representation:

For example, "Breakthrough" patents $(q^m(p_i))$ (Kelly et al. 2021) or average patent pair similarity $(q^m(p_i, p_j))$ by year (this paper).

Figure 2: Overview of the NLP pipeline

A traditional approach uses patent office classifications (e.g., Jaffe 1986; Jaffe, Trajtenberg, and Henderson 1993). These classifications, assigned by specialized patent examiners, are primarily administrative tools designed to facilitate searches for relevant prior art. The US Patent and Trademark Office (USPTO) currently uses the Cooperative Patent Classification (CPC), divided into eight top-level sections[2].

In our framework, a class-based mapping would represent each patent as a vector with 1s in the position of its assigned class(es) and 0s in all other positions. Such a representation would resemble Representation C in Figure 2. While straightforward, this approach has limitations due to its coarse granularity: it treats all patents within a class as equally similar, and all patents in different classes as equally dissimilar.

More recent approaches apply NLP models to patent text to produce numerical representations of each patent. These representations have the potential to offer finer granularity and a richer map of idea space.[3] Such representations would resemble Representations A and B in Figure 2.

TF-IDF. The workhorse model TF-IDF (Sparck Jones 1972) represents patents based on word frequency, weighted by the inverse of word frequency across all patents. The TF-IDF vector for patent i is:

$$c_{i,k}^{TFIDF} \equiv TF_{i,k} \cdot IDF_{i,k} \tag{2}$$

where $TF_{i,k} \equiv n_{i,k}/\sum_j n_{i,j}$ (Term Frequency) and $IDF_{i,k} \equiv log(\frac{\text{\# of patents in corpus}}{\text{\# of patents in corpus with word k}})$ (Inverse Document Frequency).

[2] The eight top-level sections are (a) Human Necessities, (b) Performing Operations, Transporting, (c) Chemistry, Metallurgy, (d) Textiles, Paper, (e) Fixed Construction, (f) Mechanical Engineering, Lighting, Heating, Weapons, Blasting Engines or Pumps, (g) Physics, (h) Electricity. There is also residual category for new technological developments. These eight sections are further sub-divided into over 100 "three-digit" classes.

[3] See Bochkay et al. (2023), Gentzkow, Kelly, and Taddy (2019), and Grimmer, Roberts, and Stewart (2022) for reviews of the use of NLP methods in economics and neighboring disciplines. See Smith (2020) for an accessible introduction to numerical text representations from one-hot encoding to contextual embeddings.

Neural Network-based Models. Recent advances in NLP have led to the development of sophisticated models that produce vector representations, or distributed embeddings. These models include open-source options like doc2vec (Le and Mikolov 2014; Mikolov et al. 2013), USE (Cer et al. 2018), S-BERT (Devlin et al. 2019; Reimers and Gurevych 2019), and GTE (Li et al. 2023), as well as proprietary ones such as those released by OpenAI (OpenAI 2024). Additionally, domain-specific models have been developed such as PaECTER (Ghosh et al. 2024).

Embedding sizes vary considerably across models. doc2vec typically produces embeddings of 100-300 dimensions, USE generates 512-dimensional vectors, S-BERT and GTE yield larger embeddings of 768 or 1,024 dimensions, and PaECTER, designed specifically for patents, uses 1,024-dimensional embeddings. OpenAI embeddings have dimension of 1,536 by default, but they use Matryoshka representation learning technology, allowing reductions in embedding size with limited loss in performance (Kusupati et al. 2024).

The objective functions and training processes also differ significantly. doc2vec employs a skip-gram approach, predicting context words given an input word. In contrast, USE and subsequent models involve a two-stage training process: an initial unsupervised stage followed by supervised fine-tuning on downstream tasks using supervised data. This second stage typically includes paraphrase identification and sentence similarity tasks, with the explicit goal of producing embeddings that are broadly applicable and semantically meaningful. GTE utilizes a contrastive learning objective (Li et al. 2023), which explicitly aims to both bring similar sentences closer and different ones further apart. PaECTER adapts this approach to the patent domain, fine-tuning on patent citation data. Details of the OpenAI embedding models are proprietary, but the technology and training data is likely similar to that underlying the large language model GPT-4.

The development of these models involves numerous engineering decisions that significantly impact performance. Unlike structural economic models, where evaluation often relies on examining the functional forms, the extensive engineering choices in ML make it challenging to assess models *ex ante* from first principles. Our approach circumvents this issue by focusing on the validation of model outputs. This strategy allows us to select the most effective model for our analysis, based on performance, rather than technical specifications.

2.2.1. Visualizing Idea Spaces

To develop intuition about these different representations, Figure 3 displays two-dimensional projections of high-dimensional idea spaces based on S-BERT and TF-IDF representations, chosen for their striking contrast. (See Appendix A for details on the construction of this figure.)

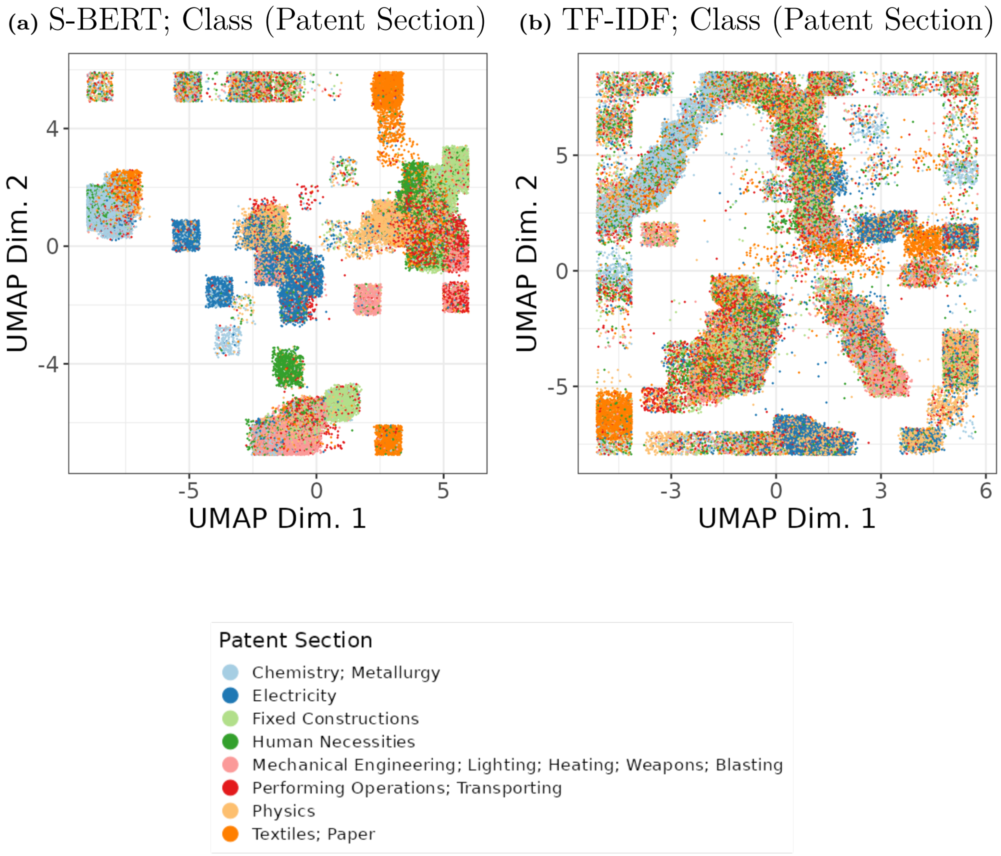
The visual differences observed in Figure 3 highlight the potential impact of representation choice on downstream analyses. By overlaying patent class colors on both S-BERT and TF-IDF representations, we demonstrate that these models differ dramatically in how they cluster patents from the same class, with S-BERT showing noticeably tighter groupings by patent class compared to TF-IDF.

Intriguingly, the S-BERT visualization reveals nuances that align with expert knowledge but are not captured by patent classifications alone. For instance, in Figure 3 Panel A, a cluster of semiconductor patents (blue, near (-5, 0)) is positioned between materials science patents (light-blue) and a broader electricity cluster. This arrangement reflects the interdisciplinary nature of semiconductor innovations, combining aspects of materials science and electricity, and it demonstrates S-BERT's ability to capture complex relationships between technological domains.

2.3. Measuring Concepts from Patent Text

The second step in our pipeline is to use numerical representations of patent text to measure a concept of economic interest. Researchers have defined a number of such measures. A core measure is pairwise similarity, defined as the cosine similarity between patent vector representations:

$$Sim^m(p_i, p_j) \equiv \frac{C_i^m \cdot C_j^m}{||C_i^m||||C_j^m||} \tag{3}$$



(a) S-BERT; Class (Patent Section) **(b)** TF-IDF; Class (Patent Section)

Figure 3: Uniform Manifold Approximation and Projection (UMAP) plots for S-BERT and TF-IDF representations

Notes: The plot is based on a sample of 111,251 patents stratified by patent class (USPTO Section) and quarter-century period. To constrain extreme values, the data were winsorized at the 5% and 95% levels along both axes.

where m denotes a specific model. This is illustrated by Step 2 in Figure 2.

Other measures from the literature can also be expressed as functions of patent representations. For instance, Kelly et al. (2021) define patent "importance" as:

$$q^m(p_i) \equiv \frac{\sum_{j\in\mathcal{F}} Sim^m(p_i,p_j)/|\mathcal{F}|}{\sum_{k\in\mathcal{B}} Sim^m(p_i,p_j)/|\mathcal{B}|} \tag{4}$$

where $\mathcal{F}$ and $\mathcal{B}$ denote sets of patents published in the 5 years after and before patent i, respectively.

Notice that the measurement of these concepts of economic interest depends both on the conceptual definition and on the choice of patent representation. Even for the same quantity, different representations can lead to different measurements. While the choice of a conceptual measure can generally be guided by theory or other *a priori* considerations, the choice of representations amounts to choosing between different "black-box" methods.

2.4. Validation-Based Selection

Given a concept of interest, how should researchers choose between alternative representations? We propose validation-based selection as a crucial step in our pipeline (Step 3 in Figure 2). This process requires external measures ("ground truth") that have a clear theoretical connection with the concept of interest, often available (or obtainable for a cost) only for a subset of the data.

Formally, given a concept c, a representation-specific measurement f^{m_i}, and a score function S, validation evaluates each representation m_i to select the best mapping:

$$V(m_i) = S\left(f^{m_i}(\mathbf{p}), c(\mathbf{p})\right) \tag{5}$$

In practice, we implement an empirically-feasible version:

$$V^j(m_i) = S^j\left(f^{m_i}(\mathbf{p^j}), g^j(\mathbf{p^j}) \mid \mathbf{p^j}\right) \tag{6}$$

where $g^j(\mathbf{p^j})$ is a ground truth, $f^{m_i}(\mathbf{p^j})$ is a measurement based on representation m_i, and S^j quantifies the correspondence between measures derived from m_i and the ground truth. The score function could be correlation, mean squared error, or other appropriate metrics. $V^j(m_i)$ provides a score or ranking for each model m_i, identifying the best representation according to validation criterion j.

If different validations suggest different representations as optimal, researchers should employ a deliberate, multi-faceted approach to make a final decision. This process should consider:

1. **Relevance to research question:** Prioritize validation tasks that most closely align with the specific economic concept or question being studied.
2. **Quality and reliability of ground truth:** Assess the reliability and representativeness of each validation task's ground truth data.
3. **Performance differentials:** Consider the magnitude of performance differences between models across validation tasks.
4. **Consistency across tasks:** Look for models that perform well across multiple validation tasks, even if not optimal in each.
5. **Domain expertise:** Leverage economic theory and subject-matter knowledge to weigh the importance of different validation tasks.

This nuanced approach acknowledges the complexity of economic concepts and the limitations of individual validation tasks. By applying our validation-based selection pipeline, researchers can identify a set of best-performing models along with their relative strengths. These relative strengths can be formalized as weights, allowing for informed model aggregation or selection based on specific research needs.

For concreteness, we can apply our framework to the interference validation task (discussed in detail in Section 3.1). Here, $\mathbf{p^j}$ represents the set of patent applications in interference cases. The ground truth function $g(\mathbf{p^j})$ creates pairwise combinations of these applications, producing a Boolean vector where entries are 1 if the corresponding pair was in an interference case. This approach leverages the expertise of patent examiners in identifying highly similar applications. The function $f^{m_i}(\mathbf{p^j})$ computes pairwise similarities based on the representation m_i. We use the Receiver Operating Characteristic Area Under the Curve (ROC AUC) or the Precision-Recall Area Under Curve (PR AUC) as our score function S^j, comparing $f^{m_i}(\mathbf{p^j})$ to $g(\mathbf{p^j})$.

For the human validation task (Section 3.2), $\mathbf{p^j}$ represents a set of patent pairs sampled to have varying levels of similarity according to each model. The ground truth function $g(\mathbf{p^j})$ is the human judgment on which pair in each comparison is more similar. This leverages non-expert perception of patent similarity. The function $f^{m_i}(\mathbf{p^j})$ computes pairwise similarities based on the representation m_i, using the same text segments shown to human annotators. Our score function S^j compares the model's ranking of pair similarities to the human judgments, measuring how well $f^{m_i}(\mathbf{p^j})$ aligns with $g(\mathbf{p^j})$.

Our approach to validation differs from some prior literature in that it is intrinsically linked with model selection. This validation-based selection, while common in fields like forecasting and machine learning, has been less prevalent in empirical economics using NLP measures. In forecasting and machine learning, it is often acknowledged that we cannot select the best model *a priori*, necessitating a structured selection procedure. Our work demonstrates that this principle extends to NLP applications in economics, where model selection can have substantial effects on results and interpretations.[4]

[4] Model selection is a well-established practice in econometrics and forecasting, often using criteria such as the Akaike Information Criterion (AIC). In machine learning, out-of-sample testing is commonly used for model selection, where models are evaluated on data not used for training. The "winning" model is typically determined by a score function, such as root mean squared error. Ash and Hansen 2023 provide examples outside of innovation economics where different text representations lead to divergent conclusions, highlighting the need for careful validation in NLP applications to economics.

3. Validation Task Results

3.1. Interferences

Our first validation task uses patent interferences. Patent interferences were a unique feature of US patent law, which applied to patents filed between 1836 and March 2013.[5]. Patent interferences were USPTO administrative proceedings that decided the priority of invention when two or more independent parties claimed to have invented the same thing at the same time. An interference was suggested by a specialized patent examiner when, during their search for relevant prior art, they encountered at least one other pending US patent applications containing the "same patentable invention" (37 CFR § 1.601). Thus, patent interferences represent expert judgment that two independent patent applications contain identical legal claims.

