Too Good to Launch: The Innovator's Dilemma at Scale

Tianshu Lyu [†] Jiawen Wu [‡] Zunda Winston Xu [§]

Junuary 30, 2026

Abstract

We construct a novel dataset linking over 360,000 U.S. utility patents held by public firms to 244,000 products those firms bring to market, using sentence embeddings to measure patent-product similarity and large language models to extract structured product information from unstructured corporate announcements. Using this dataset, we provide the first large-scale quantitative evidence for the innovator's dilemma. Within the same firm, technology class, and year, patents that are more dissimilar from the firm's existing products are substantially less likely to be commercialized and take significantly longer to reach the market when they are. The underutilization of disruptive technologies is concentrated among firms whose revenues depend most heavily on existing products, as the cannibalization hypothesis predicts. We further show that the cannibalization incentive extends upstream into R&D itself: firms with more valuable existing product portfolios generate less disruptive patent portfolios, suggesting that the anticipation of commercialization costs shapes the direction of research. Taken together, these findings imply that incumbent firms not only fail to commercialize disruptive technologies they have already developed, but also redirect their research away from disruptive directions, with consequences for the aggregate pace and direction of technological change.

JEL Codes: O31, O33, O34, L25, C55

Keywords: Patents, Products, Innovator's Dilemma, Creative Destruction, Artificial Intelligence

We thank Nicholas Barberis, Nicholas Bloom, Greg Buchak, Paul Goldsmith-Pinkham, Song Ma, Matt Marx, Amit Seru, Kelly Shue, Vikrant Vig, and participants at the NBER Innovation Information Initiative (I^3) Technical Working Group Meeting for their helpful comments.

[†]Yale School of Management; tianshu.lyu@yale.edu

[‡]Stanford Graduate School of Business; jiawenwu@stanford.edu

[§]Stanford Graduate School of Business; zundaxu@stanford.edu

1 Introduction

In 2017, a team of researchers at Google invented the Transformer, the architecture that would power virtually all modern large language models (Vaswani et al., 2017). Google also employed many of the leading AI researchers and was the first to build dedicated AI infrastructure at scale. Yet when OpenAI released ChatGPT in late 2022, Google was widely perceived as having been caught flat-footed. Sergey Brin later acknowledged publicly that Google had been "too timid" in deploying language models to consumers.[1] One reason for this timidity is not hard to identify: Google's core business is search advertising, and large language models represent a direct threat to that business model. Even if Google's LLMs were to become the leading replacement for traditional search, the advertising revenue model that generates the vast majority of Google's profits might not easily port to a conversational AI interface.

Google's experience echoes the most famous case of the innovator's dilemma: Kodak's digital camera. Kodak engineers built the first digital camera prototype in 1975 and the company was granted the foundational patent in 1978.[2] But Kodak's executives, whose profits flowed from the high-margin film business, delayed investing in digital photography for decades—long enough for competitors without legacy film revenues to capture the digital market. Christensen (2015) coined the term "innovator's dilemma" to describe this pattern: incumbent firms that possess disruptive technologies but fail to commercialize them because doing so would cannibalize their existing products.

Despite the prominence of this idea in business strategy and economic policy, systematic empirical evidence for the innovator's dilemma remains limited. The existing literature has relied primarily on case studies (Christensen, 2015; Henderson and Clark, 1990; Tripsas, 1997) or structural analyses of single industries (Igami, 2017). While these studies provide

[1] See https://www.acquired.fm/episodes/google-the-ai-company and https://www.reddit.com/r/singularity/comments/1fdy9b0/sergey_brin_admits_that_google_was_too_timid_to/.

[2] U.S. Patent No. 4,131,919, "Electronic Still Camera," filed 1975, granted 1978.

rich institutional detail, they cannot establish whether the innovator's dilemma is a general phenomenon or an artifact of particular industries and time periods. Testing the innovator's dilemma at scale requires data that link the technologies firms *invent* (patents) to the products they *sell*—data that, until now, have not existed in a systematic form.

In this paper, we construct such a dataset and use it to provide the first large-scale quantitative evidence for the innovator's dilemma. Our dataset links the universe of U.S. utility patents held by public firms to the products those firms announce, covering over 360,000 patents and 244,000 products across more than 5,000 firms from 2000 to 2025. The resulting dataset allows us to observe, for each patent a firm holds, whether and when that technology is commercialized into a product, enabling direct measurement of the utilization decisions that are at the heart of the innovator's dilemma.

The construction of this dataset is itself a contribution to economic measurement, as it requires solving several challenges that cannot be addressed with standard structured data. The first challenge is identifying which patents are embodied in which products. We address this by embedding patent text and product announcement descriptions into a common semantic space using a sentence-embedding model (all-MiniLM-L6-v2) and computing pairwise cosine similarity for each product–patent pair within a firm. To determine the similarity thresholds that distinguish utilized from unutilized patents, we validate our approach against the IPRoduct dataset (de Rassenfosse, 2020), which provides ground-truth patent–product links based on firms' Virtual Patent Marking disclosures. The validation confirms that utilized patent–product pairs exhibit significantly higher semantic similarity across all eight CPC technology sections, and we set CPC-section-specific thresholds accordingly. The second challenge is extracting structured product information from unstructured corporate announcements. A single announcement from S&P Capital IQ Key Developments may describe multiple products, and the same product may appear across different news sources. We use a large language model (Gemini 2.5 Flash) to extract individual product names from announcement text, classify products into product lines, and map products to

Compustat business segments via semantic similarity between product and segment names. These LLM-based procedures yield variables—product line entry, product–segment revenue linkages, prior sales share—that are central to measuring the cannibalization incentive but would be infeasible to construct at scale through manual coding or rule-based extraction. Together, these AI-based tools transform two largely separate data universes—patent filings and product announcements—into an integrated firm-level panel of technology commercialization decisions.

Using this dataset, we document five empirical facts. First, disruptive patents—those most dissimilar from the firm's existing products—are substantially less likely to be utilized in any subsequent product. Within the same firm, technology class, and year, a one-standard-deviation increase in our disruption measure is associated with a 19.6 percentage point decrease in utilization probability (Fact 1). Second, conditional on eventual utilization, disruptive patents take significantly longer to reach the market: a one-standard-deviation increase in disruption is associated with approximately one additional year of delay (Fact 2). Third, when disruptive patents are eventually commercialized, they disproportionately give rise to entirely new product lines rather than extensions of existing ones, validating the economic content of our disruption measure (Fact 3).

The fourth and fifth facts speak directly to the mechanism. Fact 4 shows that the utilization penalty for disruptive patents is significantly amplified in firms whose revenue depends most heavily on existing product lines. The interaction between disruption and prior sales share is large ($t = -4.64$): a firm at the 75th percentile of prior sales share faces a utilization penalty roughly 43% larger than a firm at the 25th percentile. This is the pattern predicted by the cannibalization hypothesis—firms with the most to lose from disruption are the most reluctant to commercialize it—and is difficult to rationalize under alternative explanations. Fact 5 documents that the cannibalization incentive extends upstream into R&D: within the same firm, years with higher prior sales shares are associated with less disruptive patent portfolios, suggesting that firms anticipate the commercialization costs and adjust their research

direction accordingly.

We then investigate two prominent alternative explanations and show that neither can account for these patterns. First, using the KPSS patent value measure (Kogan et al., 2017), we show that the financial market does not systematically undervalue disruptive patents at issuance, and that controlling for market valuation has essentially zero effect on the disruption coefficient. Second, using a revealed-preference measure of complementary technology availability—based on the co-utilization of patents within the same product—we show that complement readiness is a strong predictor of commercialization timing, but the disruption coefficient actually *increases* by 25% after controlling for it. The residual disruption effect in the most saturated specification, which simultaneously controls for market valuation, complementary technology, and CPC section $\times$ year fixed effects, is 0.900 ($t = 13.3$). Neither market forces nor technological constraints can explain the utilization gap; the evidence points to the firm's strategic reluctance to cannibalize its own products.

Our paper makes three contributions. First, we construct and validate a new large-scale dataset that links patents to products, enabling direct measurement of technology commercialization decisions. The dataset itself is a contribution to economic measurement: by combining sentence embeddings, large language models, and traditional structured data, we demonstrate how modern AI tools can be used to construct economic variables—patent utilization, product line entry, product–segment revenue linkages—that are central to the study of innovation but have previously been unavailable or limited to narrow industry samples. Second, we provide the first systematic, multi-industry evidence for the innovator's dilemma. While Christensen (2015) developed the concept through case studies and Igami (2017) provided structural evidence in the hard disk drive industry, our analysis spans the full cross-section of U.S. public firms and shows that the patterns are pervasive: disruptive patents are underutilized, delayed, and the penalty is concentrated in firms with the most to lose. Third, we document a novel upstream distortion: the cannibalization incentive affects not only which patents firms choose to commercialize but also which technologies they choose

to develop, with implications for the aggregate direction of technological progress.

Related Literature

Our paper relates to several strands of the literature.

Most directly, we contribute to the empirical literature on the innovator's dilemma and creative destruction. Christensen (2015) articulated the core mechanism through detailed case studies, arguing that incumbent firms fail to adopt disruptive technologies because these technologies initially serve different market segments and threaten existing revenue streams. Igami (2017) provides the first structural estimation of the innovator's dilemma, using data from the hard disk drive industry to estimate a dynamic oligopoly model in which incumbents rationally delay adoption of disruptive technologies due to cannibalization concerns. Our contribution is to provide reduced-form evidence across the U.S. public firms using a new patent-to-product dataset, documenting both the commercialization distortion and the upstream R&D distortion that the innovator's dilemma implies.

Our work also relates to the theoretical literature on creative destruction and the replacement effect. Arrow (1962) showed that a monopolist has weaker incentives to innovate than a competitive firm because the monopolist internalizes the cannibalization of its own rents. Aghion et al. (2001, 2005) extended this insight to models of endogenous growth where the intensity of creative destruction depends on the incumbent's market position. Akcigit and Kerr (2018) study whether firms choose internal (improving existing products) versus external innovations (acquiring new product lines), finding that large firms prefer internal innovation. Our Fact 4 provides a product-level analog of the replacement effect: firms whose existing products account for a larger share of revenue face stronger replacement incentives and are less willing to commercialize disruptive technologies. Fact 5 is consistent with the literature on directed technical change (Acemoglu, 2002), showing that the expected profitability of disruptive innovation—reduced by cannibalization costs—leads firms to steer their

R&D toward incremental improvements.

We also contribute to the growing literature that uses AI and natural language processing to construct economic measures of innovation. Kelly et al. (2021) use textual similarity between patents to measure technological novelty and identify breakthrough innovations. Kogan et al. (2017) estimate the private value of patents from stock market reactions. Lyu and Xu (2025) study how investors react to patent novelty and how misreactions distort innovation directions. Our paper complements these approaches by introducing a patent-to-*product* link: rather than measuring novelty relative to other patents, we measure novelty relative to the firm's actual product portfolio, and rather than inferring commercialization from citations or stock returns, we observe it directly in the product data. The use of sentence embeddings for patent–product matching and large language models for product extraction and classification demonstrates how AI tools can be deployed to construct firm-level variables at a scale that was previously infeasible.

Our dataset construction builds on a growing body of work that links patents to products. de Rassenfosse (2020) introduces the IPRoduct dataset, which uses firms' Virtual Patent Marking disclosures to identify patents underlying specific products; we use IPRoduct as an external benchmark to validate our semantic similarity approach. Mukherjee et al. (2025) use the S&P Capital IQ Key Developments database to study the determinants of new product introductions by public firms. Argente et al. (2025) link patents to products using barcode-level scanner data to study how product innovation shapes firm dynamics. Harford et al. (2025) measure patent utilization to study the value implications of firms' commercialization decisions. Our contribution relative to this literature is methodological: by combining sentence embeddings with large language models, we construct patent–product links, product line classifications, and product–segment revenue mappings at a scale and level of detail that prior approaches—whether based on patent marking disclosures, barcode data, or manual classification—have not achieved. The resulting dataset covers over 360,000 patents and 244,000 products across the U.S. public firms, enabling the first systematic analysis of

technology commercialization decisions across industries.

Finally, our findings have implications for the broader literature on declining business dynamism (Decker et al., 2016) and the reallocation of innovation across firm types (Akcigit and Ates, 2023). If incumbent firms systematically underutilize and underinvest in disruptive technologies, the commercialization of breakthroughs may depend disproportionately on new entrants who do not face the cannibalization trade-off. Our patent-to-product data provide a micro-foundation for these aggregate patterns: we observe directly that incumbent firms hold the patents for disruptive technologies but choose not to commercialize them.

The paper proceeds as follows. Section 2 describes the data sources. Section 3 presents the methodology for linking patents to products and constructing the key measures. Section 4 documents the five empirical facts. Section 5 investigates alternative mechanisms and provides further discussion for the cannibalization channel. Section 6 concludes.

2 Data

2.1 Product data

Our primary source of product information is the S&P Capital IQ Key Development database, which provides structured summaries of major corporate news and events for publicly traded companies. Following prior work (Mukherjee et al., 2025), we focus on event type 41 (Product-Related Announcements). We further restrict to announcements with headlines that indicate new product introductions ("unveil", "reveal", "launch", "introduce", "release", "present", "issue", "upgrade", "enhance", "improve", "extend", or "new"), and exclude announcements related to natural resource exploration (e.g., "drill", "explore", "excavate") to exclude operational updates in extractive industries. We focus on new-product announcements associated with 5,040 public firms with matched patent records from January 2000 to June 2025.

These new-product launch announcements provide detailed information on the relevant products, including the launching firm, the announcement date, and a brief description. However, a single announcement may cover multiple products, and the same product may appear in multiple announcements across different news sources. As a result, identifying economically meaningful product-level observations rather than announcement-level mentions cannot be achieved through simple keyword rules or deterministic parsing. To address this challenge, we use a large language model (Google's Gemini 2.5 Flash) to extract product names directly from the unstructured announcement text. We focus on the *situation* field, which provides the most detailed narrative description of the product offering, and prompt the model to identify all product names mentioned in each announcement. Detailed information on the product name extraction procedure is provided in Appendix B.

After extraction, we standardize product names by normalizing character formats, converting text to uppercase, harmonizing common formatting variations, and removing extraneous punctuation and whitespace. To further ensure the uniqueness of identified products, we treat announcements by the same firm that reference the same standardized product name within a six-month window as pertaining to a single product, and we retain the earliest associated announcement as the product's introduction. Following this procedure, we identify 244,563 unique products from the 159,519 new-product announcements in our sample.

2.2 Patent data

Our patent data comes from the U.S. Patent and Trademark Office (USPTO). Because the earliest product announcement in our sample is in 2000 and U.S. utility patents typically expire after 20 years, we focus on utility patents granted from 1980 onward, which plausibly encompass the relevant stock of patented technologies available for commercialization during our sample period. We match patents to public firms using assignee names, following the procedure in Kogan et al. (2017). The patent records include information on filing and grant dates, assignees, technology classifications, and textual descriptions, which we use in

subsequent analyses.

2.3 Benchmark product–patent links: IPRoduct

To study firms' commercialization decisions, a key input is information on which patented technologies are actually utilized in products. We establish this link by measuring semantic similarity between product and patent descriptions, based on the idea that patents embodied in a product are more likely to describe the underlying technologies referenced in product announcements.

To validate this method and estimate the similarity thresholds, we use the IPRoduct dataset developed by de Rassenfosse (2020), which links a limited set of products to the patents firms disclose as underlying those products. IPRoduct constructs these links using firms' Virtual Patent Marking (VPM) disclosures, which indicate patent protection for specific products.The IPRoduct data have been used in prior work to study the economic effects of patent protection. For example, de Rassenfosse and Zhou (2020) use VPM disclosures to show that the expiration of the 20-year patent term is associated with an 8–10% decline in product prices.

In this paper, we use IPRoduct exclusively as an external benchmark to validate our procedure for linking products to patents. In particular, we assess whether patent–product pairs identified in VPM disclosures exhibit higher semantic similarity between product and patent text than other candidate pairs. We formally describe our linking methodology and validation results in Section 3.

2.4 Public-firm data

We obtain public-firm financial and accounting information from the Compustat database. In particular, we use the Compustat Historical Segments database, which reports firms' segment-level sales by product line or business division as disclosed in financial statements.

We use this dataset to map newly introduced products to firms' business segments and to measure segment-level sales associated with firms' prior products. The procedure used to map products to segments is described in Section 3.4.

3 Methodology: Linking Patent to Product

This section describes the construction of the key measures used in the analysis. We first develop a method to link patents to products in order to measure patent utilization and commercialization timing. We then introduce additional measures that characterize the nature of firms' innovations and products, including the technological disruptiveness of patents, the introduction of new product lines, and the value of firms' existing product portfolios.

3.1 Measuring patent utilization

We begin by describing our methodology for linking products to their underlying patents. The central premise of our approach is that patents utilized in a product should be semantically more similar to that product than other unutilized patents.

Following this idea, we first identify, for each product, the set of all patents granted to the firm prior to the product's announcement date as candidate patents. We then apply a common semantic embedding model to both product descriptions and patent text, and compute pairwise similarity scores for each product–patent pair. To validate this approach, we match products from the Key Developments dataset to products in the IPRoduct dataset, which contains known patent–product links, and assess whether these ground-truth links exhibit higher semantic similarity. Based on these validation results, we determine text similarity thresholds and apply them to the full set of product-patent pairs in our sample to establish patent-product links.

Candidate patent set For each product, we define the candidate patent set as all patents granted to the firm prior to the product's announcement date. This restriction ensures that only technologies available to the firm at the time of commercialization are considered.

Semantic similarity and validation For each product-patent pair in the candidate set, we compute a semantic similarity score between the product description and the patent text. Product text is taken from the product announcement *situation* field of the product announcement, while patent text is constructed from the patent description. We embed product descriptions and patent documents using a common sentence-embedding model (all-MiniLM-L6-v2) and measure similarity using cosine similarity between embeddings.

To assess whether semantic similarity meaningfully captures patent–product relationships, we validate our approach using the IPRoduct dataset introduced in Section 2.3, which externally identified patent-product links based on firms' Virtual Patent Marking disclosures. We compare similarity scores for patent–product pairs identified in IPRoduct to those for other candidate patents held by the same firm.

To implement this validation, we first match firms and products across the Key Developments and IPRoduct datasets. After cleaning firm names, we apply fuzzy matching within the same country and manually review all candidate matches, yielding 197 confidently matched public firms. Manual inspection indicates that IPRoduct records relatively more fine-grained products (e.g., variants within a product series) than Key Developments. Accordingly, we implement a many-to-one mapping from IPRoduct products to Key Developments announcements. We retain high-confidence matches with similarity exceeding 0.55, resulting in 7,556 IPRoduct products matched to 1,551 Key Developments products.

Using these matched products, we label candidate patents as utilized or unutilized based on whether they are disclosed in IPRoduct as underlying the product. This yields 1,443 Key Developments products across 146 firms with computed product–patent similarity scores, among which 761 products in 102 firms are associated with at least one utilized patent.

We assess the validity of the patent–product matching procedure by comparing cosine similarity scores between product announcement text and candidate patent text across patents classified as utilized and unutilized for the matched products. These comparisons inform the choice of similarity thresholds used to classify patent–product links. To account for heterogeneity in patent language and product descriptions across technological domains, we allow similarity thresholds to vary by technology class and therefore conduct the validation at the level of Cooperative Patent Classification (CPC) sections.

Figure 1 plots, for each Cooperative Patent Classification (CPC) section, the distribution of product–patent similarity scores separately for patents identified as utilized and unutilized. Across all CPC sections, utilized patents exhibit systematically higher similarity, with clear separation from unutilized patents. Consistent with these patterns, Table 1 reports CPC-section–level t-tests that confirm statistically significant differences in mean similarity between utilized and unutilized patents in all sections.

Similar validation results at the product and firm levels are reported in Appendix C. Taken together, the consistent patterns observed across technology classes, products, and firms provide strong support for the validity of our semantic similarity approach in capturing economically meaningful patent–product relationships.

In our main analysis, we adopt CPC-specific similarity thresholds in our baseline specification, defined as one standard deviation above the mean similarity of unutilized patents within each CPC section in the validation sample.

Final patent–product links We apply the estimated similarity thresholds to all product-patent pairs in the full sample to construct patent–product links. A patent is linked to a product if its similarity score exceeds the relevant threshold for the product's technology class. These links allow us to characterize whether and when a patented technology is subsequently utilized in a marketed product.

We first define a binary indicator for patent utilization, which captures whether patent i

is incorporated into any product launched by its owner after the patent is filed:

$$\text{Utilized}_i = \mathbb{I}\left[\max_{p \in \mathcal{P}^{\text{post}}_{j(i)}} \text{Sim}(i,p) > \bar{s}_{c_i}\right], \tag{1}$$

where $\mathcal{P}^{\text{post}}_{j(i)} = \{p \in \mathcal{P}_{j(i)} : \text{Launch}_p > \text{Filing}_i\}$ denotes the set of products launched by firm $j(i)$ after the filing of patent i, and $\bar{s}_{c_i}$ is the similarity threshold corresponding to patent i's technology class c_i. The indicator Utilized_i equals one if the patent is linked to at least one subsequent product and zero otherwise.

For patents that are ultimately utilized, we also measure the timing of commercialization. Specifically, we define the time to utilization as the number of years between the patent filing date and the announcement date of the first product utilizing that patent:

$$\text{TimeToUtil}_i = \min_{p \in \mathcal{P}^{\text{post}}_{j(i)} : \text{Sim}(i,p) > \bar{s}_{c_i}} \left(\text{Launch}_p - \text{Filing}_i\right). \tag{2}$$

This measure captures delays in commercialization and forms the basis for our analysis of firms' technology adoption and product introduction decisions.

3.2 Measuring technology disruptiveness

To examine incumbent firms' commercialization decisions for potentially disruptive technologies, we measure how novel a new technology is relative to the firm's existing product portfolio. A technology is disruptive in this sense if it is relevant to the firm's activities but substantially different from its existing products.

We capture this idea by defining the "disruptiveness" of a patent as the distance between the patented technology and the firm's prior products. Specifically, disruptiveness measures

how different patent i is from the existing product portfolio of its owner $j(i)$:

$$\text{Disruption}_i = 1 - \max_{p \in \mathcal{P}^{\text{prior}}_{j(i)}} \text{Sim}(i, p), \tag{3}$$

where $\mathcal{P}^{\text{prior}}_{j(i)} = \{p \in \mathcal{P}_{j(i)} : \text{Launch}_p < \text{Filing}_i\}$ denotes the set of products launched by firm $j(i)$ prior to the filing of patent i. A higher value of Disruption_i indicates that the patented technology is less similar to the firm's existing products and therefore more technologically novel.

A potential concern with this measure is that a patent may appear highly disruptive simply because it is technologically unrelated to the firm's core business, rather than because it represents a novel extension of the firm's existing technologies. In such cases, low similarity to prior products would reflect irrelevance rather than disruptiveness. To address this concern, we restrict attention to patents that fall within technology classes relevant to the firm's existing product portfolio. Specifically, we pool all products previously launched by a firm and identify the set of CPC sections associated with the patents underlying those products. We then define disruptiveness only for subsequent patents that belong to these relevant CPC sections. This restriction ensures that the disruptiveness measure captures technological novelty within the firm's domain of activity, rather than entry into unrelated technological areas.

3.3 Identifying new product lines

In addition to identifying individual products, to distinguish incremental product introductions from the launch of fundamentally new product categories, we classify products into product lines using the unstructured text in product announcements. Product lines are not directly observed in standard firm-level data and are difficult to construct manually at scale, given the volume and heterogeneity of product disclosures. We therefore use a large language model (LLM)–based extraction procedure to systematically infer product line identifiers from

product announcement text.

Starting from the Key Developments product-level dataset described in Section 2.1, we use the Gemini API (Gemini 2.5 Flash) to extract a product line label for each product based on the announcement description. We then apply a standardized cleaning procedure to normalize extracted product lines, removing trademark markers, formatting artifacts, model-specific suffixes, and other non-essential tokens to obtain consistent product line names.

In cases where the LLM returns an indeterminate or missing product line, we supplement the classification using the product name reported in the announcement and apply the same cleaning rules. To address potential over-specificity and measurement noise, we construct several alternative product line definitions that vary in granularity, ranging from raw cleaned product line names to more general identifiers based on the core tokens of the product line name. We further define conservative, validated versions of these product line identifiers by restricting attention to product lines that are directly extracted by the LLM for a given firm. In our main analysis, we use a validated product line definition based on the first core token of the cleaned product line name.

Appendix D provides additional details on the LLM extraction procedure, including prompt design, post-processing steps, and illustrative examples of extracted product lines.

3.4 Product–segment mapping and existing product value

To measure a firm's dependence on its existing product portfolio, we link products to business segments using Compustat Segment sales data and aggregate segment-level sales associated with previously introduced products. Because products are not directly mapped to Compustat business segments, we establish this link using semantic similarity between product and segment names.

Specifically, for each product identified in the Key Developments data, we consider all business segments reported by the firm in each year following the product's announcement.

We embed product and segment names using a common embedding model and assign each product to the segment with the highest semantic similarity. This procedure yields, for each product and year, an associated business segment and the corresponding segment-level sales reported in Compustat Segments.

Using this mapping, we construct $\text{PriorSalesShare}_{g,t}$, the share of firm g's total revenue attributable to business segments linked to products announced prior to the filing date t of a given patent. This measure is evaluated at the time of patent filing and captures the extent to which a firm's revenues are tied to its established products when making subsequent commercialization decisions.

4 The Innovator's Dilemma: Five Facts

In this section, we document five novel empirical facts about firms' decisions to commercialize their patented technologies into products. These facts collectively provide evidence consistent with the innovator's dilemma (Christensen, 2015): incumbent firms with valuable existing product lines systematically under-utilize their most disruptive technologies, take longer to bring them to market, and even distort their upstream R&D portfolios away from disruptive innovation. The analysis leverages the patent-to-product dataset described in Section 2.

We organize the section as follows. Section 4.1 establishes that disruptive patents are less likely to be utilized in products. Section 4.2 shows that, conditional on utilization, disruptive patents take significantly longer to reach the market. Section 4.3 confirms the economic interpretation of our disruption measure by showing that disruptive patents lead to entirely new product lines rather than extensions of existing ones. Section 4.4 provides direct evidence of the product cannibalization channel: the utilization penalty for disruptive patents is significantly larger in firms whose revenue depends heavily on existing product lines. Section 4.5 documents that the cannibalization incentive extends upstream into the

R&D process, distorting firms' patent portfolios away from disruptive technologies.

4.1 Fact 1: Disruptive Patents Are Less Likely to Be Utilized

We begin by testing the most basic prediction of the innovator's dilemma: that incumbent firms are less likely to commercialize their more disruptive patents. This prediction follows from the observation that disruptive technologies, by definition, do not fit neatly into the firm's existing product portfolio. Commercializing such technologies may require new production processes, new distribution channels, and most critically, may cannibalize sales of the firm's existing products. If these costs and risks deter firms from utilizing disruptive patents, we should observe a negative relationship between a patent's disruption score and its probability of being utilized in a subsequent product.

Figure 2 provides a first look at this relationship. The figure presents a binscatter plot of the utilization rate against the disruption score, after residualizing both variables on firm and filing year fixed effects. The sample is divided into 20 equal-sized bins of the residualized disruption score, and the average residualized utilization rate is plotted within each bin. The relationship is negative, monotonic, and approximately linear: within the same firm and year, patents with higher disruption scores are substantially less likely to be commercialized.

To quantify this relationship, we estimate the following linear probability model at the patent level:

$$\text{Utilized}_i = \alpha_g + \alpha_t + \beta \cdot \text{Disruption}_i + \varepsilon_i, \tag{4}$$

where Utilized_i is an indicator equal to one if patent i is utilized in any product announced after its filing date. The key explanatory variable, Disruption_i, is defined in Equation (3). We include different sets of fixed effects across five specifications to control for distinct sources of variation and ensure that the results are not driven by confounders. Table 2 presents the results. Column 1 is a simple OLS regression without fixed effects, providing the unconditional relationship. Column 2 adds firm fixed effects (α_g), which absorb time-invariant firm

characteristics such as industry, firm size, R&D intensity, and organizational culture. This specification compares patents within the same firm, asking whether a firm's more disruptive patents are less likely to be utilized than its less disruptive ones. Column 3 instead adds filing year fixed effects (α_t), controlling for aggregate trends in patenting and product launches. Column 4 includes both firm and year fixed effects, exploiting within-firm, within-year variation. Column 5 replaces the separate year fixed effects with CPC section $\times$ year fixed effects, which control for technology-class-specific trends in utilization rates. This specification compares patents filed by the same firm, in the same technology class, in the same year. Standard errors are clustered at the firm level throughout to account for within-firm correlation in utilization decisions.

Across all five specifications, the coefficient on Disruption_i is negative and statistically significant at the 1% level. The coefficient grows in magnitude as we add fixed effects, from -1.038 in column 1 to -1.504 in column 5, suggesting that the unconditional estimate is attenuated by positive confounders (e.g., firms in high-disruption industries may also have higher utilization rates). In our preferred specification (column 5), the coefficient of -1.504 ($t = -20.9$) implies that a one-standard-deviation increase in disruption (approximately 0.13) is associated with a 19.6 percentage point decrease in the probability of utilization. Given a mean utilization rate of approximately 83%, this represents a 24% decrease relative to the mean. The effect is both statistically and economically large: within the same firm, in the same technology class and year, more disruptive patents are substantially less likely to be commercialized into products.

4.2 Fact 2: Disruptive Patents Take Longer to Utilize

The previous result establishes that disruptive patents are less likely to be commercialized overall. But the innovator's dilemma may also operate on the intensive margin: even among patents that are eventually utilized, disruptive technologies may face longer delays before reaching the market. Such delays could arise from the same cannibalization concerns that

deter utilization entirely—firms may postpone commercialization while they search for ways to introduce disruptive products without undermining existing revenue streams—or from practical challenges such as building new production capabilities or distribution channels for technologies that do not fit the existing product architecture.