3.1.1. Interference Data

We select patent applications from a database of 215 interference cases decided 1998–2014. These decisions were publicly available through the USPTO's "e-FOIA Reading Room" and encoded by Ganguli, Lin, and Reynolds (2020).[6] Each interference case involves two or more independent *parties* with competing, simultaneous claims to the same patentable invention. Each party has one or more patent *applications* corresponding to the content of the interference. In our database of 215 cases, we identify 440 distinct patent applications. Using these interference cases and applications, we construct 96,580 ($= \frac{1}{2}(440^2 - 440)$) *application pairs*. Of these application pairs, we identify 322 *interfering pairs*—meaning two patent applications from independent (opposing) parties that make overlapping claims of invention.

We represent the text of each application using the seven NLP models mentioned in the introduction: (1) TF-IDF, (2) doc2vec, (3) USE, (4) S-BERT `all-mpnet-base-v2`,

[5] More detail and institutional background on patent interferences can be found in Ganguli, Lin, and Reynolds (2020)

[6] The 215 cases are a subset of the database by Ganguli, Lin, and Reynolds (2020) selected based on the availability of the full text of applications and the claims in interference.

Table 1: Example rows from the patent pair dataset used for interference validation

		Cosine similarity based on:								
ID App. 1	ID App. 2	Class	TF-IDF	doc2vec	USE	S-BERT	Open AI	PaECTER	GTE	Int.
9885259	11064123	0.00	0.02	0.86	0.07	0.27	0.41	0.89	0.49	0
10239566	11517991	0.79	0.68	0.74	0.76	0.66	0.56	0.93	0.73	0
10116112	10909561	0.00	0.01	0.54	0.18	0.26	0.36	0.87	0.49	0
10005999	11614142	0.00	0.02	0.59	0.26	0.20	0.34	0.87	0.55	0
11255647	11428279	0.00	0.28	0.62	0.59	0.80	0.75	0.97	0.73	1

Notes: Columns show patent IDs, similarity scores from different patent representations, and a binary label indicating whether this pair was part of an interference case.

(5) OpenAI's `text-embedding-3-large`, (6) GTE, and (7) PaECTER. For every pair of applications, we compute the cosine similarity based on the vector representations from each of these seven models. We also construct an alternative measure of similarity based on the number of shared CPC classes between application pairs (for more detail on CPC classes, see Section 3.3). Table 1 shows an excerpt of the resulting application-pair database. Each row is a unique application pair. Columns are application identifiers, similarity scores, and a true interference indicator.

3.1.2. Interference Results

Next, we evaluate the performance of alternative representations. The task is to classify application pairs in interference. We provide some intuition about the economic magnitudes of performance differences by considering a hypothetical scenario where a patent examiner wants to identify application pairs that are likely to be in interference. Patent examiners can use representations of patents to compute the similarity of all pairs and rank them from most similar. In this situation, a classifier involves the ranked similarity scores and a cutoff, above which pairs will be considered interference candidates. The patent examiner and their staff would then review these interference candidates to determine which are true and which are false according to their expertise.

The examiner cares about (i) how often a classifier correctly classifies true interfering pairs (true positives) and (ii) how often a classifier incorrectly classifies non-interfering pairs as interfering pairs (false positives). They may place a different weight on each of these

Table 2: Rankings based on threshold-based metrics

(a) Separate F1-max. thresh.

Rank	Repr.	TP	FP	F1
1	PaECTER	168	58	0.67
2	GTE	170	82	0.64
3	OpenAI	182	123	0.63
4	S-BERT	143	90	0.56
5	TF-IDF	110	67	0.48
6	USE	85	58	0.40
7	doc2vec	50	72	0.25
8	Class	98	792	0.17

(b) Separate F10-max. thresh.

Rank	Repr.	TP	FP	F10
1	PaECTER	265	1,862	0.90
2	GTE	259	1,222	0.90
3	OpenAI	255	1,118	0.89
4	S-BERT	250	3,001	0.82
5	TF-IDF	253	5,306	0.77
6	USE	235	4,984	0.72
7	Class	209	6,255	0.62
8	doc2vec	198	17,944	0.44

Notes: F1/F10 scores and underlying true positives and false positives with a different thresholding strategy in each panel. The total number of patents is 440; the total number of patent pairs is 96,580; the total number of interference cases is 312.

criteria. Here, true positives represent the statutory obligation of the USPTO to determine priority of invention, while false positives incur investigation costs through examiner and staff time. Thus, there may be a trade-off between classifiers that detect many true positives but also many false positives, and those which detect fewer true positives but also fewer false positives. This can be formalized as the tradeoff between recall and precision. The recall of a classifier is the share of total interferences it correctly identifies. The precision of a classifier is the share of total pairs correctly classified as interfering.

Note that many different classifiers can be built from a given similarity measure. Different threshold levels will lead to classifiers with different performance in terms of selecting true positives and false positives.

As a starting point, consider the case where the examiner values identifying promising cases (recall) and not overburdening staff (precision) equally. The so-called F1 score does exactly this. Table 2a shows that PaECTER has the highest F1 score (67%), followed closely by GTE (64%) and OpenAI embeddings (63%), with all three significantly outperforming S-BERT (56%) and TF-IDF (48%) at each representation's F1-maximizing threshold.

Next, consider the case where the examiner places higher priority on identifying potential interferences (true positives) than on minimizing investigative effort (false positives)—i.e., staff time is considered less costly relative to missed interferences. The F10 score weights recall ten times more than precision. Table 2b shows that at each measure's F10-maximizing threshold, PaECTER, GTE, and OpenAI embeddings (with almost identical F10 scores of 90%, 90%, and 89%) retrieve 255–265 true positives, slightly outperforming S-BERT (250) and TF-IDF (253). More importantly, the top three models (PaECTER, GTE, and OpenAI) significantly reduce false positives compared to S-BERT and TF-IDF. Specifically, they reduce false positives by a factor of 1.6–2.7 compared to S-BERT and by a factor of 2.8–4.7 compared to TF-IDF, while maintaining a higher true positive rate. This dramatic reduction in false positives would significantly decrease the number of unnecessary investigations, leading to more efficient use of examiner time and resources. Text-based classifiers based on USE and doc2vec prove uncompetitive. The shared-class-based classifier consistently lags behind all NLP-based methods except doc2vec.

We next report results based on two metrics which summarize classifier performance across all possible thresholds. Receiver Operating Characteristic Area Under the Curve (ROC AUC) evaluates the trade-off between true positive and false positive rates across all possible thresholds. Precision-Recall Area Under Curve (PR AUC) measures the trade-off between precision and recall across all possible thresholds.[7]

Across both ROC AUC and PR AUC, we find that PaECTER, GTE, and OpenAI embeddings best predict interference cases, followed by S-BERT and then TF-IDF (Table 3). The PR AUC differences are more pronounced, as expected for an imbalanced binary prediction problem. PaECTER achieves a PR AUC of 0.65, closely followed by GTE at 0.64 and OpenAI at 0.62, significantly outperforming S-BERT (0.52) and TF-IDF (0.44).

Across threshold- and non-threshold-based comparisons, classifiers based on PaECTER, GTE, and OpenAI embeddings consistently demonstrate superior performance, materially outperforming even recent models like S-BERT and other approaches. The differences in

[7] See Davis and Goadrich (2006) for a comparison of the two measures.

Table 3: Rankings based on non-threshold-based metrics

(a) ROC AUC

Rank	Repr.	ROC AUC
1	PaECTER	0.99
2	GTE	0.99
3	OpenAI	0.99
4	S-BERT	0.98
5	TF-IDF	0.98
6	USE	0.96
7	Class	0.85
8	doc2vec	0.84

(b) PR AUC

Rank	Repr.	PR AUC
1	PaECTER	0.65
2	GTE	0.64
3	OpenAI	0.62
4	S-BERT	0.52
5	TF-IDF	0.44
6	USE	0.36
7	Class	0.21
8	doc2vec	0.16

Notes: ROC and PR AUC scores for different patent text representations on predicting interference cases.

performance are substantial and economically significant. Importantly, all models, including the worst-performing doc2vec, technically pass this validation task by predicting interferences better than chance. The stark performance gaps highlight generational differences in NLP technologies. The result that all models alone do better than chance underscores the critical importance of our validation-based selection pipeline and a comparative approach.

3.2. Non-Expert Human Judgment

Second, we design and implement a non-expert human validation task to assess the performance of four models: PaECTER, GTE, S-BERT, and TF-IDF.[8] This task complements the interference validation task by focusing on patents with varying levels of similarity and drawing from a broader time period. Our primary objective is to determine which model aligns more closely with the general human judgment.

A main challenge is that humans without special training struggle to place objects on absolute scales (Carlson and Montgomery 2017). Therefore, we asked research assistants (RAs) to make *relative* judgments of similarity. We presented them with two sets of patent

[8] We eliminated USE, class overlap, and doc2vec at this stage because of their non-competitiveness on the interference task. We eliminated OpenAI because its good-but-not-best performance did not justify including a proprietary and expensive model in further analyses.

pairs and asked them to select which pair contained patents more similar to each other.

To ensure that the task was feasible for human annotators, we sampled patent pairs separately for each model. The pairs were selected so that, according to the model being tested, they were at least 50 percentiles apart in terms of similarity. For example, if one pair was in the 90th percentile of similarity, the other pair could be no higher than the 40th percentile. This approach helped to create a clear distinction between the pairs, making the task more manageable for human annotators.

We provided detailed instructions to the annotators (see Appendix B for full instructions and an example). The RAs were asked to consider four factors when comparing the patent pairs: (i) the general field or domain of each patent, (ii) the specific problem each patent is trying to solve, (iii) key components of the solution each patent proposes, and (iv) any other major similarities or differences between the patents in each pair.

Each of the four annotators was asked to make 100 comparisons between two sets of patent pairs. For each patent, we presented the annotators with two specific fragments of text: the "improvement in" statement extracted from the patent and the first 500 characters of the claims section. This focused approach allowed annotators to quickly grasp the essence of each patent without being overwhelmed by technical details. The embeddings used for model comparisons in this task were calculated using the same text segments shown to the human annotators.

Annotators were encouraged to use online resources to understand unfamiliar terms or concepts, but were instructed to avoid reading parts of the patent outside the provided snippet. We also emphasized that many patent pairs might be only tenuously connected, and asked annotators to think creatively about how seemingly dissimilar patents might be solving similar problems or using related technologies. The annotators provided their final judgment as a single number (1, 2, or 0 if unsure).

Table 4: Human Agreement with Embedding-Based Similarity Rankings

	Dep. Var.: More similar pair = 1			
	PaECTER	GTE	BERT	TF-IDF
(Intercept)	0.28^{***}	0.20^{**}	0.24^{***}	0.37^{***}
	(0.07)	(0.06)	(0.07)	(0.07)
Human Choice = 1	0.51^{***}	0.62^{***}	0.54^{***}	0.35^{***}
	(0.09)	(0.08)	(0.09)	(0.10)
R^2	0.27	0.38	0.29	0.12
Adj. R^2	0.26	0.38	0.28	0.11
Num. obs.	83	90	91	89

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Notes: Regression results showing the agreement between the 4 human annotators and the relative similarity rankings of patent pairs according to different patent text representations.

3.2.1. Human Annotation Results

To analyze the agreement between human judgments and embedding-based similarity rankings, we use the following regression:

$$I[Sim(1) > Sim(2)]^{Emb} = \beta_0 + \beta_1 I[Choice = 1]^{Human} + \epsilon \tag{7}$$

where $Emb \in$ {PaECTER, GTE, BERT, TF-IDF}. The coefficient β_1 represents the increase in the likelihood that the embedding indicates pair 1 is more similar when humans choose pair 1. Higher β_1 suggests stronger human–embedding agreement.

Table 4 presents the results of our human annotation analysis. All embedding models show statistically significant agreement with human judgments, as indicated by the positive and significant coefficients on “Human Choice = 1” (β_1). However, there are clear differences in performance. GTE demonstrates the strongest alignment with human judgments, with the highest coefficient (0.62) and R^2. PaECTER and BERT are the next best performers and show very similar levels of agreement with human judgments, with coefficients of 0.51 and 0.54, respectively. TF-IDF exhibits the weakest agreement with human judgments, as evidenced by its lower coefficient (0.35) and R^2.

Interestingly, while PaECTER showed strong performance in the interference task, its performance here is lower than GTE. This may be attributed to the fact that PaECTER was fine-tuned using patent data from 1985 to 2022 (Ghosh et al. 2024), whereas this task involves historical patents from 1880-1920. GTE's strong performance across different time periods highlights its robustness and generalizability in capturing patent similarity.

3.2.2. Exploring LLMs for Scalable Patent Similarity Validation

Human annotation, while valuable, can be costly and challenging, especially when comparing technical documents like patents. To explore scalable validation methods, we investigated the use of Large Language Models (LLMs) for assessing patent similarity. However, note that we do not view LLMs as direct substitutes for human annotations. Recent research has shown that LLMs often fail to accurately reflect human judgments (Bisbee et al. 2024; Dominguez-Olmedo, Hardt, and Mendler-Dunner 2024; Goli and Singh 2024). We conduct this analysis as part of an ongoing effort in the field to incorporate LLM tools in reasonable and productive ways, acknowledging both their potential and limitations.

Our LLM-based analysis, using Claude 3.5 Sonnet and GPT-4, revealed notable differences not only from human annotations but also between the LLMs themselves in ranking embedding models. (See Appendix C for our prompt.) Claude, like human annotators, selected GTE as the best model, while GPT-4 chose S-BERT, highlighting the limitations of LLMs in capturing human intuition about technological similarity. Despite these differences, both LLMs consistently ranked newer embedding models above the traditional TF-IDF approach, aligning with our human annotation findings in this crucial aspect. This consistency suggests a potential role for LLMs in the preliminary testing of annotation tasks, potentially streamlining the process before deploying to human annotators. For a detailed discussion, refer to Appendix D.

3.3. Patent Office Classifications

Our final validation task uses patent classifications. Patent classifications are assigned to patents by specialized patent examiners. Thus, like interferences, this task relies on expert judgement. On the other hand, this task is more similar to the human annotation task in that it focuses on a coarser degree of similarity. This task also considers an extended sample period, since 1850.

The entire patent classification system, and the assignment of classes to all previously issued patents, is updated frequently. We use the CPC vintage from May 2023, which represents classifications of all patents as of that date. Since patent classifications are primarily administrative tools designed to facilitate searches for relevant prior art, they may reflect judgments of similarity that are relatively more accurate for recent patents. On the other hand, they may be less accurate for historical patents, which may be less relevant for the task of identifying prior art for current patent applications.