Figure 3 plots empirical survival curves for time to first utilization among utilized patents, separately by disruption tercile. The horizontal axis measures years from the patent filing date; the vertical axis measures the fraction of patents in each tercile that have not yet been utilized by that horizon. The three curves separate early and the gap widens monotonically: patents in the lowest disruption tercile reach 50% utilization roughly 3 years before patents in the highest tercile. The separation is visible from the earliest horizons, suggesting that the delay is not driven by a small number of exceptionally late adopters but rather reflects a systematic shift in the entire distribution of commercialization timing.

To formalize these patterns, we restrict the sample to utilized patents and estimate:

$$\ln(\text{TimeToUtil}_i) = \alpha_g + \alpha_t + \beta \cdot \text{Disruption}_i + \varepsilon_i, \tag{5}$$

where TimeToUtil_i is the number of years between the patent filing date and the announcement date of the first product utilizing that patent. We use the log transformation to account for the right-skewed distribution of time to utilization and to interpret the coefficient as an approximate percentage change. As in Fact 1, we vary the set of fixed effects across five specifications, from no fixed effects (column 1) to firm and CPC section $\times$ year fixed effects (column 5). Standard errors are clustered at the firm level.

Table 3 reports the results. Across all specifications, the coefficient on Disruption is positive and significant at the 1% level. In column 5, the coefficient is 0.854 ($t = 12.2$). To interpret this magnitude, a one-standard-deviation increase in disruption is associated with approximately 12% longer time from filing to first product utilization. Given that the median utilized patent takes 5.3 years from issuance (8.8 years from filing) to appear in a product, a

12% increase translates to roughly one additional year of delay. Even among the patents that firms eventually commercialize, the more disruptive ones face substantially longer delays.

One important consideration in interpreting the time-to-utilization regressions is selection bias: we observe time to utilization only for patents that are eventually utilized, and Fact 1 shows that more disruptive patents are less likely to be utilized at all. If the patents selected into utilization are not random, for example, if only the highest-quality disruptive patents are eventually commercialized, then the coefficient on disruption in the time-to-utilization regression may be biased. To address this concern, we estimate a Heckman selection model with the utilization equation as the selection equation. The results, reported in the Appendix (Table A.1), are qualitatively similar: the coefficient on disruption in the outcome equation remains positive and statistically significant after accounting for selection, with the inverse Mills ratio providing a modest correction.

4.3 Fact 3: Disruptive Patents Lead to New Product Lines

Facts 1 and 2 establish that disruptive patents face systematic barriers to commercialization. A natural concern is whether our disruption measure truly captures economic novelty or merely reflects noise in the embedding space. We address this by examining what happens when disruptive patents *are* eventually brought to market. If disruption captures the technology that *truly* deviates from existing products, then commercialized disruptive patents should give rise to entirely new product categories—not incremental updates within existing products. This prediction serves both as a validation of our measure's economic content and as a bridge to the mechanism we investigate next: if disruptive patents create new product lines, they also pose the greatest threat to existing ones.

To construct the dependent variable, we leverage the product line classification produced by our LLM pipeline (Section 3.3). For each firm, we sort products chronologically by

announcement date and define:

$$\text{NewLine}_p = \begin{cases} 1 & \text{if product } p \text{ is the first product in its product line for firm } g \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

When multiple products in the same product line are announced on the same date, all are coded as new product lines. In our sample of over 7 million utilized patent–product pairs, 48% of products represent new product lines—a balanced split between new categories and extensions of existing ones.

The relationship between disruption and new product lines aligns with our hypothesis. Figure 4 plots the new product line rate against the disruption score in a 20-bin binscatter. The relationship is positive and monotonic: the new product line rate rises from roughly 39% for the least disruptive patents to over 55% for the most disruptive ($\beta = 0.346$, $R^2 = 0.77$). The gradient steepens at the top of the distribution, consistent with the most disruptive patents representing qualitatively different technological directions rather than marginal departures from existing products.

To formally establish this fact, we estimate the following regression at the utilized patent–product pair level:

$$\text{NewLine}_p = \alpha_g + \alpha_t^{announce} + \alpha_s^{file} + \beta \cdot \text{Disruption}_i + \varepsilon_{ip}, \tag{7}$$

where the unit of observation is a utilized patent–product pair (i, p), α_g are firm fixed effects, $\alpha_t^{announce}$ are product announcement year fixed effects, and α_s^{file} are patent filing year fixed effects. The announcement year fixed effects control for trends in product diversification over time; the filing year fixed effects control for the possibility that patents from different cohorts have systematically different disruption profiles. Standard errors are clustered at the firm level. The key identifying variation comes from comparing, within the same firm, utilized patents with different disruption scores and the products they generate.

Table 4 presents four specifications that progressively absorb potential confounders. Column 1 (OLS without fixed effects) yields a coefficient of 0.345 ($t = 3.19$), close to the binscatter slope, confirming that the unconditional relationship is large. Column 2 adds announcement year fixed effects with little change (0.354, $t = 2.91$). The important shift comes in column 3, which adds firm fixed effects: the coefficient drops to 0.068 ($t = 2.50$), as much of the unconditional relationship reflects cross-firm differences in diversification strategies. The within-firm estimate is the more relevant one: it asks whether, for a given firm, the more disruptive patents are disproportionately associated with new product lines. In our preferred specification (column 4), which includes firm, announcement year, and filing year fixed effects, the coefficient is 0.088 ($t = 3.15$, $p = 0.002$). A one-standard-deviation increase in disruption (0.13) raises the probability of a new product line by approximately 1.1 percentage points, or about 2.4% relative to the baseline rate of 48%.

The within-firm effect is modest in absolute terms, but the result matters for two reasons. First, it validates the economic content of our disruption measure: patents that our algorithm identifies as dissimilar from existing products do in fact lead to genuinely new product categories, not merely minor extensions. Second, it implies that disruptive patents, when commercialized, create new markets—and therefore pose a direct competitive threat to the firm's existing product lines. Whether this threat deters commercialization in the first place is the subject of Fact 4.

4.4 Fact 4: Product Cannibalization Deters Utilization

Why are disruptive patents underutilized? The innovator's dilemma points to a specific mechanism: cannibalization. Commercializing a disruptive technology may create a new market—as Fact 3 confirms—but it also threatens to erode demand for the firm's current products. The stronger a firm's dependence on its existing product portfolio, the more acute this trade-off becomes. If cannibalization is one of the primary operative channels, the utilization penalty for disruptive patents should be concentrated among firms with the most

to lose from disrupting their own revenue streams.

To measure a firm's exposure to cannibalization, we construct $\text{PriorSalesShare}_{g,t}$, the share of firm g's total revenue attributable to business segments linked to products announced before the patent filing date t. The variable combines the LLM-based product-to-segment mapping with Compustat Segments sales data (see Section 3.4 for details). A firm with PriorSalesShare close to one derives almost all of its revenue from existing products and thus has the most to lose from disruption. A firm with a low share has diversified revenue sources and faces a weaker cannibalization incentive.

Figure 5 presents binscatter plots of the utilization rate against the disruption score, separately for firms with above-median (cardinal) and below-median (blue) prior sales share. Each subsample is divided into 20 equal-sized bins of the disruption score, and the average utilization rate is plotted within each bin. Both groups exhibit a negative relationship between disruption and utilization, but the slope is steeper for high prior revenue firms ($\beta = -1.45$) than for low prior revenue firms ($\beta = -1.13$). The two lines diverge as disruption increases, indicating that the cannibalization penalty grows with the degree of disruption.

To formalize this pattern, we estimate the following regression at the patent level:

$$\text{Utilized}_i = \alpha_g + \alpha_t + \beta_1 \cdot \text{Disruption}_i + \beta_2 \cdot \text{PriorSalesShare}_{g,t} + \beta_3 \cdot \text{Disruption}_i \times \text{PriorSalesShare}_{g,t} + \varepsilon_i. \tag{8}$$

The coefficient of interest is β_3, the interaction between disruption and prior sales share. The cannibalization hypothesis predicts $\beta_3 < 0$: firms with higher prior sales shares should exhibit a larger utilization penalty for disruptive patents. We include firm and year fixed effects throughout, with the most stringent specifications replacing year fixed effects with CPC section $\times$ year fixed effects. Standard errors are clustered at the firm level.

Table 5 presents the results across five specifications. Columns 1 and 3 use the continuous PriorSalesShare measure, while columns 2, 4, and 5 use a binary indicator $\text{HighPriorSales}_{g,t}$ equal to one if the firm's prior sales share exceeds the sample median. Column 5 additionally

controls for firm age and its interactions with disruption and prior sales, to address the concern that the result could be driven by systematic differences between young and old firms.

Across all specifications, the interaction term is negative and statistically significant. In column 1 (firm and year fixed effects, continuous interaction), the coefficient on Disruption $\times$ PriorSalesShare is -1.231 ($t = -4.64$). To interpret this magnitude, consider two firms that differ by one standard deviation in prior sales share (0.27). For these two firms, a one-standard-deviation increase in disruption (0.13) reduces utilization by an additional $-1.231 \times 0.27 \times 0.13 \approx 4.3$ percentage points for the firm with higher prior sales—a utilization penalty roughly 43% larger than for the low-prior-sales firm.

The binary specification in column 2 provides a more transparent comparison. The interaction coefficient on Disruption $\times$ HighPriorSales is -0.482 ($t = -2.57$), meaning the negative effect of disruption on utilization is 0.482 larger (in absolute value) for firms with above-median prior sales share. Given the main effect of disruption of -1.336, the total effect for high-prior-sales firms is $-1.336 - 0.482 = -1.818$: a one-standard-deviation increase in disruption reduces utilization by 23.6 percentage points for high-prior-sales firms versus 17.4 percentage points for low-prior-sales firms. The results are robust to CPC $\times$ year fixed effects (columns 3–4), which control for technology-class-specific time trends, and to controlling for firm age (column 5).

These results provide direct evidence for the cannibalization mechanism at the core of the innovator's dilemma. Firms whose revenue depends most heavily on existing product lines are the most reluctant to commercialize their disruptive technologies, even when they hold the patents for these technologies. The result holds within firm and within technology class $\times$ year, ruling out explanations based on industry composition or aggregate trends. In the next section (Section 5), we formally rule out two prominent alternative explanations—financial market undervaluation and the absence of complementary technologies—and confirm that the cannibalization channel is the primary driver. Before that, we document one further

consequence of the cannibalization incentive: its effect on the upstream R&D process.

4.5 Fact 5: R&D Portfolio Distortion

Fact 4 shows that cannibalization concerns deter the commercialization of disruptive patents. But a forward-looking firm that anticipates shelving its disruptive inventions may go further: it may redirect research *itself* away from disruptive technologies in the first place. Rather than investing in breakthroughs only to leave them on the shelf, such a firm would rationally tilt its R&D portfolio toward safer, more incremental innovations that complement—rather than threaten—its existing product lines. If this anticipatory channel is operative, the innovator's dilemma does not merely suppress the deployment of existing inventions; it distorts the direction of inventive activity upstream, and the full cost of cannibalization is larger than what the patent-to-product data alone reveal.

We test for this upstream distortion by asking whether firms file less disruptive patents in years when they have more revenue at stake. Figure 6 presents a binscatter plot of the mean disruption score of patents filed in a firm-year against the firm's prior sales share, after residualizing both variables on firm and year fixed effects. The sample is divided into 20 equal-sized bins of the residualized prior sales share, and the average residualized mean disruption is plotted within each bin. The relationship is negative and approximately linear ($\beta = -0.044$, $R^2 = 0.65$): within the same firm, years with higher prior sales shares are associated with less disruptive patent portfolios.

To formalize this relationship, we aggregate the patent-level data to the firm-year level and estimate:

$$\overline{\text{Disruption}}_{g,t} = \alpha_g + \alpha_t + \beta \cdot \text{PriorSalesShare}_{g,t} + \gamma' X_{g,t} + \varepsilon_{g,t}, \tag{9}$$

where $\overline{\text{Disruption}}_{g,t}$ is the mean disruption score of all patents filed by firm g in year t. The sample includes firm-years with at least two patents filed, to ensure that the mean is not

determined by a single patent. The coefficient β is identified from within-firm variation: it asks whether, in years when a firm derives a higher share of revenue from existing products, the patents it files are less disruptive. In some specifications, we add controls $X_{g,t}$, including log total sales, to ensure that the result is not driven by scale effects.

Table 6 reports the results across six specifications. Column 1 is a simple OLS regression, yielding a coefficient of -0.077 ($t = -8.04$). This estimate reflects both within- and between-firm variation. Column 2 adds year fixed effects, and column 3 adds both firm and year fixed effects. The preferred within-firm specification (column 3) yields a coefficient of -0.026 ($t = -2.30$, $p = 0.023$)—smaller than the OLS estimate, as expected, since between-firm differences in R&D orientation are absorbed. Column 4 adds log total sales as a control, with the coefficient essentially unchanged (-0.025, $t = -2.28$), confirming that the result is not driven by overall firm size. Columns 5 and 6 use alternative measures: a binary indicator for above-median prior sales share and log segment sales. These measures lack the granular variation of the continuous prior sales share, but the point estimates are directionally consistent.

To interpret the magnitude, a one-standard-deviation increase in prior sales share (0.27) is associated with a decrease of $0.026 \times 0.27 = 0.007$ in mean disruption, approximately 8% of the standard deviation of mean disruption (0.088). While modest, this effect is identified purely from within-firm variation and captures only the shift in the *average* disruptiveness of the patent portfolio. The effect may be larger at the tails—firms may be particularly unlikely to file highly disruptive patents when prior sales share is high—and understates the total distortion if firms also reduce the *number* of disruptive research projects initiated, a margin we cannot observe in the patent data.

Taken together, Facts 4 and 5 indicate that the cannibalization incentive operates at both ends of the innovation procedure. Firms with more to lose from disruption not only underutilize their disruptive patents (Fact 4) but also generate fewer disruptive patents in the first place. The total effect of cannibalization concerns on the deployment of disruptive

technologies is therefore larger than what any single margin—utilization, timing, or R&D direction—reveals on its own.

5 Mechanisms Behind the Innovator's Dilemma

In the previous section, we documented five facts consistent with the innovator's dilemma: disruptive patents are underutilized (Fact 1), delayed (Fact 2), and channel into new product lines (Fact 3), with the utilization penalty amplified by the value of existing products (Fact 4) and the distortion extending upstream into R&D portfolio composition (Fact 5). Our interpretation is that product cannibalization concerns—the risk that a disruptive technology erodes demand for the firm's existing products—are the primary driver of these patterns.

However, the patterns in Facts 1–3 are also consistent with alternative mechanisms that do not involve strategic cannibalization. In this section, we consider two prominent alternatives. The first is that the financial market systematically undervalues disruptive patents at issuance, discouraging firms from investing in their commercialization. The second is that disruptive patents lack the complementary technologies needed for commercialization, so that the utilization gap and delay reflect technological constraints rather than strategic choices. For each alternative, we lay out testable predictions, describe the data and specification, and show that the evidence does not support the alternative as the primary driver. We also present specifications that simultaneously control for both alternatives, confirming that the disruption penalty is robust.

5.1 Ruling Out Market Uncertainty

Motivation One natural concern is that disruptive patents are less likely to be utilized not because of cannibalization incentives, but because the financial market undervalues them, discouraging further investment in commercialization. This concern is directly motivated by Lyu and Xu (2025), who document that investors underreact to novel patent issuance:

novel patents predict positive long-term stock returns, suggesting that their value is not fully reflected in prices at issuance. If the market systematically undervalues disruptive technologies, then disappointing stock returns around patent issuance could reduce management's willingness to invest in bringing these technologies to market.

Several channels could transmit market undervaluation into real commercialization decisions. First, managers may be uncertain about the true value of their own patents and may learn from the stock market price, as in Goldstein (2023). If the market undervalues a disruptive patent, the manager may infer that the technology has lower commercial potential than initially expected and choose not to invest in its development. Second, firms that depend on external equity financing face a cost-of-capital channel (Baker et al., 2003): if the market undervalues the firm's patent portfolio, the firm's stock price is depressed, raising the cost of equity and making it harder to fund commercialization projects. Third, managers subject to career concerns or stock-based compensation may respond directly to short-term stock prices, even if they privately believe the market is wrong. In all three cases, market undervaluation of disruptive patents could generate the patterns documented in Facts 1–2 without any role for product cannibalization.

Testable predictions If market undervaluation were the primary channel, we would expect two specific empirical patterns:

1. *Disruptive patents should receive lower market valuations at issuance.* The stock market response to the grant of a disruptive patent should be weaker (or negative) relative to non-disruptive patents. If the market correctly values both types, or even overvalues disruptive patents, then the market undervaluation channel cannot explain the utilization gap.

2. *Controlling for market valuation should attenuate the effect of disruption on utilization and time to utilization.* If market undervaluation is the mediating variable—i.e.,

disruption reduces utilization *because* it reduces market valuation, which in turn discourages commercialization—then adding market valuation as a control should reduce the coefficient on disruption. If the coefficient is unaffected, the market channel is not the primary mechanism.

Measuring market valuation We measure market valuation using the KPSS patent value measure from Kogan et al. (2017). This measure estimates the private value of each patent by computing the change in the firm's stock market capitalization in a three-day window around the patent grant date, adjusted for the number of patents granted on that date and for market-wide movements. The KPSS measure has been widely used in the innovation literature as a proxy for the economic value of patents as perceived by the stock market. It captures investors' assessment of the patent's contribution to firm value at the time of issuance—precisely the signal that would influence managers who learn from or respond to market prices.

Specification We estimate three sets of regressions, reported in Table 7. The first set (columns 1–2) directly tests Prediction 1 by regressing log KPSS value on disruption:

$$\ln(\text{KPSS}_i) = \alpha_g + \alpha_t + \beta \cdot \text{Disruption}_i + \varepsilon_i, \tag{10}$$

where KPSS_i is the KPSS patent value for patent i. If the market undervalues disruptive patents, we expect $\beta < 0$. Column 1 includes firm and filing year fixed effects; column 2 replaces filing year fixed effects with CPC section $\times$ year fixed effects.

The second set (columns 3–5) tests Prediction 2 for the time-to-utilization outcome by adding log KPSS as a control:

$$\ln(\text{TimeToUtil}_i) = \alpha_g + \alpha_t + \beta_1 \cdot \text{Disruption}_i + \beta_2 \cdot \ln(\text{KPSS}_i) + \varepsilon_i. \tag{11}$$

Column 3 is the baseline without KPSS (reproducing Fact 2). Column 4 adds log KPSS with firm and year fixed effects. Column 5 uses CPC × year fixed effects with log KPSS. If market valuation mediates the disruption effect, β_1 should decline when β_2 is added.

The third set (columns 6–7) tests Prediction 2 for the utilization outcome:

$$\text{Utilized}_i = \alpha_g + \alpha_t + \beta_1 \cdot \text{Disruption}_i + \beta_2 \cdot \ln(\text{KPSS}_i) + \varepsilon_i. \tag{12}$$

Column 6 is the baseline without KPSS. Column 7 adds log KPSS. All specifications include firm fixed effects, and standard errors are clustered at the firm level.

Market valuation of disruptive patents Columns 1–2 of Table 7 show that the market does not systematically undervalue disruptive patents at issuance. The coefficient on disruption is 0.070 ($t = 0.42$) with firm and year fixed effects and 0.055 ($t = 0.58$) with firm and CPC × year fixed effects. Both estimates are statistically insignificant and, if anything, slightly positive—the point estimates suggest that the market values disruptive patents marginally *more* than non-disruptive ones, though the difference is not statistically distinguishable from zero.

This null result is worth discussing in light of the findings in Lyu and Xu (2025), who document significant underpricing of novel patents. The difference arises because the two studies use distinct measures. Lyu and Xu (2025) define novelty using aggregate textual similarity to all patents filed in the five years preceding the focal patent's filing year: a patent is novel if it has low textual similarity to this prior patent corpus, capturing how different a patent is from the existing technological frontier. Our measure of patent disruption is defined relative to the firm's specific product portfolio; it captures whether a patent is similar to products the firm already sells, rather than to other patents. A patent can be novel in the text-based sense (a true scientific breakthrough) without being disruptive to the firm's product portfolio (if the firm specializes in selling products embedded with the

novel technology in the relevant domain). Conversely, a patent can be highly disruptive to the firm's existing products while being technologically incremental (e.g., an extension of a competitor's technology into the firm's product space). The market appears to misprice text-based technological novelty but not product-based disruption, suggesting that the two dimensions of patent novelty have distinct implications for market valuation.

The high R^2 values in columns 1–2 (0.819 and 0.832) are also informative. The firm and year (or CPC $\times$ year) fixed effects absorb most of the variation in log KPSS, which is consistent with the finding that patent values are largely determined by firm-level and technology-class-level characteristics. After absorbing these fixed effects, the residual variation in KPSS is not systematically related to disruption.

Controlling for market valuation in time-to-utilization regressions Columns 3–5 of Table 7 test whether market valuation mediates the effect of disruption on time to utilization. In column 3 (baseline without KPSS), the disruption coefficient is 0.732 ($t = 8.30$). In column 4 (adding log KPSS), the coefficient is essentially unchanged at 0.731 ($t = 8.32$). The change is 0.001—less than 0.2% of the baseline coefficient. The coefficient on log KPSS itself is 0.008 ($t = 0.54$), economically tiny and statistically insignificant. Adding log KPSS does not improve the R^2 (0.237 in both columns). In column 5, with CPC $\times$ year fixed effects and KPSS, the disruption coefficient is 0.864 ($t = 12.6$)—larger than the baseline, as in Fact 2. The KPSS coefficient is 0.010 ($t = 0.49$), again insignificant.

These results are inconsistent with the mediation hypothesis. If the market undervaluation channel were operative, adding KPSS should absorb some of the variation currently attributed to disruption, reducing the disruption coefficient. Instead, the coefficient is unchanged. The market's assessment of patent value at issuance does not predict the speed of commercialization, conditional on disruption and firm and year fixed effects.

Controlling for market valuation in utilization regressions Columns 6–7 repeat the exercise with the utilization indicator as the dependent variable. In column 6 (baseline without KPSS), the disruption coefficient is -1.460 ($t = -17.6$). In column 7 (adding log KPSS), the coefficient is -1.459 ($t = -17.9$)—a change of 0.001, or less than 0.1% of the baseline. Controlling for KPSS has essentially zero effect on the disruption–utilization relationship.

The KPSS coefficient in column 7 is -0.021 ($t = -2.59$), which is statistically significant but economically small. A one-standard-deviation increase in log KPSS (approximately 1.5) is associated with a $0.021 \times 1.5 = 3.2$ percentage point decrease in utilization probability. The sign is negative, which is counterintuitive if one expected higher-valued patents to be more likely to be commercialized. One interpretation is that KPSS captures market "excitement" around the patent grant, and patents that generate market excitement may be in domains where the firm is less likely to have follow-through products. Regardless, the disruption coefficient is unaffected by the inclusion of KPSS.

Discussion The evidence from Table 7 rules out market undervaluation as the primary mechanism behind the innovator's dilemma patterns. The market does not assign lower values to disruptive patents (columns 1–2), and controlling for market valuation does not attenuate the effect of disruption on either time to utilization (columns 3–5) or the utilization probability (columns 6–7). While we cannot rule out that market misperceptions play *some* role in firms' innovation decisions—indeed, Lyu and Xu (2025) provide compelling evidence that they do for text-based technological novelty—the specific patterns documented in Facts 1–5 are not driven by the market's assessment of patent value. The disruption penalty for utilization and time to utilization persists fully after controlling for the market's valuation at issuance, indicating that the firm's commercialization decisions are driven by internal strategic considerations rather than external market signals.

5.2 Ruling Out Complementary Technology

Motivation A second alternative is that disruptive patents are underutilized and delayed not because of strategic cannibalization, but because they lack the complementary technologies needed for commercialization. The idea that innovation often requires a package of complementary assets is well established (Teece, 1986). A disruptive technology may be technically sound and commercially valuable in principle, but if the supporting innovations—complementary hardware, software, manufacturing processes, or distribution infrastructure—do not yet exist, the technology cannot be profitably deployed. Under this interpretation, the delay in utilizing disruptive patents reflects a technological constraint rather than a strategic choice.

This alternative is particularly important because it could potentially explain not just Facts 1–2 but also the interaction with prior sales share documented in Fact 4. Firms with large existing product lines may have developed their technologies in established domains where complementary infrastructure is mature, while disruptive technologies targeting new domains may face a "chicken-and-egg" problem: no firm invests in complementary infrastructure until the focal technology is proven, and the focal technology cannot be commercialized until the infrastructure exists. If firms with high prior sales shares tend to hold patents in more mature technology domains, the interaction between disruption and prior sales share could reflect differences in the complementary technology landscape rather than cannibalization incentives.

Testable predictions If the complementary technology channel were the primary driver, we would expect:

1. *Disruptive patents should have fewer ready complements.* Disruptive patents, being further from the firm's existing technology, should be in domains where fewer complementary technologies are available at the time of filing.

2. *Controlling for complementary technology availability should attenuate the disruption coefficient.* If the delay is primarily due to the lack of complements, adding measures of complement availability should reduce the disruption coefficient substantially.

Measuring complementary technology We exploit the patent-to-product linkages in our dataset to construct a revealed-preference measure of complementary technologies. The key observation is that products typically draw on multiple patents simultaneously: a smartphone, for example, relies on patents covering the display, processor, wireless communication, and battery technologies. We define the set of complementary patents for a focal patent i used in product p as the set of all *other* patents that are also utilized in the same product p. These are the technologies that, in practice, must be combined with patent i to produce a commercially viable product. When a patent appears in multiple products, we average the complement measures across all products in which it is used. We construct two measures:

- ComplReady_i: the fraction of complementary patents filed *before* the focal patent i's filing date. This captures the extent to which the technologies needed to commercialize patent i were already available when it was filed. A value close to one indicates that most co-utilized patents predate the focal patent, suggesting that the complementary ecosystem was already in place. A value close to zero indicates that the focal patent was filed before many of its eventual complements, suggesting that the technology may have been "ahead of its time." In our sample, the mean ComplReady is 0.50, with substantial cross-patent variation (standard deviation of 0.27).

- $\ln(N_{\text{Comp},i})$: the log number of complementary patents identified for focal patent i, regardless of their filing timing. This captures the complexity of the technology ecosystem in which the patent operates. Patents with many co-utilized patents are embedded in products that draw on a broader set of technologies.

Specification We estimate a series of regressions reported in Table 8, all with log time to utilization as the dependent variable and restricted to utilized patents:

$$\ln(\text{TimeToUtil}_i) = \alpha_g + \alpha_t + \beta_1 \cdot \text{Disruption}_i + \beta_2 \cdot \text{ComplReady}_i + \beta_3 \cdot \ln(N_{\text{Comp},i}) + \beta_4 \cdot \ln(\text{KPSS}_i) + \varepsilon_i. \tag{13}$$

We build up the specification across six columns to show how the disruption coefficient responds to the inclusion of controls. Column 1 is the baseline with only disruption, firm, and year fixed effects. Column 2 adds ComplReady. Column 3 adds both complements measures. Column 4 replaces year fixed effects with CPC section $\times$ year fixed effects. Columns 5 and 6 add log KPSS on top of the complements controls. Standard errors are clustered at the firm level throughout.

Results Table 8 yields two findings.

First, complementary technology availability is a strong predictor of commercialization speed. The coefficient on ComplReady is approximately -2.18 across all specifications (columns 2–6), with t-statistics ranging from -14.7 to -16.5. A one-standard-deviation increase in ComplReady (0.27) is associated with approximately 28% shorter time to utilization. This is a large effect, confirming that the availability of co-utilized technologies matters for when patents reach the market. The inclusion of ComplReady increases R^2 from 0.239 (column 1) to 0.440 (column 2), indicating that the complements measures explain roughly 20 additional percentage points of the variation in log time to utilization after firm and year fixed effects.

The coefficient on $\ln(N_{\text{Comp}})$ is positive and significant (approximately 0.14–0.16, $t \approx 2.5$–2.9). This may appear counterintuitive: patents with more co-utilized patents take *longer* to be commercialized, holding constant the fraction that is ready. One interpretation is that patents embedded in products requiring many complementary technologies face greater integration complexity, even when many of the individual complements are available.

Second, and more importantly, controlling for complement readiness does *not* attenuate the effect of disruption. The disruption coefficient is 0.721 in column 1 (baseline) and increases to 0.791 in column 2 (adding ComplReady), 0.796 in column 3, and 0.887 in column 4 (CPC $\times$ year fixed effects). In the most saturated specification (column 6), which simultaneously controls for complements, KPSS, and CPC $\times$ year fixed effects, the disruption coefficient is 0.900 ($t = 13.3$)—25% larger than the baseline.

The fact that the disruption coefficient *increases* after controlling for complements is informative. It implies that disruption and complement readiness are positively correlated: more disruptive patents tend to have a lower fraction of their co-utilized technologies already filed at the time of the focal patent's filing. Controlling for ComplReady removes this source of negative confounding, revealing a larger "pure" disruption effect. Part of the reason disruptive patents take longer to commercialize is indeed that their complementary technologies are less ready—this is the variation absorbed by ComplReady. But there is a large residual component of the delay that complement readiness cannot explain. Consider a thought experiment: take two patents with identical levels of complement readiness, held by the same firm, filed in the same year. The one with higher disruption—the one more dissimilar from the firm's existing products—still takes significantly longer to reach the market. This within-firm, within-complement-readiness variation is inconsistent with the complementary technology explanation and consistent with the cannibalization mechanism.