We classify patents according to whether their main classification belongs to (i) one of eight top-level CPC technology sections or (ii) one of 123 “three digit” CPC technology classes. We draw random samples of 200 patents from each of these classification groups and quarter-century period from 1850 to 2023. For each pair of patents, we form indicators for common section and common class.

As in the prior tasks, we evaluate the performance of similarity scores based on TF-IDF, S-BERT, GTE, and PaECTER representations as classifiers, in this case classifying patent pairs as belonging to the same section or class.[9]

Figure 4 shows results by level of classification detail (top-level sections versus three-digit classes) and performance metric (ROC versus PR AUC). TF-IDF is uniformly the worst performer across classification detail and performance metrics. In contrast, S-BERT's performance is competitive; it leads the other models in predicting common top-level technology section according to both ROC (Panel 4a) and PR AUC (Panel 4c). S-BERT also leads all other models in predicting common technology class according to PR AUC (Panel 4d), and it ranks second after PaECTER according to ROC AUC (Panel 4b).

[9] Feng (2020) uses doc2vec to create measures of patent similarity and uses patent classes to validate the vectors generated by doc2vec. Compared with this work, our analysis compares the performance of several models.

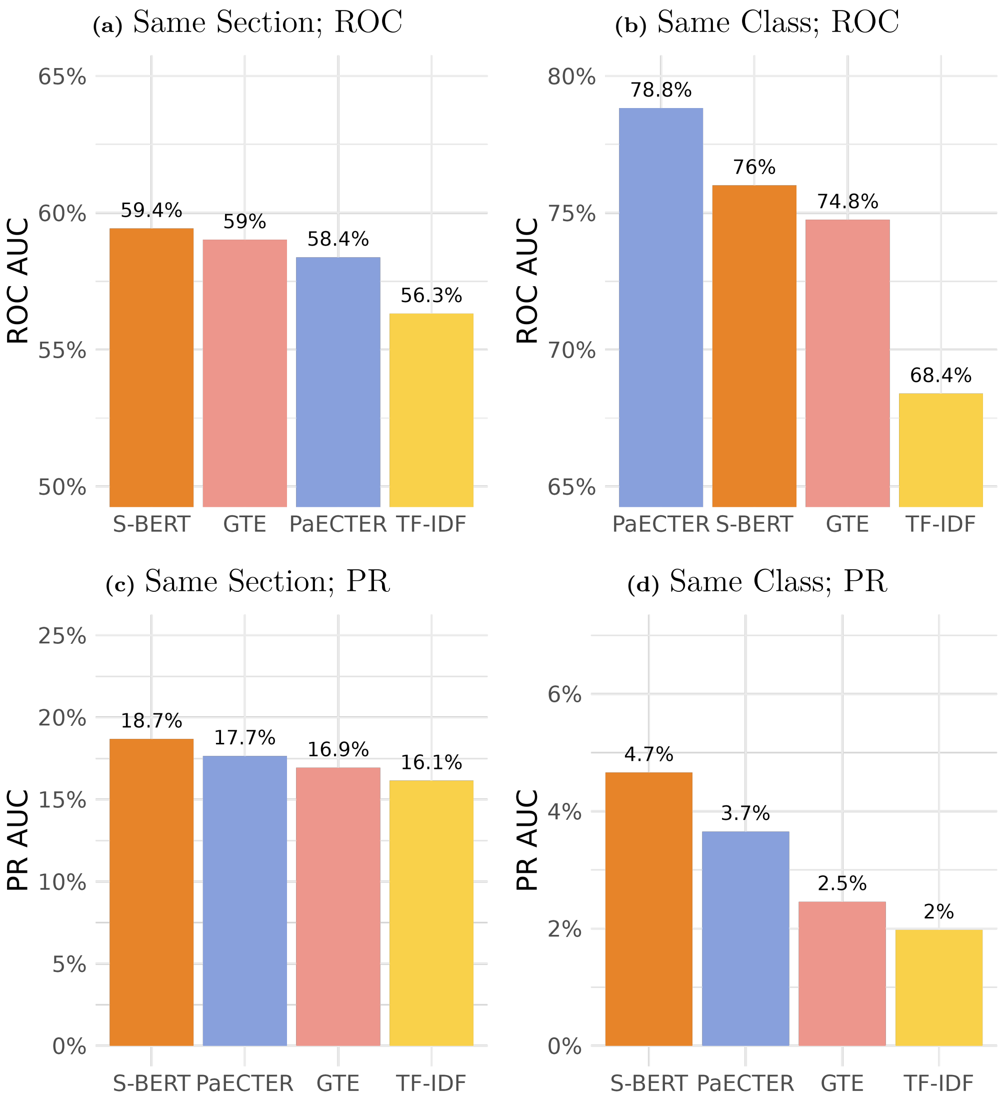


Figure 4: Representation performance on common section and common class tasks

According to both PR and ROC AUC, PaECTER representations outperform GTE in predicting common technology class (Panels 4d and 4b). However, GTE is competitive with PaECTER in predicting common technology section.

Finally, note that by construction, this validation task does not take into account between-class similarity. The design of this validation task therefore emphasizes within-class similarity at the cost of ignoring all between-class similarity. This feature potentially explains some of the divergence in performance results compared with the previous tasks. (We explore these patterns further in Section 4.1.) These results also underscore the importance of multi-faceted validation.

4. Model Selection is Critical for Downstream Economic Measurement

We measure contemporaneous invention similarity over time. We apply four text embedding models to the claims sections of US patents from 1836 to 2023 (the final complete year of patent data). Then, for each model/representation, we compute within-year average pairwise similarity, normalized by the largest similarity value in the time series.

Figure 5 illustrates how different models yield different measured average annual pairwise patent similarity. GTE exhibits a clear downward trend in patent similarity for over a century, from 1836 to 1977. PaECTER suggests that contemporary patent similarity declined from 1900 to 1950, followed by a slight, volatile increase to the present day. S-BERT suggests steadily declining patent similarity from the turn of the 20th century to today. A notable contrast is TF-IDF. TF-IDF measures indicate a sharp increase similarity until the 1950s, followed by fluctuations around a high degree of similarity.

Note that the scale of variability also differs significantly across models. GTE's minimum value is at about 90% of its maximum. PaECTER exhibits very little variability in patent similarity over time, with its minimum value at 97% of its historical maximum. S-BERT's minimum value is at 75% of its historical maximum. TF-IDF similarity is the most variable, with its minimum value at 20% of its historical maximum.

All of these models performed better than chance on our validation tasks. Thus, any single one of them *could* have been chosen to measure trends in contemporaneous pairwise similarity *without* a validation-based selection process. Put differently, a single panel from Figure 5 could have been presented as a main result, followed by individual "validations" showing correlations between that single representation and ground truths. However, our approach provides a framework for more informed model selection, allowing us to differentiate between models based on their performance across multiple tasks.

We can confidently eliminate TF-IDF due to its significant underperformance on all

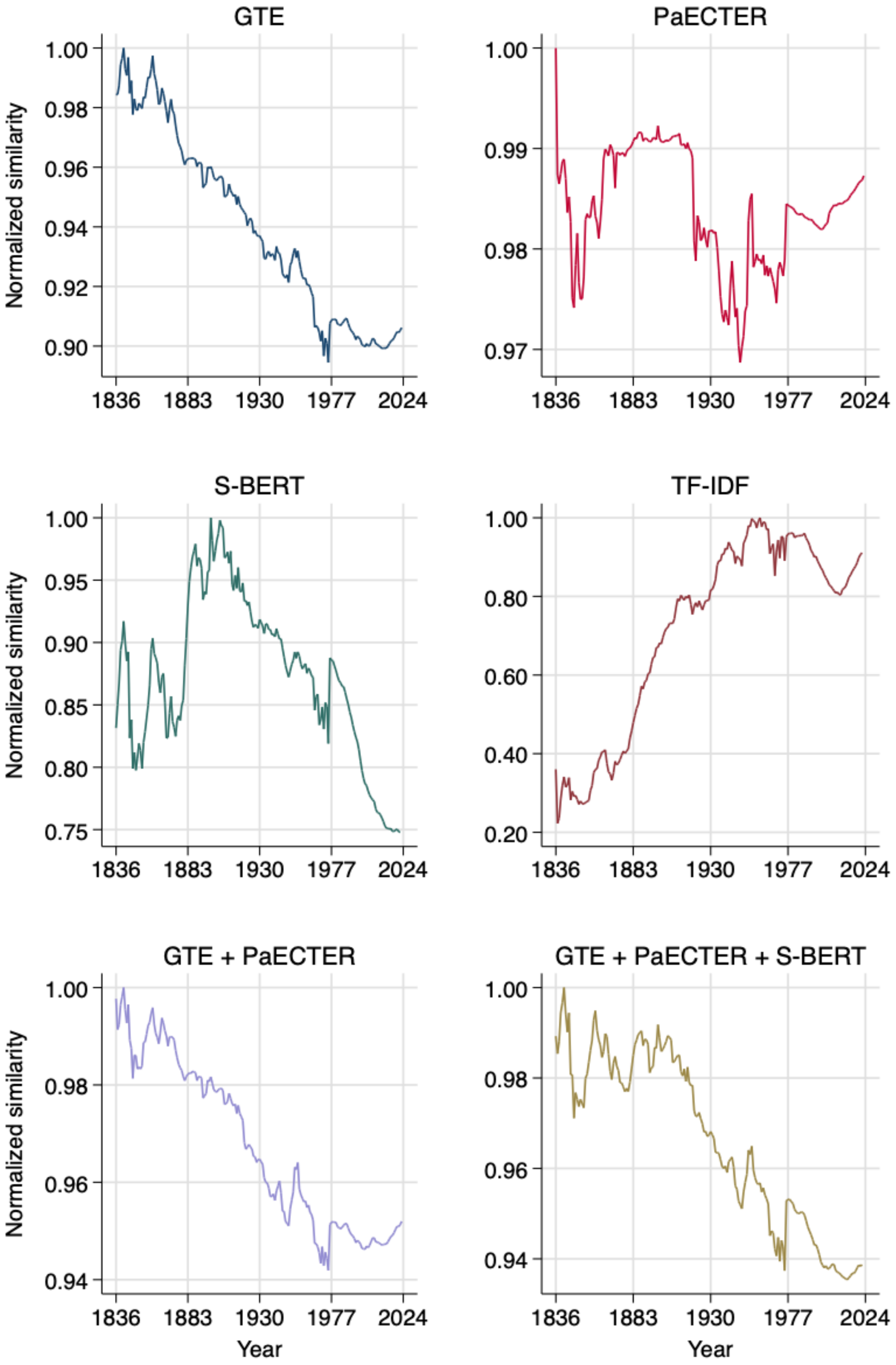


Figure 5: Average pairwise patent claims similarity by representation and year

validation tasks. In order to compare and aggregate the results from the other models, it is useful to review the validation task results.

The interference task used expert judgment about near-identical similarity using modern data (Section 3.1). In this task, PaECTER marginally outperformed GTE. S-BERT was substantially worse at predicting interferences.

The human annotation task used non-expert judgment about a broader sense of similarity using historical data (Section 3.2). In this task, GTE significantly outperformed the other models. S-BERT and PaECTER performed similarly on this task.

Finally, the patent classification task used expert judgment about a coarse sense of similarity using both historical and modern data (Section 3.3). In this task, S-BERT performed best at predicting common technology sections and classes; PaECTER was competitive with S-BERT, and GTE lagged.

Overall, we view GTE and PaECTER as the strongest models across our validation tasks, with S-BERT lagging behind. Both GTE and PaECTER suggest patents declined in contemporaneous similarity in the first half of the 20th century, but they differ in other aspects. Given GTE's success at predicting historical human annotations and PaECTER's strong results in predicting modern patent interferences , one might weigh GTE's historical trends and PaECTER's modern trends more. Alternatively, a equally weighting both GTE and PaECTER indexes (Figure 5) captures these dynamics well: a long-run historical decline in patent similarity from 1836 to 1977, followed by relatively stability. Note that PaECTER's scale implies minimal movement in similarity overall. Because of that, a simple average of the GTE and PaECTER indexes results in a trend similar to GTE's, but on a smaller scale.

On the other hand, one might want to give some weight to the S-BERT results, given its strong performance on the patent class task. The final panel in Figure 5 reports a weighted average of S-BERT (10% weight), GTE (45% weight), and PaECTER (45% weight). Compared with the simple average of GTE and PaECTER, this index shows more stable contemporaneous similarity in the middle 19th century, and a steady decline in similarity since 1977.

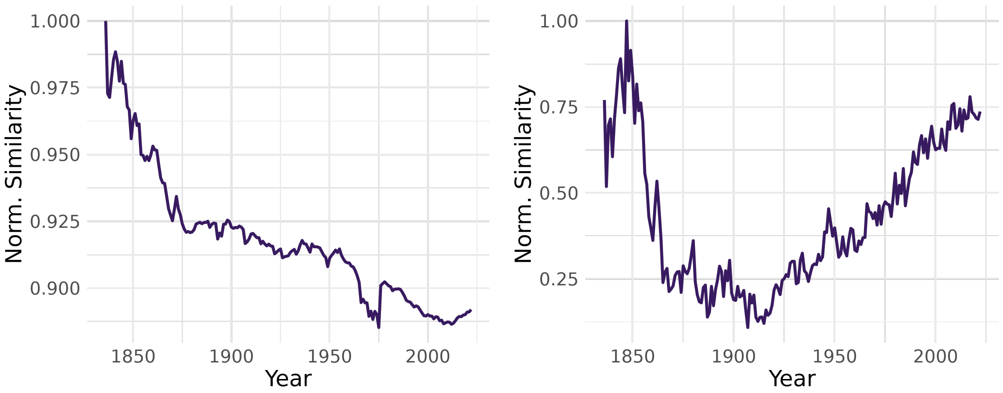
These findings highlight the importance of careful model selection and validation in economic research, particularly when analyzing long-term trends in technological change.

Our findings have significant implications for understanding long-run invention dynamics. In a companion paper (Ganguli et al. 2024), we develop a theory that links the decline in contemporaneous invention similarity to other notable trends in innovation economics, including the increasing burden of knowledge (Jones 2009), rising R&D spending (Hirschey, Skiba, and Wintoki 2012), declining R&D productivity (Bloom et al. 2020), and constant R&D spillovers (Lucking, Bloom, and Van Reenen 2019). This theory posits that as the burden of knowledge increases, inventors spread out over an expanding knowledge frontier, leading to a situation where ideas become harder to find (Bloom et al. 2020) due to weaker knowledge spillovers from more distant neighbors in idea space.