Controlling for both alternatives simultaneously Columns 5–6 present the most saturated specifications, controlling for both complements and market valuation simultaneously. In column 5 (firm and year fixed effects), the disruption coefficient is 0.809 ($t = 9.82$). In column 6 (CPC $\times$ year fixed effects), it is 0.900 ($t = 13.3$). Neither market valuation nor complementary technology—separately or jointly—can account for the disruption penalty. The KPSS coefficient is small and insignificant in both specifications, consistent with the previous subsection. The R^2 in column 6 is 0.502, meaning that the full set of controls

explains roughly half of the variation in log time to utilization.

5.3 The Cannibalization Channel

Having ruled out financial market undervaluation and complementary technology constraints, we return to the maintained hypothesis: product cannibalization is the primary driver of the patterns documented in Facts 1–5.

Summary of evidence The evidence rests on three pillars. First, the mechanism tests in Sections 5.1 and 5.2 rule out the two most prominent alternatives. The market does not systematically undervalue disruptive patents at issuance, and controlling for market valuation does not attenuate the disruption penalty. Complementary technology availability is a strong predictor of commercialization timing, but the disruption coefficient actually *increases* by 25% after controlling for complements, indicating that the baseline estimate is biased downward by omitted variable bias. The residual disruption effect is 0.900 ($t = 13.3$) in the most saturated specification.

Second, Facts 4 and 5 provide positive evidence for the cannibalization channel rather than simply ruling out alternatives. Fact 4 shows that the utilization penalty is amplified in firms whose revenue depends most heavily on existing products ($\beta_3 = -1.231$, $t = -4.64$). This is the pattern predicted by the cannibalization hypothesis and is difficult to rationalize under alternative mechanisms that do not involve the firm's product portfolio. Fact 5 shows that the cannibalization incentive extends upstream into R&D: firms with higher prior sales shares file less disruptive patents (-0.026, $t = -2.30$), suggesting that firms anticipate the commercialization costs and adjust their research direction accordingly.

Third, Fact 3 provides supporting evidence by showing that disruptive patents, when eventually commercialized, disproportionately give rise to new product lines (0.088, $t = 3.15$). Disruptive technologies represent new categories that could substitute for or render obsolete the firm's current offerings, which is precisely why firms resist commercializing them.

Connection to the literature The cannibalization mechanism we document is a central element of the innovator's dilemma as articulated by Christensen (2015), who argued that incumbent firms fail to adopt disruptive technologies because these technologies initially serve different market segments and threaten to undermine the firm's existing business. Our contribution is to provide systematic, large-sample quantitative evidence for this mechanism using patent-to-product data, complementing the case study evidence in Christensen (2015) and the industry-specific analyses in subsequent work.

Our findings also relate to the literature on creative destruction and the "replacement effect" (Arrow, 1962; Aghion et al., 2005). In Arrow's framework, a monopolist has weaker incentives to innovate than a competitive firm because the monopolist internalizes the cannibalization of its own rents. Our Fact 4 provides a product-level analog: firms whose existing products account for a larger share of revenue face stronger replacement effects and are less willing to commercialize disruptive technologies. The heterogeneity across firms with different prior sales shares is consistent with models of endogenous innovation where the intensity of creative destruction depends on the incumbent's market position (Aghion et al., 2001; Akcigit and Kerr, 2018).

The R&D distortion documented in Fact 5 connects to the literature on directed technical change (Acemoglu, 2002). In this framework, firms choose the direction of their innovation effort based on expected profitability, and market size effects influence which technologies are developed. Our finding that firms with higher prior sales shares file less disruptive patents can be interpreted through this lens: the expected profitability of disruptive innovation is lower for firms with large existing product lines, leading them to direct their R&D toward incremental improvements of existing technologies.

Firm-level evidence and aggregate implications While our analysis is conducted at the patent and firm level, the patterns we document have aggregate implications for the pace and direction of technological progress. If incumbent firms systematically under-utilize

and under-invest in disruptive technologies, then the commercialization of breakthroughs may depend disproportionately on new entrants who do not face the cannibalization trade-off. This is consistent with the empirical finding in Akcigit and Ates (2023) that the share of innovation by young firms has declined over time, and with the broader concern about declining business dynamism in the US economy (Decker et al., 2016). Our data, by linking patents to products within firms, provides a micro-foundation for these aggregate patterns: we observe directly that incumbent firms hold the patents for disruptive technologies but choose not to commercialize them.

5.4 Other Potential Mechanisms

While our evidence supports the cannibalization channel, we acknowledge that other mechanisms may also contribute to the underutilization of disruptive patents. We discuss several additional channels that are consistent with the patterns in our data and that represent directions for future research.

Managerial inertia and organizational rigidity A large literature in organizational economics argues that incumbent firms face organizational barriers to adopting disruptive technologies, even when the economic case for adoption is clear. Henderson and Clark (1990) show that established firms struggle with "architectural" innovations that reconfigure the linkages between components, because the firm's organizational structure and routines are optimized for the existing technology. Tripsas (1997) document that the typesetter industry's incumbents failed to adopt digital technology despite having the requisite technical capabilities, suggesting that organizational inertia rather than technological incompetence was the binding constraint.

In our context, managerial inertia could manifest as a reluctance to reorganize production lines, retrain the workforce, or restructure distribution channels. This mechanism is distinct from cannibalization—a firm could face organizational barriers to commercializing

a disruptive patent even if the new product does not directly compete with existing ones. However, the two mechanisms are likely mutually reinforcing: firms with valuable existing products have both a stronger cannibalization incentive *and* more entrenched organizational structures, making it difficult to empirically separate the two channels. The fact that our Fact 4 interaction is identified from within-firm variation (comparing the same firm in periods with different prior sales shares) provides some reassurance that the result is not solely driven by time-invariant organizational characteristics, but we cannot fully rule out time-varying organizational rigidity as a contributing factor.

Financial constraints and investment frictions Firms may underutilize disruptive patents because they face financial constraints that prevent them from funding the commercialization of new, uncertain technologies. Commercially deploying a disruptive technology typically requires large upfront investments in new production facilities, marketing, and distribution, with uncertain and potentially distant payoffs. Financially constrained firms may prefer to invest in incremental improvements to existing products, which offer more predictable returns.

This channel is related to, but distinct from, the market undervaluation channel tested in Section 5.1. The market undervaluation channel operates through the stock price signal, while the financial constraints channel operates through the availability of internal or external funds. A firm could face binding financial constraints even if the market correctly values its patents, if for example it lacks sufficient internal cash flow and faces high costs of external financing (Hall, 2002; Brown et al., 2009). In our setting, if financial constraints were the primary driver, we would expect the disruption penalty to be larger for firms with lower cash reserves or higher leverage. We do not formally test these predictions in the current paper, but note that the firm fixed effects in our specifications absorb time-invariant differences in financial constraints across firms.

Managerial short-termism and agency costs Managers evaluated on short-term financial performance may rationally avoid commercializing disruptive technologies whose payoffs are long-term and uncertain (Stein, 1989; Edmans, 2009). If managers face pressure to meet quarterly earnings targets, they may prefer incremental products that generate predictable near-term revenue rather than disruptive products whose commercial success is uncertain.

This mechanism is closely related to the findings in Lyu and Xu (2025), who show that stock market misreactions to patent novelty influence firms' subsequent innovation directions. In our context, a manager contemplating whether to commercialize a disruptive patent may anticipate that the new product will initially underperform existing products, generating negative earnings surprises and stock price pressure. The combination of cannibalization and short-termism could be especially effective in deterring commercialization.

Consumer acceptance and demand uncertainty Firms may delay commercializing disruptive technologies because they are uncertain about consumer acceptance. A disruptive product differs significantly from the firm's existing offerings and may require consumers to change their behavior, learn new usage patterns, or accept initially inferior performance in some dimensions. Firms facing such uncertainty may rationally delay commercialization until more information about demand becomes available.

This demand-side mechanism is particularly relevant for our research agenda. Our planned integration of Consumer Reports product ratings and NielsenIQ sales data will allow us to directly measure consumer evaluations and observe product-level sales trajectories, enabling us to test whether disruptive products cannibalize existing product sales as predicted by the innovator's dilemma. These extensions will help decompose the innovator's dilemma into its supply-side (cannibalization and organizational inertia) and demand-side (consumer acceptance) components.

6 Conclusion

In this paper, we construct a novel dataset linking firm patents to individual products using large language models and semantic embeddings. The dataset is general purpose, covering all technology classes and industries, and is validated against ground-truth patent–product links from Virtual Patent Markings. The LLM-based pipeline involves product extraction from corporate news, product line classification, business segment mapping, and semantic similarity computation—demonstrating that modern language models can be used to construct firm-level datasets linking innovation inputs (patents) to innovation outputs (products) at a scale and generality that was previously infeasible.

Using this dataset, we document five stylized facts that collectively provide evidence for the innovator's dilemma in incumbent firms. First, disruptive patents—those most dissimilar from a firm's existing product portfolio—are significantly less likely to be commercialized into products. Second, conditional on eventual utilization, disruptive patents take substantially longer to reach the market, with a one-standard-deviation increase in disruption associated with roughly one additional year of delay. Third, when disruptive patents are commercialized, they disproportionately give rise to entirely new product lines rather than extensions of existing ones, validating the economic content of our disruption measure. Fourth, the utilization penalty for disruption is amplified in firms whose revenue depends most heavily on existing product lines, providing direct evidence of the product cannibalization channel. Fifth, cannibalization concerns distort not only commercialization decisions but also upstream R&D: firms with higher prior sales shares file systematically less disruptive patents. We rule out two prominent alternative mechanisms—financial market undervaluation and the absence of complementary technologies—showing that neither can account for the patterns we document.

Our findings contribute to the growing literature on the use of AI and large language models for economic measurement. The patent-to-product dataset fills an important gap

between the large literature on patents as measures of innovation input and the smaller literature on new products as measures of innovation output. By linking these two sides of the innovation process at the firm level, we can directly study the "black box" of how firms translate their R&D efforts into commercial products—a question that has been largely inaccessible to empirical research due to data limitations. The fact that our approach works across technology classes and industries, and can be validated against an independent ground-truth dataset, suggests that it may be a useful building block for future research on innovation and industrial organization.

Several directions for ongoing and future work follow naturally. First, we plan to supplement the product data with Consumer Reports product ratings and test results, as well as NielsenIQ sales and pricing data, to study the consumer-side view of the innovator's dilemma. Consumer Reports buys and independently tests approximately 9,000 products, providing product specifications, features, ratings, reviews, and expert test results. The unique feature of consumer reviews and expert test ratings allows us to examine whether novel technologies are initially less well received by consumers or expert evaluators, providing a demand-side mechanism for why firms delay commercializing disruptive patents. NielsenIQ provides actual product-level sales, quantities, and prices in the consumer packaged goods sector, which would allow us to study the revenue consequences of commercializing disruptive technologies and to directly quantify cannibalization effects.

Second, we aim to develop a structural model of the firm's joint R&D and commercialization decision. The five facts documented in this paper characterize the equilibrium relationship between disruption, utilization, and prior product value, but they do not pin down the welfare consequences of the innovator's dilemma. A structural model that allows firms to choose both the direction of their R&D (the composition of disruptive vs. incremental patents) and the timing of commercialization (when to bring a patented technology to market) would enable us to quantify the welfare losses arising from innovation misallocation driven by cannibalization concerns. One promising approach is to model the firm as facing a

dynamic trade-off between short-term cannibalization costs and long-term competitive benefits of commercializing disruptive technologies, with the key friction being that incumbent firms internalize the cannibalization externality while potential entrants do not.

Third, building on the evidence in Lyu and Xu (2025) that stock market misreactions to patent novelty distort firms' innovation directions, we plan to examine how financial market misperceptions interact with product market cannibalization incentives. Our mechanism tests rule out market undervaluation as the primary driver of the innovator's dilemma patterns, but this does not mean that financial market frictions are irrelevant. It is possible that market misperceptions *amplify* the cannibalization incentive: if the market undervalues disruptive technologies, then firms face both a product market cost (cannibalization) and a financial market cost (short-term stock price penalties) of commercializing disruption. Understanding the joint role of these two channels in shaping firms' innovation trajectories is an important direction for future research.

References

Acemoglu, D. (2002). Directed technical change. *The review of economic studies 69*(4), 781–809.

Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt (2005). Competition and innovation: An inverted-u relationship. *The quarterly journal of economics 120*(2), 701–728.

Aghion, P., C. Harris, P. Howitt, and J. Vickers (2001). Competition, imitation and growth with step-by-step innovation. *The Review of Economic Studies 68*(3), 467–492.

Akcigit, U. and S. T. Ates (2023). What happened to us business dynamism? *Journal of Political Economy 131*(8), 2059–2124.

Akcigit, U. and W. R. Kerr (2018). Growth through heterogeneous innovations. *Journal of Political Economy 126*(4), 1374–1443.

Argente, D., S. Baslandze, D. Hanley, and S. Moreira (2025). Patents to products: Product innovation and firm dynamics. Working Paper.

Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In *Readings in industrial economics: Volume two: Private enterprise and state intervention*, pp. 219–236. Springer.

Baker, M., J. C. Stein, and J. Wurgler (2003). When does the market matter? stock prices and the investment of equity-dependent firms. *The quarterly journal of economics 118*(3), 969–1005.

Brown, J. R., S. M. Fazzari, and B. C. Petersen (2009). Financing innovation and growth: Cash flow, external equity, and the 1990s r&d boom. *The journal of finance 64*(1), 151–185.

Christensen, C. M. (2015). *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Review Press.

de Rassenfosse, G. (2020). Linking products to patents using virtual patent marking. Mimeo, Ecole polytechnique fédérale de Lausanne.

de Rassenfosse, G. and L. Zhou (2020). Patents and supra-competitive prices: Evidence from consumer products. Available at SSRN 3756359.

Decker, R. A., J. Haltiwanger, R. S. Jarmin, and J. Miranda (2016). Declining business dynamism: What we know and the way forward. *American Economic Review 106*(5), 203–207.

Edmans, A. (2009). Blockholder trading, market efficiency, and managerial myopia. *The journal of finance 64*(6), 2481–2513.

Goldstein, I. (2023). Information in financial markets and its real effects. *Review of Finance 27*(1), 1–32.

Hall, B. H. (2002). The financing of research and development. *Oxford review of economic policy 18*(1), 35–51.

Harford, J., Q. He, and B. Qiu (2025). The value of corporate patent utilization. Available at SSRN 5107971.

Henderson, R. M. and K. B. Clark (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative science quarterly*, 9–30.

Igami, M. (2017). Estimating the innovator's dilemma: Structural analysis of creative destruction in the hard disk drive industry, 1981–1998. *Journal of Political Economy 125*(3), 798–847.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021). Measuring technological innovation over the long run. *American Economic Review: Insights 3*(3), 303–320.

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

Lyu, T. and Z. W. Xu (2025). Taking the road less traveled? market misreaction and firm innovation directions. Available at SSRN 4758236.

Mukherjee, A., B. Pellegrino, Y. Ren, T. Thornquist, and A. Žaldokas (2025). New products. Working Paper.

Stein, J. C. (1989). Efficient capital markets, inefficient firms: A model of myopic corporate behavior. *The quarterly journal of economics 104*(4), 655–669.

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research policy 15*(6), 285–305.

Tripsas, M. (1997). Unraveling the process of creative destruction: Complementary assets and incumbent survival in the typesetter industry. *Strategic management journal 18*(S1), 119–142.

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

Figure 1: Product–patent similarity distributions by CPC section

This figure plots the distribution of cosine similarity between product–patent pairs, separately for utilized pairs (orange) and non-utilized pairs (blue), across the eight CPC sections. Cosine similarity is computed using sentence embeddings (all-MiniLM-L6-v2) of product and patent descriptions. Each panel corresponds to one CPC section. The dashed vertical line marks the CPC-section-specific threshold used to classify a patent as utilized by a given product; the threshold value is shown above the line. Utilized pairs shift rightward relative to non-utilized pairs in every section, confirming that the similarity measure captures meaningful product–patent linkages. Thresholds range from 0.21 (Mechanical Engineering) to 0.34 (Human Necessities), reflecting cross-section variation in the textual overlap between patent and product language.

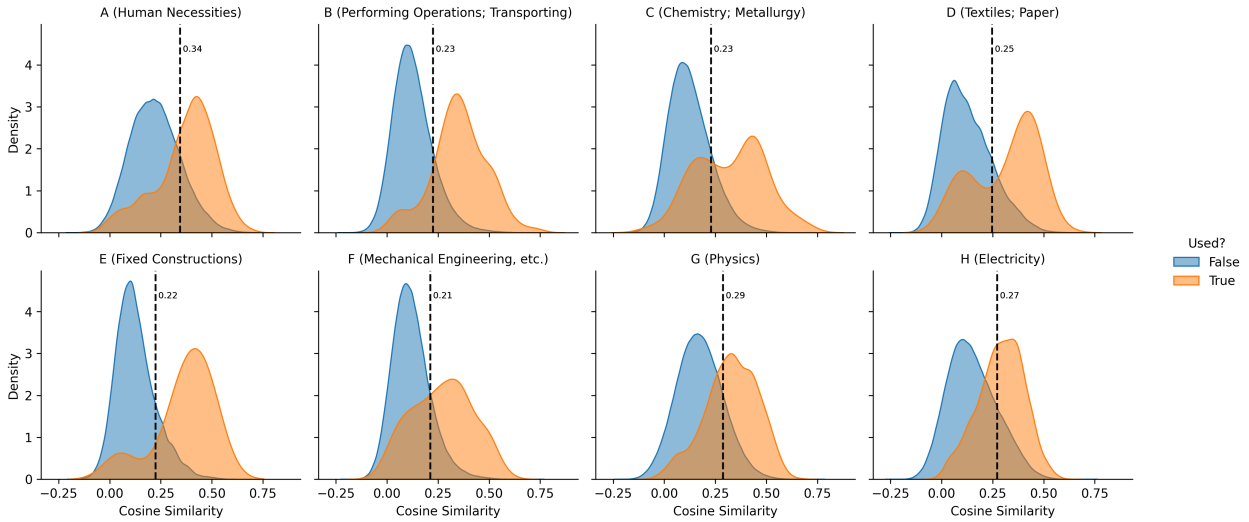


Figure 2: Disruptive Patents Are Less Likely to Be Utilized

This figure presents a binscatter plot of the patent utilization rate against the disruption score, after residualizing both variables on firm and filing year fixed effects. We estimate $\text{Utilized}_i = \alpha_g + \alpha_t + \beta \cdot \text{Disruption}_i + \varepsilon_i$, where $\text{Utilized}_i = 1$ if the patent is utilized in any product announced after its filing date and $\text{Disruption}_i = 1 - \max_{j \in \mathcal{P}_g^{prior}} \text{CosSim}(i, j)$. The sample includes all patents in our dataset. The figure divides the residualized disruption scores into 20 equal-sized bins and plots the average residualized utilization rate within each bin. The strong negative and approximately linear relationship indicates that, within the same firm and year, patents with higher disruption scores are substantially less likely to be commercialized.

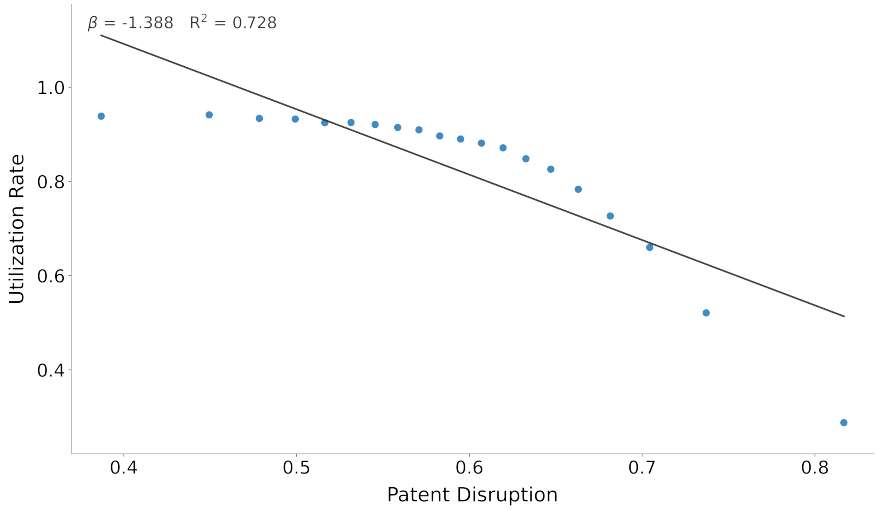


Figure 3: Disruptive Patents Take Longer to Utilize

This figure plots empirical survival curves for time to first product utilization among utilized patents, separately by disruption tercile. The horizontal axis measures years from patent filing date; the vertical axis measures the fraction of patents not yet utilized. Patents are sorted into three terciles based on their disruption score: low, medium, and high. The curves fan out sharply: patents in the highest disruption tercile have a median time to utilization that is roughly 3 years longer than patents in the lowest tercile, consistent with the regression evidence in Table 3. The separation between curves is visible from the earliest horizons and persists throughout the observation window.

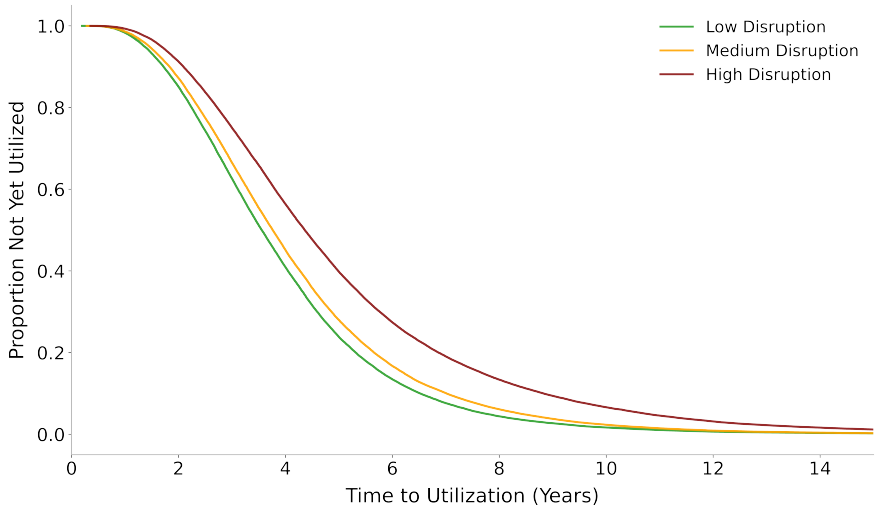


Figure 4: Disruptive Patents Lead to New Product Lines

This figure presents a binscatter plot of the new product line rate against the disruption score, using the sample of over 7 million utilized patent–product pairs. $\text{NewLine}_p = 1$ if product p is the first product in its product line for the firm; product lines are classified using a large language model. The figure divides the disruption score into 20 equal-sized bins and plots the average new product line rate within each bin. The relationship is positive and monotonic ($\beta = 0.346$, $R^2 = 0.77$): the new product line rate rises from roughly 39% for the least disruptive patents to over 55% for the most disruptive. This confirms that patents our algorithm identifies as more dissimilar from existing products do in fact lead to genuinely new product categories rather than extensions of existing lines.

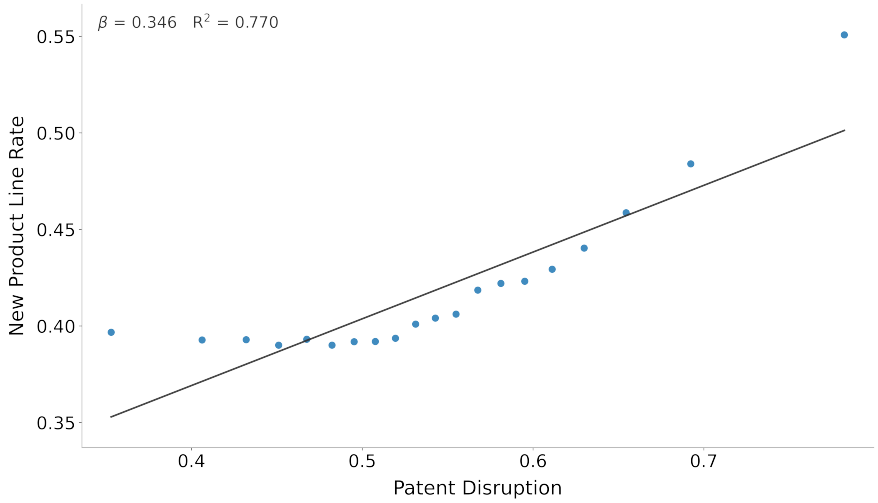


Figure 5: Product Cannibalization: Utilization Penalty by Prior Sales Share

This figure presents binscatter plots of patent utilization rate against the disruption score, separately for firms with above-median (cardinal) and below-median (blue) prior sales share. $\text{PriorSalesShare}_{g,t}$ is the share of firm g's total revenue (from Compustat Segments) attributable to business segments linked to products announced before the patent filing date. Both groups exhibit a negative relationship between disruption and utilization, but the slope is steeper for high prior revenue firms ($\beta = -1.45$) than for low prior revenue firms ($\beta = -1.13$). This gap visualizes the interaction effect in Table 5: firms whose revenue depends heavily on existing product lines are disproportionately less likely to commercialize their disruptive patents, consistent with the cannibalization mechanism.

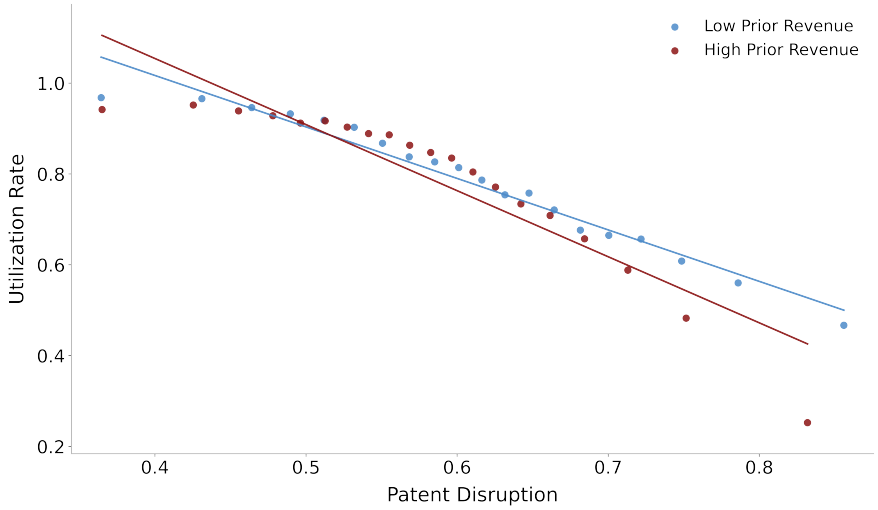


Figure 6: R&D Portfolio Distortion: Prior Sales Share and Mean Disruption

This figure presents a binscatter plot of the mean disruption score of patents filed in a firm-year against the firm's prior sales share, after residualizing both variables on firm and filing year fixed effects. We estimate $\overline{\text{Disruption}}_{g,t} = \alpha_g + \alpha_t + \beta \cdot \text{PriorSalesShare}_{g,t} + \varepsilon_{g,t}$ at the firm-year level, requiring at least two patents per firm-year. The figure divides residualized prior sales share into 20 equal-sized bins and plots the average residualized mean disruption within each bin. The negative slope indicates that, within the same firm, years with higher prior sales shares are associated with less disruptive patent portfolios. This is consistent with the hypothesis that cannibalization concerns distort not only commercialization decisions but also upstream R&D: firms with more at stake in their existing product lines steer their patenting activity toward less disruptive technologies.

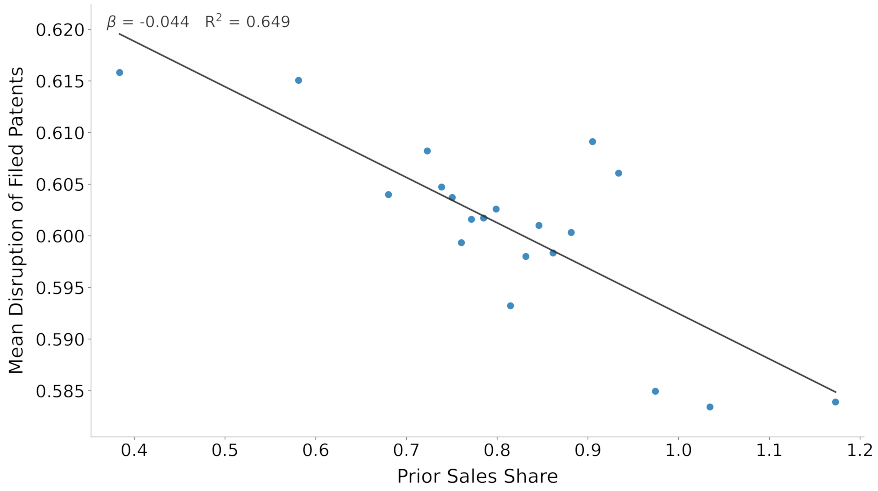


Table 1: Utilized vs. Unutilized patent similarity by CPC section

This table reports mean cosine similarity scores between product announcement text and candidate patent text, separately for patents classified as utilized and unutilized, by CPC section. The final column reports the difference in mean similarity between utilized and unutilized patents, along with t-statistics from two-sided tests of equality of means in parentheses. Higher similarity indicates greater semantic overlap between product descriptions and patent text.

CPC Section	Utilized Patents Mean	Unutilized Patents Mean	Utilized − Unutilized (t-stats)
A (Human Necessities)	0.370	0.223	0.147*** (41.02)
B (Performing Operations; Transporting)	0.352	0.127	0.225*** (33.99)
C (Chemistry; Metallurgy)	0.322	0.125	0.197*** (18.39)
D (Textiles; Paper)	0.312	0.129	0.183*** (8.90)
E (Fixed Constructions)	0.365	0.127	0.238*** (6.74)
F (Mechanical Engineering, etc.)	0.270	0.119	0.151*** (14.09)
G (Physics)	0.328	0.174	0.153*** (56.49)
H (Electricity)	0.290	0.151	0.139*** (74.50)

Notes: *** indicates significance at the 1% level.