To further demonstrate that measurement depends on representation, we revisited the analysis of breakthrough patents by Kelly et al. 2021. (See details in Appendix E.) Overall, our analysis confirms the Kelly et al. (2021) finding that the rate of breakthrough inventions is higher today compared with prior decades. That said, the choice of representation still matters. Compared with the TF-BIDF representations used by Kelly et al. (2021), GTE-based measures suggest that the recent increase in breakthrough inventions is less unusual compared with historical patterns. Moreover, GTE-based measures appear to be more robust and less sensitive to decisions about how to process and residualize the data. In Appendix F, we further explore why deep learning models perform better compared with traditional methods such as TF-IDF.

4.1. Similarity Dynamics Within and Across Patent Office Technology Classifications

Figure 6 decomposes patent claim similarity into within-class and between-class components. We group patents according to their primary CPC "three-digit" class. Then, we calculate average pairwise similarity for patent claims in the same class and for claims in different classes, using our validated GTE-based representations. Within-class similarity



(a) Within classes

(b) Between classes

Figure 6: Average pairwise patent claims similarity by year according to GTE embeddings, decomposed into within and between class similarity.

declines throughout the sample, closely mirroring the overall trend observed in Figure 5. In contrast, between-class similarity shows more dramatic fluctuations, with a sharp initial decrease followed by a substantial increase since the early 20th century.

These findings highlight the limitations of classification-based approaches to measuring patent similarity (where, by definition, the patents in the same class are always similar). They also demonstrate a potential weakness of our patent class validation task, since, by construction, between-class similarity is ignored. Instead, text-based models can better capture dynamics both within and across technological fields, providing a more comprehensive picture of the evolving landscape of innovation.

4.2. Declines in Interferences

To further substantiate our GTE results on declining similarity, we construct a time series of interference rates spanning 150 years. This analysis provides an independent confirmation of our finding of declining patent similarity over time. While we used post-1998 interference cases to validate GTE-based measures of similarity, this section documents trends in interference rates from 1864 to 2014, offering additional out-of-sample empirical support for our main results.

We estimate the annual rate of interferences per issued patent, which approximates the probability that an issued patent was involved in an interference. Our analysis combines four distinct data sources spanning different time periods. Our earliest data come from the *Registers of Interferences* (1864–1901), which we purpose-digitized from the USPTO Records in the National Archives. We recorded 19,388 interference cases, averaging 504 interferences terminated annually. For the period 1950–1962, we rely on summary statistics from Di Simone, Gambell, and Gareau (1963), which report an average of 640 interferences terminated annually. Data from Calvert and Sofocleous (1982, 1986, 1989, 1992, 1995) show an average of 237 interferences terminated yearly from 1980–1994. Finally, our most recent data from Ganguli, Lin, and Reynolds (2020) indicate an average of 76 interferences terminated annually from 1998–2014[10].

Figure 7 illustrates a striking decline in the rate of interference over the 150-year period from 1864 to 2014. The average rate of interference fell from 2.71% in 1864–1901 to 1.43% in 1950–1962, then to 0.30% in 1980–1994, and finally to 0.05% in 1998–2014. Interestingly, the greater variability in the 19th century in the rate of interference, followed by a steady decline in interference since the middle 20th century, most closely resembles the trend in pairwise patent similarity illustrated by the weighted average of S-BERT, GTE, and PaECTER shown in Figure 5. This substantial and consistent decline in interference rates provides independent support for a long-run decline in invention similarity.

5. Conclusion

This paper presents a systematic approach to the construction and validation-based selection of text-based measures for economic concepts, focusing on patent similarity. Our analysis of NLP model selection and validation for economic research yields important recommendations for economists. We advocate testing multiple models, especially when some of the candidate NLP models are older or proprietary. Researchers should design domain-

[10] This figure likely slightly undercounts the actual number of interferences, as some were terminated before reaching the Board of Patent Interferences.

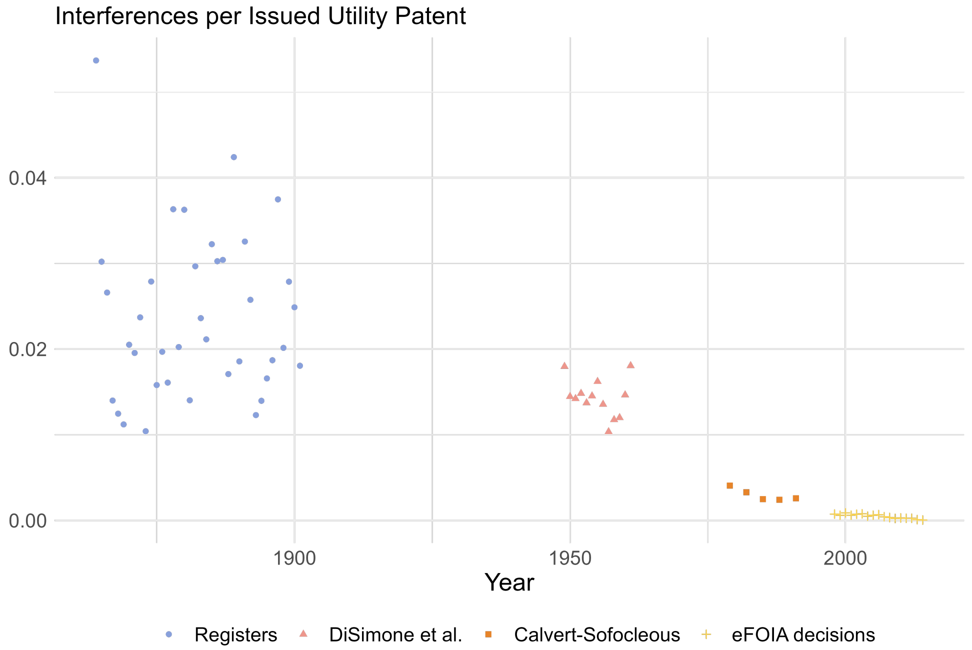


Figure 7: Interferences per issued patent

specific validation tasks tailored to their research questions to guide model selection. For studies measuring the similarity of patents, or similar concepts, GTE and PaECTER embeddings can serve as benchmark embeddings, as long as newer alternatives do not exist. As new generations of models emerge or when developing custom models, we encourage using our validation tasks alongside the research question-specific ones.

These recommendations stem from our finding of substantial performance differences between models on specific tasks and their divergent trends in key economic measures. This approach can allow economists to conduct more reliable and more robust analyses using methods from the dynamic and evolving field of NLP.

References

Akcigit, Ufuk, William R. Kerr, and Tom Nicholas (2017). *The Mechanics of Endogenous Innovation and Growth: Evidence from Historical US Patents*. Working Paper. Harvard University. URL: https://economics.harvard.edu/files/economics/files/kerr-william_mechanics_of_endogenous_innovation_patents_sbbi-2-3-17_0.pdf.

Arts, Sam, Bruno Cassiman, and Juan Carlos Gomez (2018). "Text Matching to Measure Patent Similarity". In: *Strategic Management Journal* 39.1, pp. 62–84. DOI: 10.1002/smj.2699.

Arts, Sam, Jianan Hou, and Juan Carlos Gomez (2021). “Natural Language Processing to Identify the Creation and Impact of New Technologies in Patent Text: Code, Data, and New Measures”. In: *Research Policy* 50.2, p. 104144. DOI: 10.1016/j.respol.2020.104144.

Ash, Elliott and Stephen Hansen (2023). “Text Algorithms in Economics”. In: *Annual Review of Economics* 15.1, pp. 659–688. DOI: 10.1146/annurev-economics-082222-074352.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S. Graff Zivin (Aug. 2019). “Does Science Advance One Funeral at a Time?” In: *American Economic Review* 109.8, pp. 2889–2920. DOI: 10.1257/aer.20161574.

Berkes, Enrico and Ruben Gaetani (Sept. 2020). “The Geography of Unconventional Innovation”. In: *The Economic Journal* 131.636, pp. 1466–1514. DOI: 10.1093/ej/ueaa111.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson (2024). “Synthetic Replacements for Human Survey Data? The Perils of Large Language Models”. In: *Political Analysis*, pp. 1–16. DOI: 10.1017/pan.2024.5.

Bloom, Nicholas, Charles I. Jones, John Van Reenen, and Michael Webb (2020). “Are Ideas Getting Harder to Find?” In: *American Economic Review* 110.4, pp. 1104–1144. DOI: 10.1257/aer.20180338.

Bochkay, Khrystyna, Stephen V. Brown, Andrew J. Leone, and Jennifer Wu Tucker (2023). “Textual Analysis in Accounting: What’s Next?” In: *Contemporary Accounting Research* 40.2, pp. 765–805. DOI: 10.1111/1911-3846.12825.

Calvert, Ian A. and Michael Sofocleous (1982). “Three Years of Interference Statistics”. In: *Journal of the Patent Office Society* 64, p. 699.

— (1986). “Interference Statistics for Fiscal Years 1983 to 1985”. In: *Journal of the Patent & Trademark Office Society* 68, p. 385.

— (1989). “Interference Statistics for Fiscal Years 1986 to 1988”. In: *Journal of the Patent & Trademark Office Society* 71, p. 399.

— (1992). “Interference Statistics for Fiscal Years 1989 to 1991”. In: *Journal of the Patent & Trademark Office Society* 74, p. 822.

— (1995). “Interference Statistics for Fiscal Years 1992 to 1994”. In: *Journal of the Patent & Trademark Office Society* 77, p. 417.

Carlson, David and Jacob M. Montgomery (2017). “A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts”. In: *American Political Science Review* 111.4, pp. 835–843. DOI: 10.1017/S0003055417000302.

Carmody, Sean (2023). *Ngramr: Retrieve and Plot Google n-Gram Data.* Manual.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil (Nov. 2018). “Universal Sentence Encoder for English”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Brussels, Belgium: Association for Computational Linguistics, pp. 169–174. DOI: 10.18653/v1/D18-2029.

Cheng, Zhaoqi, Dokyun Lee, and Prasanna Tambe (2022). *InnoVAE: Generative AI for Understanding Patents and Innovation.* Working Paper. SSRN. DOI: 10.2139/ssrn.3868599.

Clancy, Matthew S. (2018). “Inventing by Combining Pre-Existing Technologies: Patent Evidence on Learning and Fishing Out”. In: *Research Policy* 47.1, pp. 252–265. DOI: 10.1016/j.respol.2017.10.015.

Dasgupta, Partha and Eric Maskin (1987). “The Simple Economics of Research Portfolios”. In: *The Economic Journal* 97.387, pp. 581–595. DOI: 10.2307/2232925.

Davis, Jesse and Mark Goadrich (2006). “The Relationship between Precision-Recall and ROC Curves”. In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML ’06. New York, NY, USA: Association for Computing Machinery, pp. 233–240. DOI: 10.1145/1143844.1143874.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Di Simone, Daniel V., James B. Gambell, and Charles F. Gareau (1963). “Characteristics of Interference Practice”. In: *Journal of the Patent Office Society* 45, pp. 503–591.

Dominguez-Olmedo, Ricardo, Moritz Hardt, and Celestine Mendler-Dunner (2024). *Questioning the Survey Responses of Large Language Models*. arXiv: 2306.07951 `[cs.CL]`.

Feng, Sijie (July 2020). “The Proximity of Ideas: An Analysis of Patent Text Using Machine Learning”. In: *PLOS ONE* 15.7, pp. 1–19. DOI: 10.1371/journal.pone.0234880.

Fleming, Lee (2001). “Recombinant Uncertainty in Technological Search”. In: *Management Science* 47.1, pp. 117–132. DOI: 10.1287/mnsc.47.1.117.10671.

Ganguli, Ina, Jeffrey Lin, Vitaly Meursault, and Nicholas Reynolds (2024). *Declining Invention Similarity: Theory, Implications, and Evidence*. Work in Progress. Federal Reserve Bank of Philadelphia.

Ganguli, Ina, Jeffrey Lin, and Nicholas Reynolds (2020). “The Paper Trail of Knowledge Spillovers: Evidence from Patent Interferences”. In: *American Economic Journal: Applied Economics* 12.2, pp. 278–302. DOI: 10.1257/app.20180017.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). “Text as Data”. In: *Journal of Economic Literature* 57.3, pp. 535–574. DOI: 10.1257/jel.20181020.

Ghosh, Mainak, Sebastian Erhardt, Michael E. Rose, Erik Buunk, and Dietmar Harhoff (2024). *PaECTER: Patent-level Representation Learning using Citation-informed Transformers*. arXiv: 2402.19411 `[cs.IR]`.

Goli, Ali and Amandeep Singh (2024). “Frontiers: Can Large Language Models Capture Human Preferences?” In: *Marketing Science* 43.4, pp. 709–722. DOI: 10.1287/mksc.2023.0306.

Griliches, Zvi (1979). “Issues in Assessing the Contribution of Research and Development to Productivity Growth”. In: *The Bell Journal of Economics* 10.1, pp. 92–116. URL: http://www.jstor.org/stable/3003321.

Grimmer, J., M.E. Roberts, and B.M. Stewart (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Hirschey, Mark, Hilla Skiba, and M Babajide Wintoki (2012). “The Size, Concentration and Evolution of Corporate R&D Spending in US Firms from 1976 to 2010: Evidence and Implications”. In: *Journal of Corporate Finance* 18.3, pp. 496–518. DOI: 10.1016/j.jcorpfin.2012.02.002.

Hsieh, Cheng-Yu, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister (2023). *Distilling Step-*

by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. arXiv: 2305.02301 `[cs.CL]`.

Jaffe, Adam B. (1986). “Technological Opportunity and Spillovers of R&D: Evidence from Firms’ Patents, Profits, and Market Value”. In: *The American Economic Review* 76.5, pp. 984–1001. URL: http://www.jstor.org/stable/1816464.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson (Aug. 1993). “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations”. In: *The Quarterly Journal of Economics* 108.3, pp. 577–598. DOI: 10.2307/2118401.

Jones, Benjamin F. (2009). “The Burden of Knowledge and the “Death of the Renaissance Man”: Is Innovation Getting Harder?” In: *The Review of Economic Studies* 76.1, pp. 283–317. DOI: 10.1111/j.1467-937X.2008.00531.x.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy (Sept. 2021). “Measuring Technological Innovation over the Long Run”. In: *American Economic Review: Insights* 3.3, pp. 303–20. DOI: 10.1257/aeri.20190499.

Kusupati, Aditya, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi (2024). *Matryoshka Representation Learning.* arXiv: 2205.13147 `[cs.LG]`.

Le, Quoc and Tomas Mikolov (22–24 Jun 2014). “Distributed Representations of Sentences and Documents”. In: *Proceedings of the 31st International Conference on Machine Learning.* Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, pp. 1188–1196. URL: https://proceedings.mlr.press/v32/le14.html.