Table 2: Disruptive Patents Are Less Likely to Be Utilized

This table presents the OLS estimates of regressing an indicator for patent utilization on the patent-level disruption score. The unit of observation is a patent. We estimate:

$$\text{Utilized}_i = \alpha_g + \alpha_t + \beta \cdot \text{Disruption}_i + \varepsilon_i,$$

where $\text{Utilized}_i = 1$ if patent i is utilized in any product announced by the firm after the patent filing date, and $\text{Disruption}_i = 1 - \max_{j \in \mathcal{P}_g^{prior}} \text{CosSim}(i, j)$, which measures the dissimilarity of the patent to the firm's existing product portfolio. A patent is classified as utilized if the cosine similarity between its embedding and any subsequent product embedding exceeds a CPC-section-specific threshold, calibrated using the IPRoduct ground-truth validation sample. Column 1 reports OLS without fixed effects. Column 2 adds firm fixed effects, which absorb time-invariant firm characteristics such as industry, size, and R&D intensity. Column 3 adds filing year fixed effects. Column 4 includes both firm and filing year fixed effects. Column 5 replaces the separate year fixed effects with CPC section × filing year fixed effects, which control for technology-class-specific trends in utilization rates. Standard errors are clustered at the firm level.

	Utilized				
	(1)	(2)	(3)	(4)	(5)
Disruption	-1.038***	-1.098***	-1.366***	-1.450***	-1.504***
	(0.115)	(0.096)	(0.106)	(0.081)	(0.072)
Firm FE	No	Yes	No	Yes	Yes
Year FE	No	No	Yes	Yes	No
CPC× Year FE	No	No	No	No	Yes
Observations	237,939	237,934	237,939	237,934	237,675
Adj. R^2	0.105	0.182	0.212	0.293	0.356

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Disruptive Patents Take Longer to Utilize

This table presents the OLS estimates of regressing the log time to first utilization on the disruption score. The sample is restricted to utilized patents (i.e., patents with $\text{Utilized}_i = 1$). We estimate:

$$\ln(\text{TimeToUtil}_i) = \alpha_g + \alpha_t + \beta \cdot \text{Disruption}_i + \varepsilon_i,$$

where TimeToUtil_i is the number of years between the patent filing date and the announcement date of the first product utilizing that patent. We use the log transformation to account for the right-skewed distribution of time to utilization. The coefficient β can be interpreted as the approximate percentage increase in time to utilization associated with a unit increase in disruption. Columns 1–5 progressively add fixed effects as in Table 2. The median utilized patent in our sample takes 5.3 years from issuance (8.8 years from filing) to appear in a product. Standard errors are clustered at the firm level.

	ln(Time to Utilization)				
	(1)	(2)	(3)	(4)	(5)
Disruption	0.976***	1.053***	0.569***	0.721***	0.854***
	(0.138)	(0.108)	(0.156)	(0.089)	(0.070)
Firm FE	No	Yes	No	Yes	Yes
Year FE	No	No	Yes	Yes	No
CPC× Year FE	No	No	No	No	Yes
Observations	196,660	196,654	196,660	196,654	196,400
Adj. R^2	0.039	0.169	0.137	0.239	0.304

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Disruptive Patents Lead to New Product Lines

This table presents the OLS estimates of regressing a new product line indicator on the disruption score of the associated patent. The unit of observation is a utilized patent–product pair. We estimate:

$$\text{NewLine}_p = \alpha_g + \alpha_t^{announce} + \alpha_s^{file} + \beta \cdot \text{Disruption}_i + \varepsilon_{ip},$$

where $\text{NewLine}_p = 1$ if product p is the first product in its product line for the firm. Product lines are classified using a large language model (Gemini 2.5 Flash) based on product name, type, and announcement context. When multiple products in the same line launch on the same date, all are coded as new product lines. In our sample, 48% of utilized patent–product pairs are associated with a new product line. Column 1 is OLS without fixed effects. Column 2 adds product announcement year fixed effects. Column 3 adds firm fixed effects, which absorb cross-firm differences in product diversification strategies. Column 4 adds patent filing year fixed effects. Standard errors are clustered at the firm level.

	New Product Line			
	(1)	(2)	(3)	(4)
Disruption	0.345***	0.354***	0.068**	0.088***
	(0.108)	(0.122)	(0.027)	(0.028)
Firm FE	No	No	Yes	Yes
Year FE	No	Yes	Yes	Yes
Filing Year FE	No	No	No	Yes
Observations	7,490,171	7,490,171	7,490,170	7,490,170
Adj. R^2	0.00	0.02	0.09	0.09

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Product Cannibalization Deters Utilization of Disruptive Patents

This table presents the OLS estimates of regressing patent utilization on disruption interacted with the firm's prior sales share. We estimate:

$$\begin{aligned}\text{Utilized}_i = \alpha_g + \alpha_t + \beta_1 \cdot \text{Disruption}_i + \beta_2 \cdot \text{PriorSalesShare}_{g,t} \\ + \beta_3 \cdot \text{Disruption}_i \times \text{PriorSalesShare}_{g,t} + \varepsilon_i,\end{aligned}$$

where $\text{PriorSalesShare}_{g,t}$ is the share of firm g's total revenue (from Compustat Segments) attributable to business segments linked to products announced before the patent filing date t. The linking of products to business segments is performed using a large language model. The coefficient β_3 tests whether the utilization penalty for disruption is amplified in firms with higher dependence on existing products (the cannibalization hypothesis predicts $\beta_3 < 0$). Columns 1 and 3 use the continuous prior sales share; columns 2, 4, and 5 use a binary indicator $\text{HighPriorSales}_{g,t} = 1$ if the firm's prior sales share exceeds the sample median. Columns 1–2 include firm and filing year fixed effects. Columns 3–4 replace filing year fixed effects with CPC section × filing year fixed effects. Column 5 additionally controls for firm age (a binary indicator for firms younger than 5 years) and its interactions with disruption and high prior sales. Standard errors are clustered at the firm level.

	Utilized				
	(1)	(2)	(3)	(4)	(5)
Disruption	-0.567***	-1.336***	-0.696***	-1.414***	-1.340***
	(0.215)	(0.146)	(0.216)	(0.123)	(0.153)
Prior Sales Share	0.795***		0.759***		
	(0.164)		(0.151)		
Disruption × Prior Sales Share	-1.231***		-1.148***		
	(0.265)		(0.264)		
High Prior Sales		0.253**		0.238**	0.250**
		(0.110)		(0.091)	(0.112)
Disruption × High Prior Sales		-0.482**		-0.447***	-0.478**
		(0.188)		(0.162)	(0.194)
Young					-0.124
					(0.138)
Disruption × Young					0.120
					(0.244)
High Prior Sales × Young					0.065
					(0.171)
Disruption × High Prior Sales × Young					-0.097
					(0.318)
Firm FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	No	No	Yes
CPC× Year FE	No	No	Yes	Yes	No
Observations	142,098	142,098	141,823	141,823	142,098
Adj. R^2	0.332	0.329	0.407	0.405	0.330

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: R&D Portfolio Distortion: Prior Product Value Reduces Disruptiveness of New Patents

This table presents the OLS estimates of regressing the mean disruption score of patents filed by a firm in a given year on the firm's prior sales share. The unit of observation is a firm-year. We estimate:

$$\overline{\text{Disruption}}_{g,t} = \alpha_g + \alpha_t + \beta \cdot \text{PriorSalesShare}_{g,t} + \gamma' X_{g,t} + \varepsilon_{g,t},$$

where $\overline{\text{Disruption}}_{g,t}$ is the average disruption score of all patents filed by firm g in year t, and $\text{PriorSalesShare}_{g,t}$ is the share of firm revenue from business segments linked to existing products. The sample includes firm-years with at least two patents filed. Column 1 is OLS without fixed effects. Column 2 adds filing year fixed effects. Column 3 adds both firm and year fixed effects, exploiting within-firm variation. Column 4 adds log total sales as a control for overall firm size. Column 5 uses a binary indicator for above-median prior sales share. Column 6 uses log segment sales as an alternative measure. Standard errors are clustered at the firm level. The sample consists of 1,350 firm-year observations from 150 unique firms.

	$\overline{\text{Disruption}}_{g,t}$					
	(1)	(2)	(3)	(4)	(5)	(6)
Prior Sales Share	-0.077***	-0.063***	-0.026**	-0.025**		
	(0.010)	(0.016)	(0.011)	(0.011)		
High Prior Sales					-0.009	
					(0.006)	
ln(Seg. Sales)						-0.001
						(0.003)
ln(Sales)				0.005		
				(0.007)		
Firm FE	No	No	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes	Yes
Observations	1,350	1,350	1,340	1,335	1,340	1,335
Adj. R^2	0.054	0.146	0.642	0.644	0.641	0.641

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Ruling Out Market Uncertainty: KPSS Patent Value and Utilization

This table tests whether the financial market undervalues disruptive patents at issuance and whether market valuation mediates the relationship between disruption and utilization. The KPSS patent value measure (Kogan et al., 2017) captures the stock market's response to patent issuance by estimating the change in firm market capitalization in a three-day window around the patent grant date. Columns 1–2 regress log KPSS value on disruption. A negative and significant coefficient would indicate that the market undervalues disruptive patents. Instead, we find small and insignificant coefficients, indicating that the market does not systematically penalize disruptive patents at issuance. Columns 3–5 regress log time to utilization on disruption, with and without controlling for log KPSS value; the disruption coefficient is unattenuated by the inclusion of KPSS as a control. Columns 6–7 repeat the exercise with the utilization indicator as the dependent variable; again, controlling for KPSS does not attenuate the disruption effect. All specifications include firm fixed effects. Columns 1, 3–4, and 6–7 include filing year fixed effects; columns 2 and 5 include CPC section × filing year fixed effects. Standard errors are clustered at the firm level.

	ln(KPSS)		ln(Time to Util.)			Utilized	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Disruption	0.070	0.055	0.732***	0.731***	0.864***	-1.460***	-1.459***
	(0.168)	(0.095)	(0.088)	(0.088)	(0.069)	(0.083)	(0.082)
ln(KPSS)				0.008	0.010		-0.021**
				(0.014)	(0.021)		(0.008)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	No	Yes	Yes	No	Yes	Yes
CPC× Year FE	No	Yes	No	No	Yes	No	No
Observations	233,985	233,730	190,448	190,448	190,194	230,872	230,872
Adj. R^2	0.819	0.832	0.237	0.237	0.303	0.294	0.295

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Ruling Out Complementary Technology

This table tests whether the availability of complementary technology explains the longer time to utilization for disruptive patents. The dependent variable is log time to utilization in all columns. Following Kelly et al. (2021), we identify complementary patents for each focal patent and construct two measures: ComplReady_i is the fraction of complementary patents that were filed before the focal patent (capturing the extent to which complementary technology is already available at the time of filing), and $\ln(N_{\text{Comp},i})$ is the log number of complementary patents. Column 1 is the baseline with firm and filing year fixed effects. Columns 2–3 add complements controls. Column 4 replaces filing year fixed effects with CPC section × filing year fixed effects. Columns 5–6 additionally control for KPSS patent value. If the disruption penalty were driven by the absence of complementary technologies, controlling for ComplReady should attenuate the disruption coefficient. Instead, the disruption coefficient *increases* after adding complements controls, indicating that the two channels operate independently. Standard errors are clustered at the firm level.

	ln(Time to Utilization)					
	(1)	(2)	(3)	(4)	(5)	(6)
Disruption	0.721***	0.791***	0.796***	0.887***	0.809***	0.900***
	(0.089)	(0.095)	(0.082)	(0.069)	(0.082)	(0.068)
Complements Ready		-2.181***	-2.095***	-2.128***	-2.152***	-2.182***
		(0.148)	(0.133)	(0.132)	(0.128)	(0.130)
ln(N Complements)			0.156***	0.144***	0.152**	0.140***
			(0.059)	(0.049)	(0.059)	(0.049)
ln(KPSS)					-0.014	-0.010
					(0.013)	(0.016)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	No	Yes	No
CPC× Year FE	No	No	No	Yes	No	Yes
Observations	196,650	196,650	196,650	196,396	190,444	190,190
Adj. R^2	0.239	0.440	0.451	0.498	0.456	0.502

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A Additional Figures and Tables

Table A.1: Heckman Selection Correction: Time to Utilization

This table addresses potential selection bias in the time-to-utilization regressions (Table 3). Because we observe time to utilization only for patents that are eventually utilized, and Fact 1 shows that more disruptive patents are less likely to be utilized, the OLS estimates may be biased. We estimate a Heckman selection model in which the selection equation models the utilization decision and the outcome equation models log time to utilization. The exclusion restriction is the number of prior products announced by the firm before the patent filing date, which predicts utilization (more products increase the likelihood of a match) but should not directly affect the speed of commercialization conditional on utilization. Odd-numbered columns report OLS without selection correction; even-numbered columns report the selection-corrected estimates. Columns 1–2 restrict the sample to young firms (less than 5 years old). Columns 3–4 restrict to firms at least 5 years old. Columns 5–6 restrict to firms at least 10 years old. Columns 7–8 use the full sample. Column 2 uses the two-step estimator; columns 4, 6, and 8 use maximum likelihood with clustered standard errors. Standard errors are in parentheses.

	Young (<5 yrs)		Old (≥5 yrs)		Old (≥10 yrs)		Full Sample	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Outcome equation: ln*(Time to Utilization)*								
Disruption	0.759***	0.067	0.704***	0.359*	0.732***	0.378*	0.684***	0.353
	(0.121)	(0.122)	(0.199)	(0.213)	(0.213)	(0.223)	(0.194)	(0.216)
Selection equation: Utilized								
Disruption		-11.520***		-6.960***		-6.748***		-6.872***
		(0.433)		(0.655)		(0.632)		(0.658)
Selection Correction	No	Yes	No	Yes	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,828	4,369	156,137	191,166	146,162	179,744	159,965	195,535

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.2: Heckman Selection Correction: Cannibalization and Time to Utilization

This table addresses potential selection bias in the cannibalization regressions (Table 5) by estimating a Heckman selection model for time to utilization with disruption interacted with prior sales share. The selection equation models the utilization decision as a function of disruption, prior sales share, and the exclusion restriction (number of prior products). Odd-numbered columns report OLS without selection correction; even-numbered columns report the selection-corrected estimates. Columns 1–2 use the low-revenue subsample with continuous prior sales share. Columns 3–4 use the high-revenue subsample without the interaction. Columns 5–6 use the full sample with the binary high prior sales indicator and its interaction with disruption. Column 2 uses the two-step estimator; columns 4 and 6 use maximum likelihood with clustered standard errors. Standard errors are in parentheses.

	Low Revenue		High Revenue		Full Sample	
	(1)	(2)	(3)	(4)	(5)	(6)
Outcome equation: ln*(Time to Utilization)*						
Disruption	1.489***	0.510***	0.375**	-0.113	0.969***	0.510**
	(0.345)	(0.097)	(0.174)	(0.228)	(0.212)	(0.202)
Disruption × Prior Sales Share	-1.112***	-0.571***				
	(0.355)	(0.094)				
Prior Sales Share	0.510**	0.336***				
	(0.242)	(0.054)				
Disruption × High Prior Sales					-0.686***	-0.628***
					(0.240)	(0.218)
High Prior Sales					0.315**	0.293***
					(0.121)	(0.113)
Selection equation: Utilized						
Disruption		-5.473***		-8.440***		-7.134***
		(0.068)		(1.022)		(0.800)
Selection Correction	No	Yes	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	42,230	54,041	69,606	88,062	111,836	142,103

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B LLM-based Product Name Extraction

This appendix provides additional details on the large language model (LLM)–based procedure used to extract information on newly introduced products from unstructured product announcement text.

B.1 LLM setup

Model and generation parameters We use the Gemini API (Gemini 2.5 Flash) to perform product name extraction. We set the temperature to zero to minimize randomness in the model's responses. The maximum output length is set sufficiently high to accommodate announcements that reference multiple products. Content-based safety filtering is disabled to prevent inadvertent suppression of valid content in corporate product announcements.

Input text For each product announcement, we provide the model with two fields from the S&P Capital IQ Key Developments database: the *headline* and the *situation*. The headline typically highlights the primary product launch, while the situation field contains a more detailed narrative description and may reference multiple newly introduced products.

Extraction task and outputs The model is instructed to extract only newly introduced products and to exclude references to previously released products, generic technology descriptions, or umbrella terms that collectively refer to multiple products. From the headline, the model extracts the main newly introduced product name. From the situation field, the model extracts the name of each newly introduced product along with a short description of its product type, defined as a concise phrase summarizing the product's function or technical category. If no newly introduced product can be identified in a given field, the model is instructed to return an empty string. The same prompt and output format are applied uniformly across all announcements in the sample.

In the empirical analysis, we primarily rely on products extracted from the situation field, which provides more detailed and comprehensive descriptions of product functionality and therefore facilitates better identification of unique products.

B.2 Prompt text

In the following prompt, the *headline* and *situation* fields are provided to the model as input text, and the extracted product name(s) and product type(s) constitute the model

output.

Prompt text:

> I am about to provide you with a product announcement. From the text, extract only the newly introduced product(s) — do not include existing or previously released products. Please extract:
>
> **Headline Product Name**: The main new product name from the HEADLINE.
>
> **Product Name** & **Product Type**: One or more newly introduced products described in the **SITUATION** section. Ignore high-level summary mentions or umbrella terms that refer collectively to multiple products. Only list the individual products that are clearly described with their own functions or features.
>
> **Product Type** should be a short phrase (3–7 words) summarizing the product's function or technical category (e.g., "mobile payment app", "electric vehicle battery").
>
> If no relevant product name or type can be found, return an empty string "" instead of using placeholders like "N/A" or "none".
>
> **HEADLINE:** "headline"
>
> **SITUATION:** "situation"
>
> Respond in this strict format:
>
> `headline_product`: [product name]
>
> Then, for each newly introduced product in the situation:
>
> `name`: [product name], `type`: [product type]

B.3 Illustrative examples

Figure B.1 below presents illustrative examples that demonstrate how the LLM-based product extraction procedure is applied in practice.

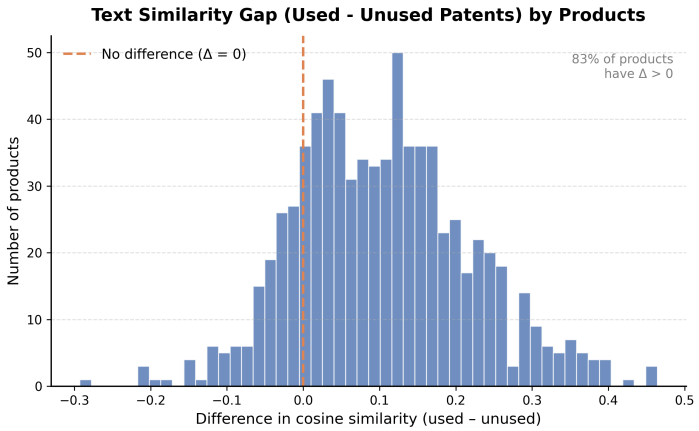
Figure B.1: Illustrative examples of LLM-based product extraction

Input: Headline	Input: Situation	Output: Headline Product	Output: Situation Product(s)	Output: Situation Product Type(s)
Nokia Corp. Introduces Nokia 20Lives	Nokia Corp. introduces a captivating new interactive adventure, Nokia 20Lives, following the highly popular Nokia Game. Nokia 20Lives presents a next generation online and mobile experience, combining elements from games, movies, fashion and music business. Starting on September 19, participants of the new game will have a unique opportunity of stepping into somebody else's shoes - not only once, or twice, but 20 times. Nokia 20Lives presents 20 fictional characters and gives the players a chance of living an exciting day in these characters' lives. Provided by Nokia, the web-based Nokia 20Lives will be played in 21 countries across Europe in 11 languages.	Nokia 20Lives	Nokia 20Lives	online and mobile interactive game
Thermo Fisher Scientific Inc. Launches In-Air SARS-CoV-2 Surveillance Solution	Thermo Fisher Scientific Inc. announced the launch of the Thermo Scientific AerosolSense Sampler, a new surveillance solution designed to deliver fast and highly reliable insight into the presence of in-air pathogens, including SARS-CoV-2. Individual testing will continue to be a cornerstone in the COVID-19 pandemic response because it enables and informs clinical diagnoses, patient triage and treatment decisions. As society returns to pre-pandemic activities, expanded and complementary monitoring of environments will be key in providing useful insights into virus presence. The AerosolSense Sampler is the next step in providing institutions with highly reliable information to help keep their employees and the public safe. The AerosolSense Sampler is an in-air pathogen surveillance solution, which collects representative aerosol samples of ambient air and traps in-air pathogens on a collection substrate. The sample can be readily analyzed through subsequent laboratory testing using polymerase chain reaction (PCR) methodology. Confirming the presence of pathogens like SARS-CoV-2 in a space during a known time interval enables decision makers in hospitals and other indoor facilities to take actions to protect their employees and provide the public with confidence.	In-Air SARS-CoV-2 Surveillance Solution	Thermo Scientific AerosolSense Sampler	in-air pathogen surveillance solution
General Motors Corporation Launches New SUVs	General Motors Corporation announced that, new spacious and fuel-efficient SUVs begin rolling out of a $1.5 billion assembly plant just outside Lansing, Mich. The three new models - the Buick Enclave, which debuted at the Los Angeles Auto Show, and the GMC Acadia and Saturn Outlook - offer a combination of fuel economy, styling, passenger room and standard features. The new vehicles have the profile of a traditional SUV and a car-like unibody chassis that reduces weight and improves interior space and ride comfort.	New SUVs	Buick Enclave	spacious fuel-efficient unibody SUV
			GMC Acadia	spacious fuel-efficient unibody SUV
			Saturn Outlook	spacious fuel-efficient unibody SUV

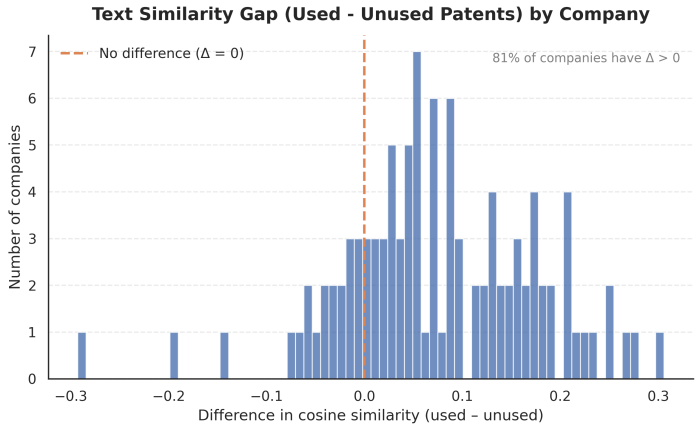
C Additional validation for the patent-product link

Figure C.1: Distribution of text similarity gaps between utilized and unutilized patents

(A) Product-level text similarity gap



(B) Firm-level text similarity gap



Notes: For each product (panel A) and firm (panel B), the text similarity gap is defined as the difference between the mean cosine similarity of patents classified as utilized and the mean cosine similarity of unutilized patents, computed using product announcement text. Positive values indicate higher similarity between products and utilized patents.

D LLM-based Product Line Extraction

This appendix provides additional details on the large language model (LLM)–based procedure used to extract and classify product lines from product announcement text.

D.1 LLM Setup

Model and generation parameters We use the Gemini API (Gemini 2.5 Flash) to perform product line extraction. We set the temperature to zero to minimize randomness in the model's responses. The maximum output length is set sufficiently high to accommodate announcements that reference multiple products. Content-based safety filtering is disabled to prevent inadvertent suppression of valid content in corporate product announcements.

Input text For each product announcement, we provide the model with the *situation* field from the S&P Capital IQ Key Developments database, as well as the target product name extracted following the steps described in Section 2.1.

Extraction task and outputs The model is instructed to identify the overarching product line or family to which the announced product belongs, rather than a specific model, version, or configuration. The prompt emphasizes semantic understanding of product functionality and positioning and explicitly discourages overly granular or model-specific outputs. The model returns a single product line label for each product. If the announcement text does not clearly specify a product line or family, the model is instructed to return an unknown label.

The final prompt used in the analysis is shown below. We experimented with several prompt variants during development to ensure stable and interpretable outputs and fix the prompt prior to constructing the product line classifications used in the paper.

D.2 Prompt text

> **System role.** You are an expert in identifying product lines, brands, or series (family names) from product announcements.
>
> **Task.** From the announcement text below, extract the *product line / brand / trademark* (i.e., the **series or family name**) associated with the **target product**. Return `UNKNOWN` if the announcement does not clearly specify a product line.

Firm information.

- GVKEY: {row['gvkey']}

Target product (name extracted from the announcement).

{row['situation_name_clean']}

Announcement text.

```{row['situation']}```

Definition (important).

- "Product line / brand" refers to a reusable *series or family name* shared across multiple versions or models.
- Output the **shortest unambiguous series or family name** that appears *verbatim* in the announcement text.

Illustrative examples (series name → model/version).

- Consumer electronics: `iPhone` → `iPhone 15 Pro Max`; `Galaxy` → `Galaxy S24 Ultra`
- Laptops/PCs: `ThinkPad` → `ThinkPad X1 Carbon Gen 12`; `PowerEdge` → `PowerEdge R760`
- Semiconductors: `Ryzen` → `Ryzen 9 9950X`; `GeForce RTX` → `GeForce RTX 5090`
- Consumer packaged goods: `Tide` → `Tide Pods`; `Fusion` → `Fusion5 ProGlide`
- Pharmaceuticals: `Ozempic` → `Ozempic 2.0 mg`; `Keytruda` → `Keytruda combination therapy`
- Medical devices: `da Vinci` → `da Vinci Xi`; `Libre` → `Libre 3`
- Automobiles: `Civic` → `Civic Type R`; `F-150` → `F-150 Lightning`

Important exclusions.

- Do *not* output the full model or version string if a shorter series name exists.
- Do *not* output the firm name, a technology or feature name, or a generic descriptor unless it is clearly the product line.
- If no clear product line or brand is specified, return `UNKNOWN`.
```

Rules.

1. Focus *only* on the target product. Ignore other products mentioned.
2. Use only information explicitly present in the announcement text. Do not infer from outside knowledge.
3. Prefer the shortest unambiguous series or family name appearing verbatim.
4. If multiple candidate series names appear, select the one most clearly presented as the product family rather than a feature or technology term.
5. If identification is uncertain, return `UNKNOWN`.

Output format (strict). Return *one line only* in exactly the following format:

```
product_line:  <VERBATIM series name OR UNKNOWN>
```

D.3 Post-processing and cleaning

The raw product line labels extracted by the LLM are standardized using a uniform cleaning procedure to ensure consistency across firms and over time. The cleaning steps include:

- removing trademark and registration markers (e.g., ™, ®, TM);
- normalizing punctuation and whitespace;
- removing suffix metadata after pipe separators (e.g., " — HD" in "Pro Tools — HD")
- removing trailing parenthetical content, which often contains model codes or technical specifications;
- removing explicit generation or series indicators (e.g., "Generation 3," "Series 7");
- stripping trailing model- or version-specific tokens;
- removing generic product-type descriptors (e.g., "platform," "system," "solutions") when they appear at the ends suffixes.

In cases where the LLM returns an indeterminate or missing product line label, we supplement the classification using the product name reported in the announcement, prioritizing names extracted from headline text when available. The same cleaning rules are applied to these supplemented product line labels.

D.4 Alternative product line definitions

To address potential over-specificity and measurement noise, we construct multiple product line definitions that vary in granularity. In addition to the raw cleaned product line name, we generate more general product line identifiers based on the core tokens of the product line label (e.g., using the first one, two, or three tokens after cleaning). For each definition, we construct a conservative, validated version by retaining only product line labels that are directly extracted by the LLM for a given firm.

Our baseline specification uses the validated product line definition based on the first core token of the cleaned product line name.

D.5 Illustrative examples

Figure D.1 provides illustrative examples of product announcement text and the corresponding product line labels extracted by the LLM.

Figure D.1: Illustrative examples of LLM-based product line extraction

Input: Product Name	Input: Situation	Output: Product Line
LOGITECH R QUICKCAM FOR NOTEBOOKS DELUXE	Logitech International SA unveiled the latest additions in its growing family of portable peripherals, designed to help people get the most out of their notebook PCs. The new Logitech(R) Premium Notebook Headset, the Logitech(R) QuickCam(R) for Notebooks Pro and Logitech(R) QuickCam for Notebooks Deluxe webcams make it easy to participate in Internet phone or video calls. The enhanced webcams now feature Logitech(R) RightLight(TM) 2 Technology, which offers superior image quality, even in the harsh lighting people may encounter as they use their notebook PCs in different places. The new Logitech(R) V450 Laser Cordless Mouse for Notebooks offers Logitech's laser-precision tracking technology and alleviates the frustration of trying to navigate complex digital content when using only a touchpad.	QuickCam
FIBER ONE 90 CALORIE CHEWY BARS	General Mills has announced the introduction of new 90 calorie chewy bars under the brand name, Fiber One. Each bar contains 90 calories, five grams of fiber and 2.5gm of fat. The new Fiber One 90 calorie chewy bars are available in two varieties, chocolate and chocolate peanut butter and come in five-count boxes.	Fiber One
NOKIA 5320	Nokia Corp. unveiled two new music-enabled mobile phones, Nokia 5320 and Nokia 5220, as additions to its XpressMusic range. The Nokia 5320 and 5220 cost MYR 1,085 and MYR 750 respectively.	XpressMusic