Lee, Jieh-Sheng and Jieh Hsiang (2019). *PatentBERT: Patent Classification with Fine-Tuning a Pre-Trained BERT Model.* arXiv: 1906.02124 `[cs.CL]`.

Li, Zehan, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang (2023). *Towards General Text Embeddings with Multi-stage Contrastive Learning.* arXiv: 2308.03281 `[cs.CL]`.

Lucking, Brian, Nicholas Bloom, and John Van Reenen (2019). “Have R&D Spillovers Declined in the 21st Century?” In: *Fiscal Studies* 40.4, pp. 561–590. DOI: 10.1111/1475-5890.12195.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). *Efficient Estimation of Word Representations in Vector Space.* arXiv: 1301.3781 `[cs.CL]`.

Miller, George A. (1992). “WordNet: A Lexical Database for English”. In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992.*

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn (2017). “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict”. In: *Political Analysis* 16.4, pp. 372–403. DOI: 10.1093/pan/mpn018.

Murata, Yasusada, Ryo Nakajima, Ryosuke Okamoto, and Ryuichi Tamura (Dec. 2014). “Localized Knowledge Spillovers and Patent Citations: A Distance-Based Approach”. In: *The Review of Economics and Statistics* 96.5, pp. 967–985. DOI: 10.1162/REST_a_00422.

Olsson, Ola (2000). “Knowledge as a Set in Idea Space: An Epistemological View on Growth”. In: *Journal of Economic Growth* 5, pp. 253–275. DOI: 10.1023/A:1009829601155.

OpenAI (2024). *text-embedding-3-large.* URL: https://openai.com/index/new-embedding-models-and-api-updates/ (visited on 06/02/2024).

Park, Michael, Erin Leahey, and Russell J. Funk (2023). "Papers and Patents are Becoming Less Disruptive Over Time". In: *Nature* 613.7942, pp. 138–144. DOI: 10.1038/s41586-022-05543-x.

Reimers, Nils and Iryna Gurevych (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* arXiv: 1908.10084 [cs.CL].

Schnoebelen, Tyler, Julia Silge, and Alex Hayes (2022). *Tidylo: Weighted Tidy Log Odds Ratio.* Manual.

Smith, Noah A. (May 2020). "Contextual Word Representations: Putting Words into Computers". In: *Communications of the ACM* 63.6, pp. 66–74. DOI: 10.1145/3347145.

Sparck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and its Application in Retrieval". In: *Journal of Documentation* 28.1, pp. 11–21. DOI: 10.1108/eb026526.

Thompson, Peter and Melanie Fox-Kean (Mar. 2005). "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment". In: *American Economic Review* 95.1, pp. 450–460. DOI: 10.1257/0002828053828509.

U.S. Patent and Trademark Office (Feb. 2023). *Data Download Tables.* PatentsView. URL: https://patentsview.org/download/data-download-tables.

Wang, Jian, Reinhilde Veugelers, and Paula Stephan (2017). "Bias Against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators". In: *Research Policy* 46.8, pp. 1416–1436. DOI: 10.1016/j.respol.2017.06.006.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2024). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems.* NIPS '22. New Orleans, LA, USA: Curran Associates Inc.

Appendix A. Visualization of Embedding Spaces

This appendix describes the process we followed to generate the visualizations discussed in Section 2.

The raw data are obtained using the same sampling strategy outlined in the class and period validation section (3.3). This strategy involves sampling patents from specified classes as categorized by the USPTO, across distinct 25-year periods ranging from 1850 to 2023.

We then plot 2-dimensional projections of the embedding spaces, where individual patents are marked with color according to their respective class or period. This visualization technique provides a geometrically intuitive perspective of the innovation space. It also lays a visual foundation for comparing the efficacy of different embedding techniques like S-BERT and TF-IDF.

A.1. Methodology

The primary method we employ for visualization is dimensionality reduction through the Uniform Manifold Approximation and Projection (UMAP) technique. UMAP is noted for its ability to preserve both global and local structures during reduction, making it, roughly speaking, a non-linear variant of Principal Component Analysis (PCA).

To speed up the computation, we conduct the initial dimension reduction using PCA, which reduces the dimensionality of the S-BERT and TF-IDF representations to 50. Subsequently, UMAP is applied to these reduced representations. This two-step process harnesses the computational efficiency of PCA while benefiting from the geometric qualities of UMAP.

We manually tuned UMAP hyperparameters to achieve a more clustered representation that looked more like an “archipelago” than a singular “continent.” This tuning aids in better visual separation among clusters within the innovation space.

A.2. Plotting

One of the challenges encountered during visualization was the overlapping of data points, especially in dense clusters. To mitigate this, a jittering technique was employed which disperses each point slightly within its local neighborhood to reduce overlap, hence enhancing the visibility of individual clusters. The jittering results in a boxier scatter plot, which is a compromise for better clarity.

The plots (refer to Figure 3) primarily serve as illustrative tools, providing a more tangible notion of the idea space. We use color coding to denote different patent classes and 25-year periods in both S-BERT and TF-IDF projections. Despite the inherent distortions, some observations could hint at underlying structural differences between the representations.

At first glance, it's clear how the representations reflect the class and period structure. S-BERT representations show clearer class boundaries compared to TF-IDF representations, suggesting that patent clustering is closer to the class structure. On the other hand, TF-IDF periods seem less mixed compared to S-BERT periods, although this difference is more subtle. These visual patterns match the results we discussed in Section 3.3, where we evaluated how well the representations classify patent pairs into the same class and same period categories. This consistency between visual observations and analytical findings is encouraging.

It is harder to draw conclusions from the general layout because of the distortions inherent in the projection project. However, some observations stand out. For example, TF-IDF has more "dust" compared to S-BERT, which has more of an "empty space." Also, the extended x and y tails in TF-IDF, hidden due to winsorizing, hint at a possible trend where variability in expressing similar ideas with different words pushes these representations farther from the core.

Lastly, we explored the clusters qualitatively using an interactive tool. While we don't expect every aspect of patent positions to be interpretable, some interesting observations came to light. For instance, in Panel A of Figure 3, a blue square around (-5, 0), representing the electricity class, contains many semiconductor patents. This square sits between the light blue square on its left representing materials science patents (Chemistry and Metallurgy) and a more general blue electricity patent cluster on its right. Although such observations are anecdotal, they help build trust in the model, especially when supported by more rigorous analyzes. Such qualitative insights, alongside quantitative evaluations, enrich our understanding of the embedding spaces and their ability to capture the complex nature of innovation.

The visualizations provide insight into how different representations can result in meaningfully different similarity measures, highlighting the importance of making grounded choices in representations when studying innovation.

Appendix B. Instructions for the Non-Expert Human Judgement Task

You will be comparing the similarity of two pairs of patents to determine which pair is more similar to each other. Read through each pair carefully. Then compare the key aspects of each pair of patents, including the following (feel free to use "scratchpad" column to take notes, but that's not necessary):

- The general field or domain the patents relate to
- The specific problem each patent is trying to solve
- The key components of the solution each patent proposes
- Any other major similarities or differences between the patents in each pair

Based on analyzing these factors, assess the overall similarity of the patents in each pair. Determine which pair of patents you think is more similar to each other.

If you don't understand the text enough to assess the above, feel free to google to understand meaning of unfamiliar words or concepts. But try to avoid reading parts of the patent that are outside the snippet (for example, using google patents).

In the "anno_more_similar_1_or_2_or_0" column, put only the number of the pair (1 or 2) that you judge to be more similar. If you are unsure about which is better, put 0 there.

To make it easier to annotate in excel, adjust the width of the text_pair_1 and text_pair_2 columns and click the "wrap text" button.

Example

Pair 1

IMPROVEMENT: Improvements in Train-Binding Harvesters and Mowers

CLAIMS: The combination of the wedge-shaped platform 15, secondary platform 47, door 35, carriage 46, pivoted reciprocating extension-rake 41, chain 64, and the pulleys 60, these members constructed and operating substantially as and for the purposes herein specified. 2. In combination with the main frame B, the detachable arm 63, having the binder mounted thereon, substantially as and for the purposes herein specified. 3. The combination of the arm 63, eyebolt H

IMPROVEMENT: Improvement in Incandescent Electric Lamps

CLAIMS: 1. The combination, with the incandescing conductor of an electric lamp and the key for controlling the circuit thereof, of an adjustable resistance located within the base of the lamp and cut in or out of the circuit in any desired proportion by the key, so that the lamp may be used at any desired power less than its normal capacity, substantially as set forth. 2. A carbon resistance made substantially as described, and provided with a series of metallic contacts, in combination with a keyhavin

Pair 2

IMPROVEMENT: Improvements in Wire Fences

CLAIMS: 1. In a wire fence a vertical brace or tie having two legs, a horizontal wire having horizontal bends disposed between said two legs, a plate having at each end a pair of horizontally-extending prongs or fingers with spaces between the same, and a connecting-portion d, the back side of said connecting portion being disposed within said horizontal bend, the horizontal wire passing throughsaid spaces, and the front side of said prongs or fingers being clamped around said legs, substantially as and

---

IMPROVEMENT: improvements in hitches

CLAIMS: 1. A trailer hitch comprising a bar, means for rigidly securing said bar vertically on a vehicle bumper, a loop loosely mounted on the lower portion of the bar, said bar having an opening in its upper portion, a bracket removably mounted on the bar, said bracket including a second vertical bar engaged at its lower end in the loop, a forwardly projecting rigid pin on the upper end portion of the second-named bar engaged in the opening of the first-named bar, and a ball rigidly mounted on the seco

Possible Reasoning

Pair 1

The first patent relates to harvesting/mowing equipment, while the second is about incandescent electric lamps. Very different domains. The first patent aims to improve the binding mechanism on a harvester/mower. The second allows adjusting the power level of an electric lamp. The first uses components like platforms, doors, carriages, rakes and pulleys in its solution. The second uses an adjustable resistance, metallic contacts, and a key. The two patents are solving very different problems in unrelated fields using dissimilar components and mechanisms.

Pair 2

Both patents relate to connection/attachment mechanisms, the first for wire fences and the second for trailer hitches. More related domains than Pair 1. The first patent aims to provide an improved way to brace and tie together wires in a fence. The second provides an improved trailer hitch mechanism. Both make use of bars, loops, brackets, and engagement of components to create their attachment solutions. While the specific applications differ, both patents essentially aim to solve connection/attachment problems using some similar components like bars, loops and brackets.

Conclusion

The patents in Pair 2 seem to have more in common in terms of their general domain, the type of problem they are solving, and some of the key components used, compared to the very different patents in Pair 1. Pair 2 appears more similar overall.

More difficult pairs

Many patent pairs will be more tenuously connected than others, even when patent pairs seem dissimilar try to think about how they might be trying to solve similar problems or using similar technology.

Here are some examples of dissimilar things that might still be the more similar patent pair in a row:

- Sewing Machines and Closet Hanging Rods are very different technologies, but are both related to clothing/home goods
- Flutes and Tube Sprinklers are very different technologies, but are both tubes with holes in them

Often the patents themselves are small but complicated improvements in technologies you are already familiar with. Even if it is hard to understand the improvement, try to think about how you can connect the technologies in each pair of patents (even tenuously), keeping in mind again:

- The general field or domain the patents relate to
- The specific problem each patent is trying to solve
- The key components of the solution each patent proposes
- Any other major similarities or differences between the patents in each pair

Appendix C. LLM prompt for patent similarity assessment

You will be comparing the similarity of two pairs of patents to determine which pair is more similar to each other.

Here is the first pair of patents:

<pair1> {PAIR1} </pair1>

And here is the second pair of patents:

<pair2> {PAIR2} </pair2>

Read through each pair carefully. Then, in a <scratchpad>, compare the key aspects of each pair of patents, including:

- The general field or domain the patents relate to
- The specific problem each patent is trying to solve
- The key components of the solution each patent proposes
- Any other major similarities or differences between the patents in each pair

Based on analyzing these factors, assess the overall similarity of the patents in each pair. Determine which pair of patents you think is more similar to each other.

In an <answer> tag, output only the number of the pair (1 or 2) that you judge to be more similar. If you are unsure about which is better, output 0. Do not include any other text or explanation. Close the answer tag with </answer>. You shouldn't have a bias towards answering either 1 or 2, the answer should be only evidence-based. If you don't have a reasonable level of confidence, it's better to output a 0.

Appendix D. LLMs for patent similarity assessment

Human annotation, while valuable, can be costly and challenging, especially when comparing technical documents like patents. To address these limitations and provide a scalable approach to our validation setup, we explore the use of Large Language Models (LLMs) for annotation tasks. While this approach introduces its own set of limitations, it offers potential benefits in terms of scalability and cost-effectiveness.

It's crucial to note that we do not view this as an exercise in using LLMs as survey respondents. Recent research across various disciplines has shown that LLMs often do not reflect human judgments in statistically accurate ways (Bisbee et al. 2024; Dominguez-Olmedo, Hardt, and Mendler-Dunner 2024; Goli and Singh 2024). In light of these findings, we cannot assume that LLMs have the same underlying concept of idea similarity as humans. Rather, we explore whether this is the case to a useful degree by comparing LLM results with human annotations, allowing us to assess the potential utility of LLMs in this context.

Our approach is conceptually similar to the distillation techniques used in LLM research, where outputs from larger models are used to improve or evaluate smaller models (Hsieh et al. 2023). In our case, we're not improving capabilities but testing them, using larger LLMs to evaluate the performance of smaller embedding models that share many elements with LLMs.

We employed two state-of-the-art (as of July 2024) language models, Claude 3.5 Sonnet (`claude-3-5-sonnet-20240620`) and GPT-4o (`gpt-4o-2024-05-13`), to perform the same similarity judgment task as human annotators. We provided the models with identical patent pair comparisons, using carefully designed prompts based on the human annotator instructions (see Appendix Appendix C for the full prompt).

Our prompts were structured to mirror the human annotation process closely, incorporating a "chain of thought" (CoT) approach (Wei et al. 2024). The LLMs were instructed to analyze key aspects of each patent pair in a "scratchpad" section before making a final judgment, mirroring the format of human annotations.

Table D.5: LLM Agreement with Embedding-Based Similarity Rankings

	PaECTER		GTE		S-BERT		TF-IDF	
	Claude	GPT	Claude	GPT	Claude	GPT	Claude	GPT
(Intercept)	0.14	0.17	0.08	0.14	0.16	0.11	0.16	0.31^{*}
	(0.10)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.13)
Claude=1	0.52^{***}		0.60^{***}		0.58^{***}		0.54^{***}	
	(0.11)		(0.10)		(0.10)		(0.10)	
GPT4o=1		0.57^{***}		0.58^{***}		0.71^{***}		0.35^{*}
		(0.12)		(0.11)		(0.10)		(0.15)
R^2	0.19	0.26	0.28	0.28	0.28	0.43	0.23	0.08
Num. obs.	92	72	91	76	90	67	94	68

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Notes: Regression results showing the agreement between Claude 3.5 Sonnet (`claude-3-5-sonnet-20240620`), GPT-4o (`gpt-4o-2024-05-13`), and the relative similarity rankings of patent pairs according to different patent text representations.

D.0.1. LLM-based Results

To analyze the agreement between LLM judgments and embedding-based similarity rankings, we use the following regression setup:

$$I[Sim(2) > Sim(1)]^{Emb} = \beta_0^{LLM} + \beta_1^{LLM} I[Response = 2]^{LLM} + \epsilon \tag{D.1}$$

where LLM $\in$ {Claude, GPT} and Emb $\in$ {PaECTER, GTE, BERT, TF-IDF}. The coefficient β_1 represents the increase in the probability that the embedding indicates pair 2 is more similar when the LLM chooses pair 2. Higher β_1 suggests a stronger LLM-embedding agreement.

Each LLM produced outputs for 100 comparisons. However, the number of observations in our regressions is lower, reflecting the removal of cases where the LLM responded with 0 (indicating it couldn't decide). This ensures that our analysis focuses on clear judgments made by the LLMs.

We present the results of our LLM-based regressions in Table D.5. The ranking of representations differs between the two LLMs and from our human annotation results. For Claude, the ranking is GTE >S-BERT >PaECTER >TF-IDF, while for GPT-4o, it's S-

BERT >GTE >PaECTER >TF-IDF. This contrasts with the human annotation ranking of GTE >BERT >PaECTER >TF-IDF. Despite these differences, both LLMs consistently show that newer embedding models (PaECTER, GTE, S-BERT) outperform the traditional TF-IDF approach, aligning with our human annotation findings in this crucial aspect.

The variability in results between human annotators and different LLMs underscores the potential limitations of using LLMs as proxies for human judgment in this context. However, the consistent underperformance of TF-IDF across all evaluation methods (human and LLM) provides strong evidence for the superiority of newer embedding techniques in capturing patent similarity. This suggests a potential use for LLMs as a cost-effective way to test the validation tasks before deploying them to human annotators, streamlining the overall validation process.

Appendix E. Revisiting Breakthrough Patents with Validated Patent Representations

This section demonstrates how the choice of patent representation can significantly impact economic measurement by revisiting the analysis of "breakthrough" patents in Kelly et al. (2021).

Our investigation not only provides a robustness check on their findings but also underscores the critical importance of model selection in economic research. Kelly et al. (2021) employ a backward-looking variant of TF-IDF, which they term TF-BIDF, to identify breakthrough patents. These are defined as patents dissimilar to past inventions but highly similar to future ones. Their method involves creating TF-BIDF representations of patent texts, residualizing this measure on year fixed effects, identifying the top 10% of patents in the residualized measure, and plotting the rate of breakthrough patents normalized by total US population.

Our analysis explores the sensitivity of their results along three key dimensions: (i) using GTE versus TF-BIDF for representations, (ii) residualizing the breakthrough measure on year fixed effects, and (iii) normalizing the rate of breakthrough patents by total US population. This comprehensive approach allows us to isolate the impact of representation choice while also examining other methodological decisions.

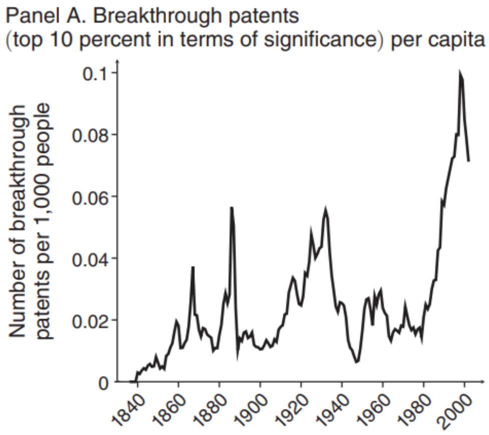
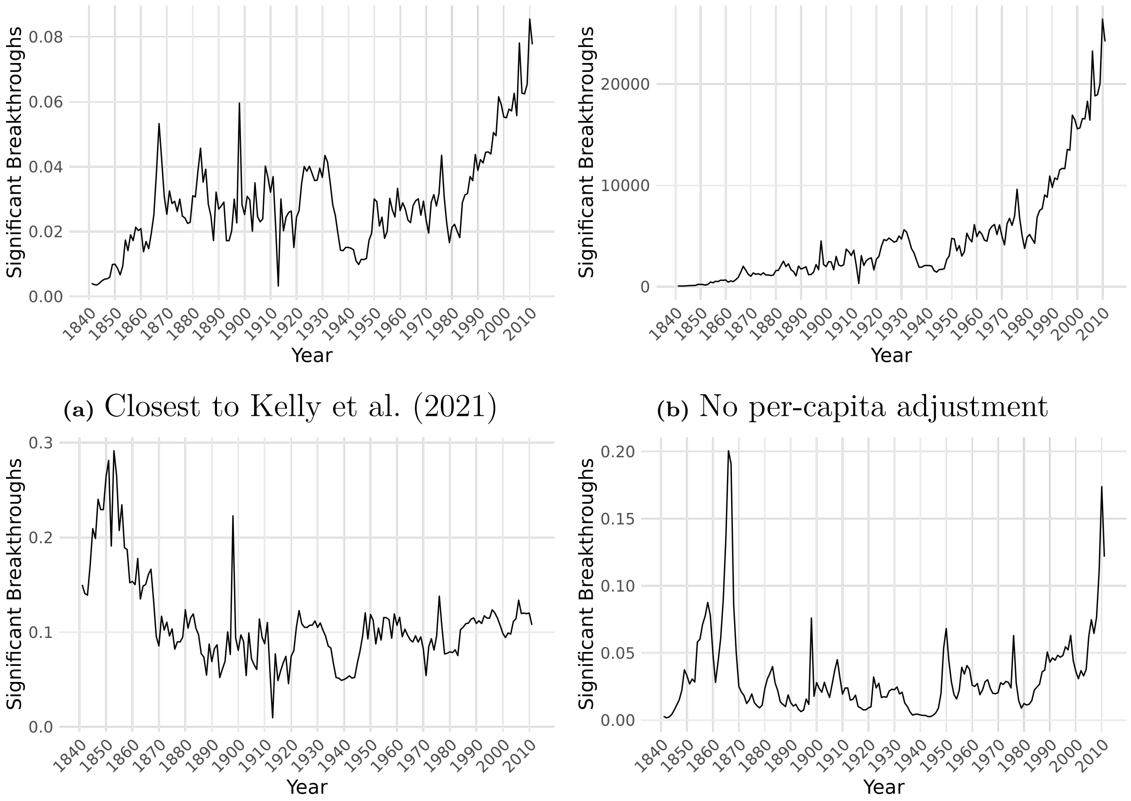


Figure E.8: Reproduction of Kelly et al. (2021), Figure 4, Panel A



(a) Closest to Kelly et al. (2021)

(b) No per-capita adjustment

(c) Adjusting for number of patents instead of population

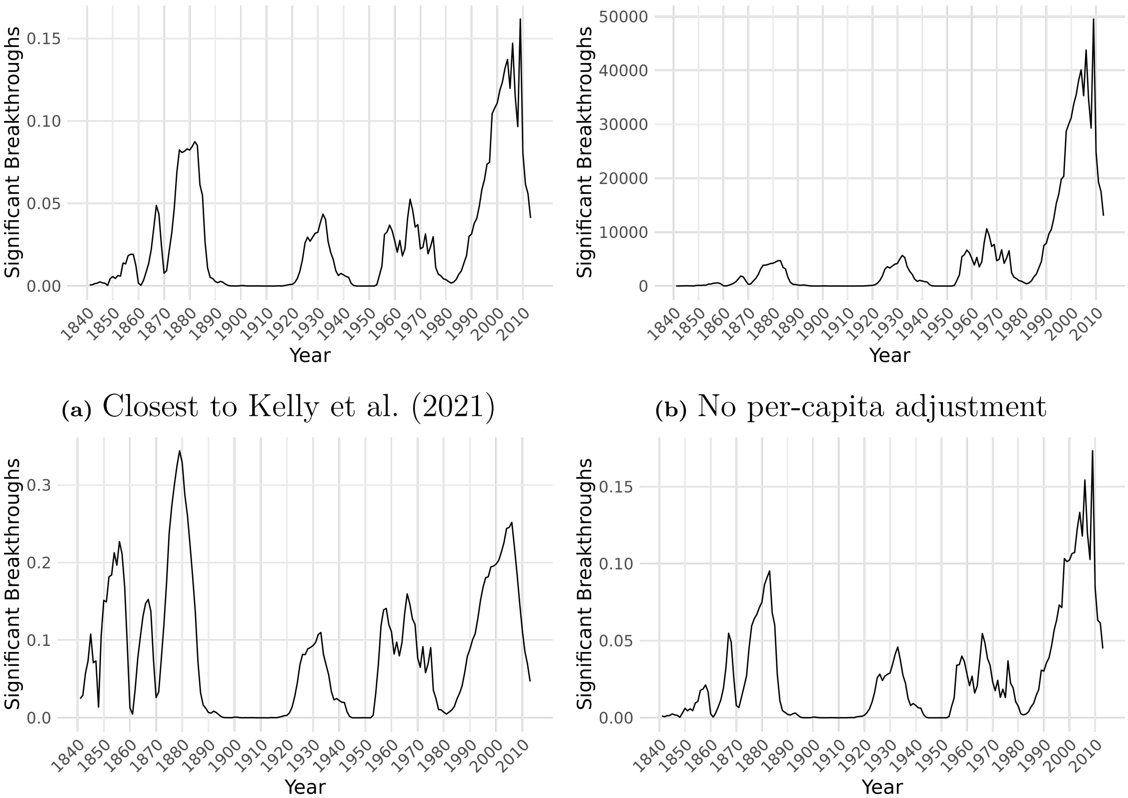
(d) No year fixed effects

Figure E.9: Replication and robustness of Kelly et al. (2021) using TF-BIDF representations

Figure E.8 reproduces the key result from Kelly et al. (2021), showing the rate of breakthrough patents per capita over time. Our replication, shown in Figure E.9, Panel A, closely mirrors their findings despite some methodological differences.[11] The qualitative dynamics are very similar, with fluctuations in the rate (per US population) of breakthrough patents, followed by a sharp increase starting around 1980.

Examining the robustness of the Kelly et al. (2021) results reveals several insights: The choice of normalization significantly affects the interpretation of results. While the per-capita measure shows a sharp increase in breakthrough patents since 1980 (Figure E.9, Panel A),

[11] The primary distinctions are in the source corpus and IDF computation. We use the ProQuest database of patent claims, whereas Kelly et al. (2021) employed Google Patents digitized text. Additionally, for computational efficiency, we simplify the calculation of backward-looking IDFs to the prior five calendar years, while Kelly et al. (2021) computed a backward IDF for each patent up to five years prior to its issue date.



(a) Closest to Kelly et al. (2021)

(b) No per-capita adjustment

(c) Adjusting for number of patents instead of population

(d) No year fixed effects

Figure E.10: Replication and robustness of Kelly et al. (2021) using GTE representations

normalizing by the total number of patents issued in that year reveals that the peak rate of breakthroughs occurred before 1870 (Figure E.9, Panel C). Additionally, residualizing on year fixed effects alters the historical pattern of breakthrough patents (Panel D), producing two peaks. These comparisons highlight the sensitivity of the results to methodological choices.

Figure E.10 presents the same analysis using GTE representations instead of TF-BIDF. This comparison yields several important observations.

GTE-based measures confirm the general trend of increased breakthrough patent rates in recent decades, lending support to the Kelly et al. (2021) findings. However, GTE representations suggest that the recent increase in breakthrough inventions is less exceptional when compared to historical patterns. GTE identifies similar, albeit more modest, booms in breakthrough patents in the 1870s, 1930s, and 1960s.

GTE-based measures also appear more robust to methodological choices. For instance, the decision to residualize on year fixed effects has less impact on the overall trends when using GTE (Figure E.10, Panel D) compared to TF-BIDF. This enhanced stability suggests that GTE may provide a more reliable foundation for analyzing patent data across different methodological approaches, potentially offering a more consistent view of technological change over time.

Our analysis corroborates Kelly et al. (2021)'s finding of elevated breakthrough invention rates in recent decades, while demonstrating the crucial role of representation choice in economic measurement. GTE-based measures, unlike TF-BIDF, reveal that the recent surge in breakthrough inventions is less exceptional in a broader historical context. Crucially, GTE representations show greater robustness and reduced sensitivity to data processing and residualization decisions. These findings underscore the importance of validation-based model selection in research on technological progress.

Appendix F. Why are Deep Learning Models Better? An In-Depth Look at Why S-BERT is Better than TF-IDF.

In this section, we explore the performance differences between S-BERT and TF-IDF. First, we compare a 21st-century bicycle patent and a 19th-century velocipede patent to illustrate S-BERT's ability to identify semantic similarities. Second, we examine unigram frequencies in the Google Books Ngram database. Unigrams characteristic of patent pairs with high TF-IDF similarity overweight period-specific language similarities, rather than similarity of ideas represented by the patents. We then present details of the characteristic unigram methodology, an additional Google Books Ngram analysis, and a synonym-based analysis that further highlights S-BERT's ability to capture semantic similarity.

F.1. Example: Bicycle versus Velocipede

Figure F.11 shows a bicycle patent from the 21st century and a velocipede patent from the 19th century. Despite these patents originating from different time periods and employing distinct terminologies, S-BERT successfully identifies them as similar, positioning them in the 87th percentile of similarity. At the same time, the similarity according to TF-IDF is 0. This example illustrates the S-BERT's ability to capture semantic nuances and contextual similarities despite changes in language.

Both patents introduce improvements in the design or function of two-wheeled vehicles. A velocipede is an archaic term for a type of bicycle. Although Patent 1 focuses on the "front frame for a bicycle" while Patent 2 is more broadly about an "improved velocipede," they both involve common mechanical features such as tubes, frames, and axles. However, the patents do not share many common terms. Patent 1 talks about "front frame," "inner tubes," "upper tube," while Patent 2 mentions "friction-clutch," "spurs," "arms," etc.

S-BERT takes into account not just specific words, but also the context in which these words appear. Words with similar meaning that frequently appear in similar contexts will be assigned similar S-BERT vectors. Thus, S-BERT representations reflect that both patents

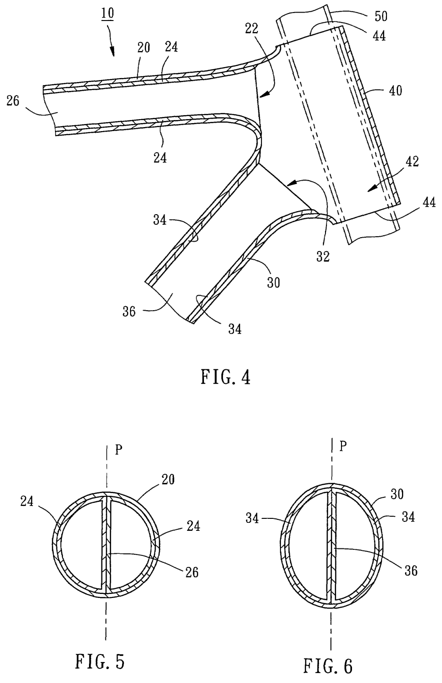
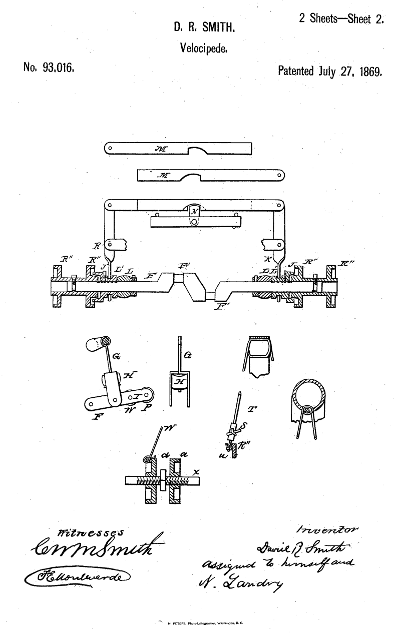
Patent 1: US7562890B2 (2009)	**Patent 2: US93016A (1869)**
Front frame for a bicycle. 1. A front frame for a bicycle, comprising: two first inner tubes abutted together; two second inner tubes abutted together; an upper tube of cured multiple layers of fiber reinforced rein material wound around the two first inner tubes so that there is no crack between the upper tube and . . .	IMPROVED VELOCIPEDE. In the velocipede as constructed, and in combination therewith, the friction-clutch, spurs, arms, cross-bar, cam, guide-wheel, with hollow rim and axle, arranged and operated substantially as described. In witness whereof, I have hereunto set my hand and seal.
	

Figure F.11: A conceptually similar pair of patents from different time periods

Notes: Velocipede is a type of bicycle. The text is truncated to the title and the beginning of the claims section of the patents. Typos due to OCR were fixed for this illustrative example. According to S-BERT, these patents are in the 87th percentile of similarity, whereas according to TF-IDF, the similarity is 0.

are about two-wheeled vehicles, even if they use different terms. S-BERT is trained on a diverse dataset, which includes technical language. It can therefore encode terms like “frame,” “tubes,” and “axle” as related in general, even if they appear in different contexts.

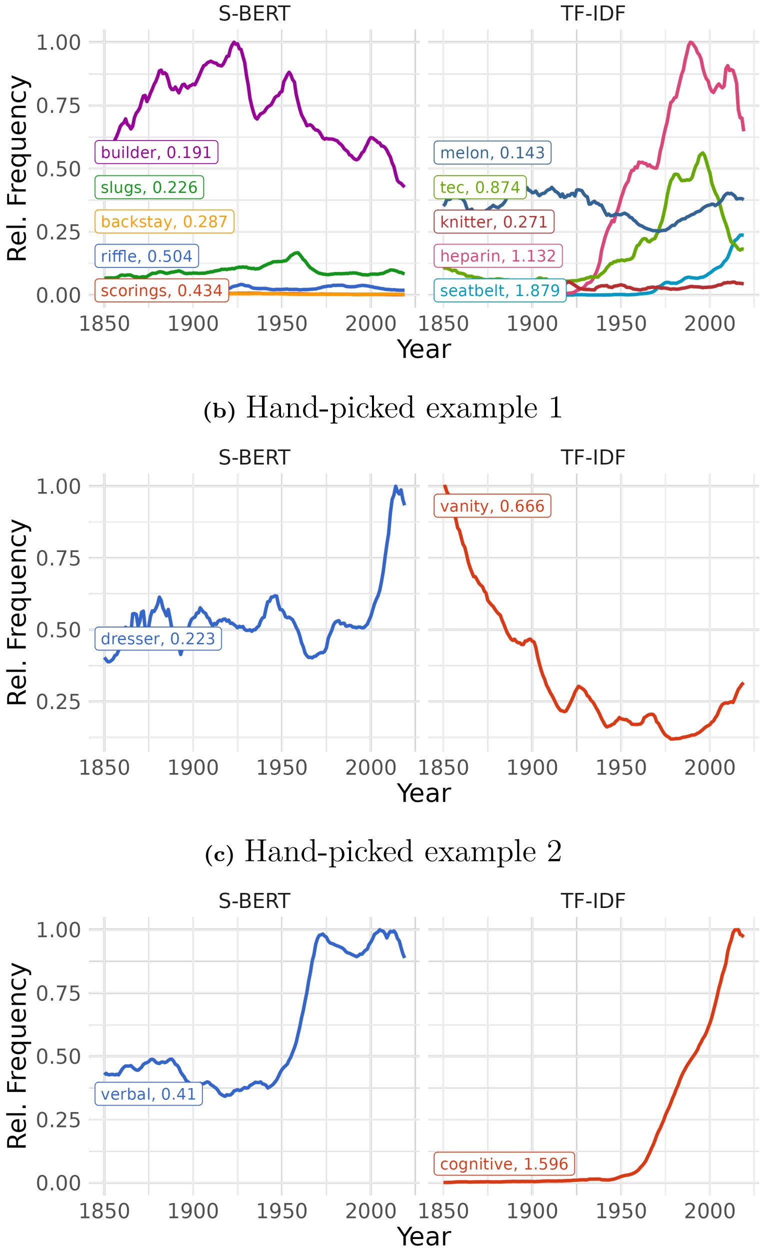
TF-IDF is a simpler bag-of-words model that does not capture meaning in the same way (see Smith 2020). It considers only the frequency of individual words in each document and in the corpus as a whole. TF-IDF treats distinct terms such as “bicycle” and “velocipede” as unrelated concepts. In sum, S-BERT is able to better capture the semantic and contextual similarities between these two patents that describe similar inventions but do not share a common vocabulary.

F.2. TF-IDF Overweights Period-Specific Words versus Universal Synonyms

The bicycle/velocipede example suggests that TF-IDF overweights period-specific terms like velocipede, leading it to assign low similarity to pairs that might describe the same idea with different terms. Here we extend that analysis. We hypothesize that terms used in patent pairs assigned high similarity by TF-IDF should have a higher variance of usage over time. These period-specific terms might be archaic or modern, or they may have irregular fluctuations in usage.

Figure F.12 presents some illustrative examples of unigram frequencies over time. Among the top-five most characteristic unigrams, TF-IDF unigrams are more volatile, which indicates more time-specific word usage.

We further hand-picked examples of conceptually-similar words in panel (b). “Dresser,” characteristic of S-BERT similar pairs, exhibits moderate use with little variation until the 2000s. In contrast, “vanity,” characteristic of TF-IDF similar pairs, exhibits more volatility, steadily dropping in usage throughout the period between 1850 and 1970, followed by a small rise. Another example is shown in panel (c). “Verbal” and “cognitive” both increase after 1950. But the increase is more dramatic for “cognitive,” and therefore this term characteristic of TF-IDF similar pairs has a larger coefficient of variation.



(a) Top-5 characteristic unigrams for each representation

(b) Hand-picked example 1

(c) Hand-picked example 2

Figure F.12: Frequency of characteristic unigrams of the pairs of patents classified as similar by S-BERT and TF-IDF

Notes: The plot is based on the Google Ngram Corpus (1850–2019). Frequency is normalized to the largest frequency on each plot. The number after the unigram label is the coefficient of variation, defined as the standard deviation divided by the mean. The characteristic unigrams are computed using the Monroe, Colaresi, and Quinn 2017 algorithm.

F.3. Google Ngrams Analysis

To gain insights into the time-specific nature of the words that TF-IDF focuses on, we turn to examining the tokens characteristic of patent pairs located closely in the TF-IDF space through the lens of Google Ngrams data. We identify characteristic tokens that differentiate patent pairs based on their similarity scores. Our analysis categorizes patent pairs into three groups: (i) those identified as similar by both S-BERT and TF-IDF, (ii) those recognized as similar only by S-BERT, and (iii) those recognized as similar only by TF-IDF. We exclude pairs with mutual agreement between models and determine characteristic unigrams for the latter two categories.

This analysis demonstrates that the unigrams characteristic of patent pairs with high TF-IDF similarity tend be more heavily used in specific time periods compared to the S-BERT unigrams, which can explain the outperformance of TF-IDF in the period classification task.

The Google Books Ngrams dataset is a collection of word frequencies derived from the Google Books corpus,[12] which contains a vast array of books published over several centuries. This dataset enables the analysis of the usage patterns of words and phrases over time, providing a valuable resource for studying the evolution of language.

In NLP, characteristic tokens or words are specific lexical features that are highly indicative of a particular category, topic, or sentiment. These tokens serve as markers that can help in classifying or differentiating texts based on the target concept of interest, such as the party alignment of a political speech, or, in our case, whether a patent pair is deemed similar by S-BERT or TF-IDF. We use the Monroe, Colaresi, and Quinn (2017) method implemented in the Schnoebelen, Silge, and Hayes (2022) R library to systematically identify characteristic words. The method employs Bayesian shrinkage and regularization techniques to select and evaluate the relative importance of words that capture the target semantic concept.

[12] Specifically, we use the "English 2019" corpus accessed using *ngramr* library in R programming language (Carmody 2023).

Finding characteristic words requires a corpus of text split according to a categorical variable, which we obtain the following way. From the corpus of 11,200 patents used in the class and period validation task, we selected pairs that were in the top quartile of similarity scores according to S-BERT, TF-IDF, or both. We then categorized these pairs into three classes:

1. The representations agree
2. S-BERT identifies as similar, but TF-IDF does not *S-BERT Yes* category
3. TF-IDF identifies as similar, but S-BERT does not *TF-IDF Yes* category

We discard the pairs where both representations agreed and use the rest of the pairs as the input to Monroe, Colaresi, and Quinn (2017) algorithm to find unigrams most characteristic of S-BERT and TF-IDF similarity. The output of the algorithm is the list of characteristic words for the categories *S-BERT Yes* and *TF-IDF Yes* along with the weighted log-odds that quantify the extent to which a unigram is more likely to appear in one category of patent pairs compared to the other.

Once the characteristic unigrams are obtained, we analyze their frequency from 1850 to the present using the Google Books Ngram corpus. For each unigram, we calculate the mean and standard deviation of its frequency over time. To obtain a measure of variation that is comparable between different unigrams we compute the coefficient of variation, defined as the standard deviation divided by the mean.

Figure F.13 demonstrates the average coefficient of variation for *S-BERT Yes* and *TF-IDF Yes* characteristic unigrams. The difference is large, especially for the unigrams with the highest weighted log-odds. For the top 100 unigrams, the S-BERT coefficient of variation is 0.7 compared to 1.2 for TF-IDF (which means that the average standard deviation is 70% and 120% of the mean, respectively). As we increase the number of unigrams we include in the computation, the difference becomes smaller, but is always large: for all unigrams, the S-BERT coefficient of variation is 0.74 compared to 0.95 for TF-IDF.

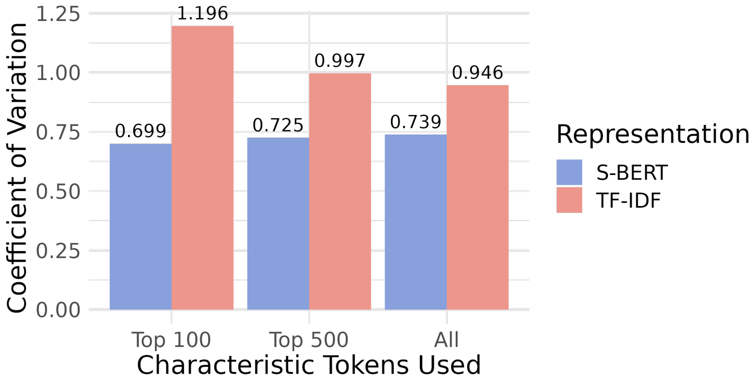


Figure F.13: Average over-time coefficient of variation of the frequency of characteristic unigrams of the pairs of patents classified as similar by S-BERT and TF-IDF

Notes: The unigram frequency information is from the Google Ngram Corpus (1850–2019). The coefficient of variation is defined as the standard deviation divided by the mean. The characteristic unigrams are computed using the Monroe, Colaresi, and Quinn 2017 algorithm.

The higher coefficient of variation of unigrams in the *TF-IDF Yes* category suggests that TF-IDF is sensitive to the linguistic peculiarities of specific time periods. This provides strong evidence for why TF-IDF is more effective at categorizing patents based on their temporal context.

F.4. Synonyms Analysis

The objective of this analysis further explore the contrasting types of similarity captured by S-BERT and TF-IDF, particularly focusing on why S-BERT excels in class validation while TF-IDF shines in the period task. Our hypothesis posits that S-BERT, unlike TF-IDF, assigns a relatively lower weight to exactly overlapping words when determining similarity between patent pairs, and leans more towards semantic similarity and other forms of word "interchangeability." This distinction becomes apparent when analyzing patents within the same period that tend to exhibit period-specific overlapping language, even if they belong to different classes. Conversely, patents from the same class but different periods are more likely to exhibit similarity at a conceptual or idea level, which is the main type of similarity we aim to capture.

In preparing the data for analysis, we further stratified patent pairs from the Class/Period validation sample into two strata: tfidf_yes, S-BERT_yes, and agree (using the 75th percentile similarity cutoff for yes). For instance, S-BERT_yes implies that according to S-BERT this pair is similar, but according to TF-IDF, it is not. We further categorized them as same_class, same_period, both_same, and neither_same. To focus on informative cases, pairs in agree, both_same, and neither_same categories were excluded. A sample of 200 pairs from each of the 4 strata (800 pairs in total) was selected.

To enrich our analysis, we employed WordNet, a lexical database of English (Miller 1992). In WordNet, nouns, verbs, adjectives, and adverbs are grouped into sets of synonyms (synsets), each expressing a distinct word sense. These synsets are interlinked by means of semantic relations. The relations include hypernyms (more abstract terms), hyponyms (more specific terms). For each word in each patent, we listed all word senses. For each word sense, we found the set of synonyms, hypernyms, and hyponyms. These, along with the original word, were concatenated. For instance, for the word "air," we obtained a set of related terms encompassing synonyms like "breeze," hypernyms like "gas," and hyponyms like "zephyr."

Each patent was then represented as the set of unique tokens in it (each counted once) and separately as the set of unique tokens plus their synonyms, hypernyms, and hyponyms. For each document pair, we calculated the exact word overlap and the word plus synonym plus hypernym plus hyponym overlap (Word+ overlap).

We then conducted a pair of analyzes with the aim of investigating whether the same text characteristics drive both S-BERT similarity and belonging to the same_class category, as well as TF-IDF similarity and belonging to the same_period category. In the first analysis of the pair, we ran regressions with S-BERT and TF-IDF on the LHS and the text characteristics (exact word overlap and Word+ overlap) on the RHS. This analysis aimed to explore the relationship between the similarity scores generated by S-BERT and TF-IDF and the text characteristics.

In the second analysis of the pair, we conducted a PR AUC analysis with same_class and same_period categories as the dependent variables and the text characteristics as predictors. This analysis aimed to explore how well the text characteristics predict the categorization of patents into same_class and same_period categories.

The findings from both analyzes exhibited similar patterns: S-BERT similarity and same_class categorization were both driven by Word+ overlap, while TF-IDF similarity and same_period categorization were both driven by direct word overlap. These patterns led us to conclude that S-BERT's superior performance in same_class categorization can be attributed to its ability to capture the semantic similarity of words present in the patents, whereas TF-IDF's superior performance in same_period categorization can be attributed to its ability to capture direct word overlap.

The findings are shown in Table F.6 and Figure F.14, exhibiting expected patterns. Table F.6 quantitatively shows how WordNet-derived measures relate to S-BERT and TF-IDF similarity scores. The regression coefficients indicate that S-BERT's similarity scores are negatively associated with direct word overlap but positively associated with Word+ overlap, suggesting a stronger emphasis on semantic similarity (the negative coefficient on direct word overlap is not surprising, given our sampling strategy's focus on patent pairs where the two models disagree). Conversely, TF-IDF's similarity scores are positively associated with direct word overlap, indicating a preference for exact lexical matching.

Following the tabular analysis, Figure F.14 visually represents the Precision-Recall Area Under Curve (PR AUC) values for Word and Word+ overlap measures across same_class and same_period categorizations. In the same_class categorization, it is discernible from the figure that Word+ overlap (sim_combined) yields a higher PR AUC value of 0.49 compared to the Word overlap (sim_1_2) value of 0.43, underscoring the importance of capturing semantic relationships in addition to exact word overlap for classifying patents within the same class. Conversely, in the same_period categorization, Word overlap outperforms Word+ overlap with a PR AUC value of 0.588 against 0.512, indicating that direct word overlap is more pertinent for capturing period-specific similarities. The Figure also shows that, S-BERT

Table F.6: Regression results for similarity scores and Wordnet-based measures on the `S-BERT_yes` and `tfidf_yes` patent sample

	TF-IDF	S-BERT
(Intercept)	0.31^{***}	0.58^{***}
	(0.02)	(0.02)
Word Overlap	0.39^{***}	-0.29^{***}
	(0.04)	(0.04)
Word+ Overlap	-0.01	0.13^{**}
	(0.04)	(0.04)
R^2	0.15	0.06
Adj. R^2	0.15	0.06
Num. obs.	800	800

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

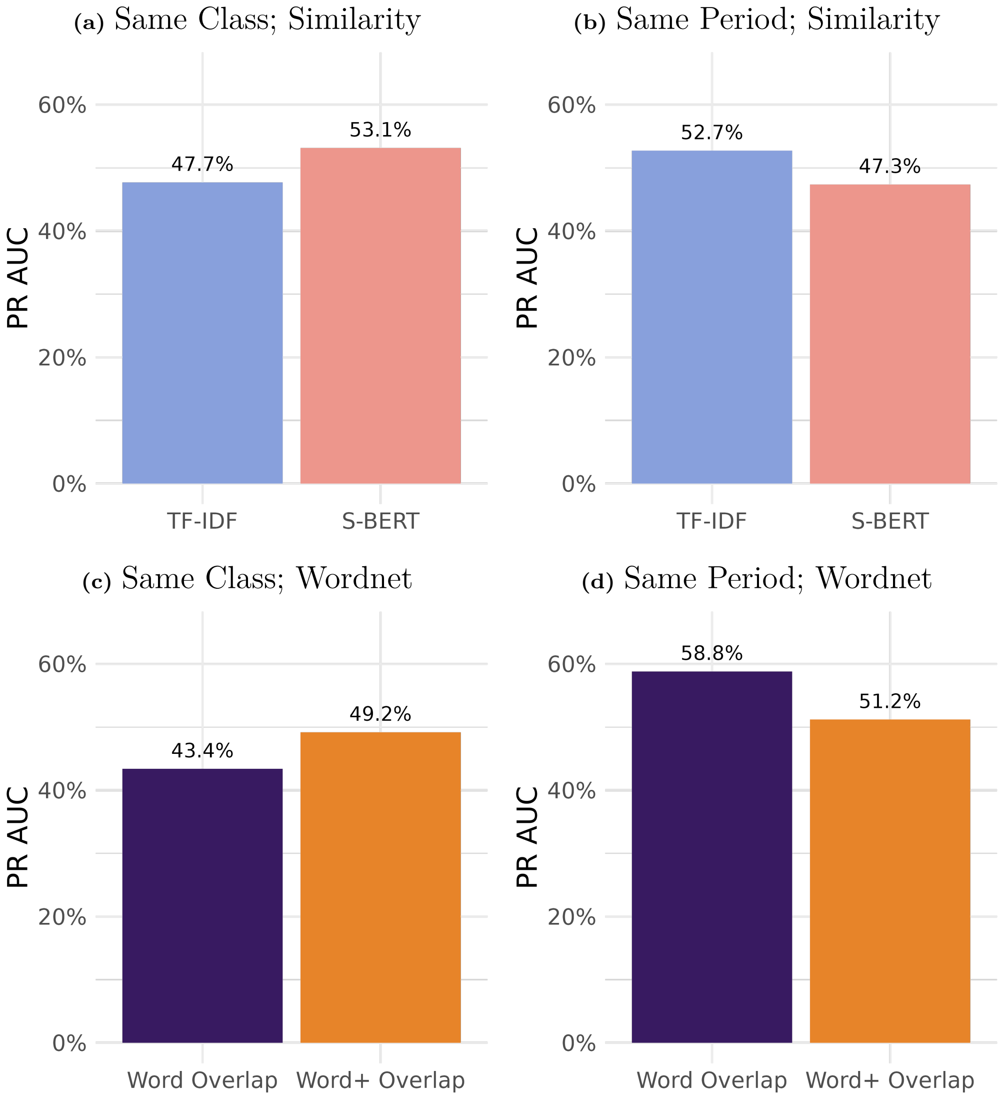
Notes: The table presents the coefficients from a regression analysis where the dependent variables are the similarity scores generated by TF-IDF and S-BERT. The independent variables are Word Overlap, representing the exact word overlap between patent pairs, and Word+ Overlap, representing the overlap including synonyms, hypernyms, and hyponyms. The negative coefficients for S-BERT on Word Overlap and for TF-IDF on Word+ Overlap are observed due to the sampling strategy focusing on patents where the two models disagree.

performs best on `same_class` task and TF-IDF performs `same_period` task on the sub-sample used in this analysis, conforming with the full sample results discussed in Section 3.3.

In conclusion, one of the mechanisms through which S-BERT better captures idea similarity is through its ability to assign similar vectors to words located closely in the semantic graph (synonyms, hypernyms, hyponyms). This is consistent with the properties theoretically expected from S-BERT based on its architecture and training procedure. Our results show that these properties are useful in innovation economics by allowing S-BERT to capture the similarity of ideas in a way that transcends period-specific language.

F.5. Why is S-BERT Better? Conclusion

The Google Ngrams analysis and the patent pair example collectively offer robust evidence to support our initial observations. TF-IDF's strength lies in identifying patents from the same time period, primarily due to its sensitivity to words that are popular within specific temporal contexts. Conversely, S-BERT proves superior at classifying patents into


(a) Same Class; Similarity

(b) Same Period; Similarity

(c) Same Class; Wordnet

(d) Same Period; Wordnet

Figure F.14: Similarity scores based on the S-BERT and TF-IDF representations and Wordnet-based measures for categorizing patent pairs as belonging to the same class and period

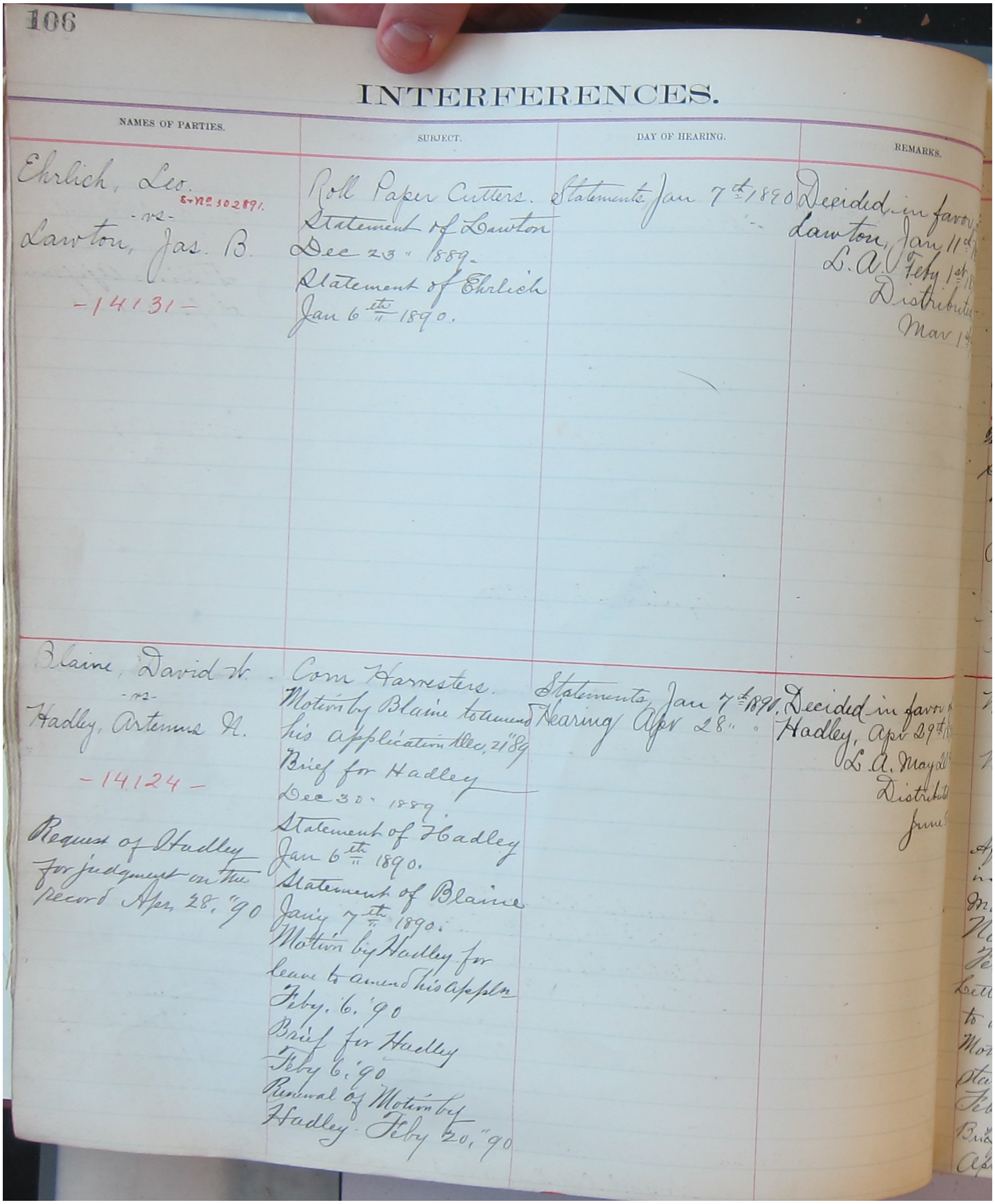
Notes: The sample includes patent pairs in the `S-BERT_yes` and `tfidf_yes` categories. We evaluate how well patent pairs can be classified as belonging to the same class or the same quarter-century period using two sets of similarity scores, based on S-BERT and TF-IDF representations, and two sets of Wordnet-based measures, Word Overlap and Word+ Overlap. "Word" represents exact word overlap and "Word+" encompasses word overlap along with their synonyms, hypernyms, and hyponyms as derived from Wordbet, a lexical database grouping English words into sets of synonyms and recording their semantic relationships.

the same technical class, given its ability to understand and capture the semantic essence of the text, highlighted by its association with synonym, hypernym, and hyponym overlap as opposed to the exact word overlap. These insights are important for choosing the more appropriate model for specific downstream tasks.

Appendix G. Miscellanea

G.1. Photograph of the Register of Interferences

Figure G.15 shows an example page from one of the Register volumes. It displays two cases. Both cases record hearing dates of January 7, 1890. The subject of the first case was roll paper cutters and the competing inventors were named Ehrlich and Lawton. The case was decided in favor of Lawton on January 11. The subject of the second case, Blaine v. Hadley, was corn harvesters; the case was decided in favor of Hadley on April 29th.



106

INTERFERENCES.

NAMES OF PARTIES.	SUBJECT.	DAY OF HEARING.	REMARKS.
Ehrlich, Leo. S-No. 302891. -vs- Lawton, Jas. B. -14131-	Roll Paper Cutters. Statement of Lawton Dec 23" 1889. Statement of Ehrlich Jan 6th" 1890.	Statements Jan 7th 1890	Decided in favor Lawton, Jan 11th L.A. Feby 1st Distributed Mar 1
Blaine, David W. -vs- Hadley, Artemus N. -14124- Request of Hadley for judgment on the record Apr. 28, '90	Corn Harvesters. Motion by Blaine to amend his application Dec, 21 '89 Brief for Hadley Dec 30" 1889 Statement of Hadley Jan 6th" 1890. Statement of Blaine Jany 7th" 1890. Motion by Hadley for leave to amend his appln Feby. 6. '90 Brief for Hadley Feby 6. '90 Renewal of Motion by Hadley. Feby 20, '90	Statements Jan 7th 1890. Hearing Apr 28" "	Decided in favor Hadley, Apr 29th L.A. May Distributed June

Figure G.15: Example page from Register of Interferences