Measuring Creative Destruction†

Ali Kakhbod — UC Berkeley
Leonid Kogan — MIT, NBER
Peiyao Li — UC Berkeley
Dimitris Papanikolaou — Northwestern, NBER

Abstract

We investigate how innovations by other firms affect a focal firm's future performance, using text-based measures of innovation displacement—i.e., how relevant one firm's innovations are to another's operations. Our findings show that when recent innovations by other major innovators overlap with the focal firm's technologies, the focal firm's profit growth declines over the subsequent seven years. This displacement phenomenon remains robust across different types of firms and model specifications. Moreover, we show that our measure can contribute to practical applications such as out-of-sample prediction of profit growth.

Keywords: Innovation, Displacement, Patents, Machine learning.
JEL Classification: C1, G10, G11, O3.

†First Draft: October 2024. This draft: February 2025. We are grateful to Maxime Bonelli, Amir Kermani, Martin Lettau, and Dmitry Livdan for helpful comments. We thank Borys Babiak for outstanding research assistance. Ali Kakhbod and Peiyao Li are with the Haas School of Business at UC Berkeley. Leonid Kogan is with the Sloan School of Management at MIT and NBER. Dimitris Papanikolaou is with the Kellogg School of Management at Northwestern University and NBER. E-mails: Kakhbod (akakhbod@berkeley.edu), Kogan (lkogan@mit.edu), Li (ojhfklsjhl@berkeley.edu), Papanikolaou (d-papanikolaou@kellogg.northwestern.edu).

1 Introduction

Innovation accelerates the growth of firms by improving the quality of their products or reducing costs. Yet, through creative destruction, new inventions can devalue incumbents' existing technologies and intangible assets. This process of creative destruction has been a significant focus in both macroeconomics and finance.[1] Motivated by the need to measure firm-level exposures to creative destruction empirically, we proposed the first text-based granular measure of Innovation Displacement Exposure (IDE) at the firm level, which we construct using generative AI.

We use a flexible underlying model: any innovation by any firm may be associated with any other firm's growth. The link strength between each pair of firms depends on two factors. First, take two firms where A is innovating and B is potentially displaced. The size of the resulting displacement depends on the relevance of A's recent innovations to B's daily operations and research. For example, innovation in semiconductors is likely associated with a shock in the growth of a GPU manufacturer. However, this same innovation is less likely to be related to changes in a restaurant chain's performance. Secondly, the strength of the link depends on firm A's current aggregate innovation value. This reflects the premise that innovations by a major innovator, such as Apple, may have a more widespread effect on other firms. In contrast, innovations by a lower-ranking innovator may have a more limited impact on other firms.

The first component of our study measures the similarity between firm A's recent innovations and the technologies that firm B currently employs in its operations and research. We consider each issued patent as an innovation and define a firm's assets in the innovation economy by the technologies it utilizes in its daily activities. Drawing on patent data from USPTO and textual descriptions of each firm's technologies from annual 10-K filings, we summarize and embed this information using large language models. Our primary displacement measure is based on the similarity between firm A's recent innovations and firm B's existing technologies.

To incorporate the second component of IDE – the current aggregate market value of A's innovations – into our measure, we weigh the above similarity measure by the aggregate firm-level innovation value in each year, as computed by Kogan et al. (2017). The final weighted values constitute our measure of IDE.

We study the association between IDE and firms' profit growth to validate that our measure captures firms' exposure to technological displacement. We have three main

[1] Aghion and Howitt (1992); Klette and Kortum (2004); Kogan et al. (2017); Kogan, Papanikolaou and Stoffman (2020).

findings. First, we find that IDE is significantly negatively associated with firms' profit growth, with the strength of the effect increasing over the seven-year horizon. These patterns of heterogeneity in predictable firm profit growth cannot be well explained by common firm characteristics. We find that our displacement measure is associated with several traditional firm characteristics such as profitability and R&D expenditures. Nonetheless, a list of 9 innovation-related firm characteristics and over 500 fixed effects can only explain about 15% of the variation in our displacement measure (13% *adjusted* R^2). Moreover, the residuals of the IDE measure (after regressing on firm characteristics) are significantly negatively associated with future profit growth.

Secondly, we demonstrate that our displacement measure is significantly associated with the profit growth of both innovative and noninnovative firms, with the gap between high- and low-IDE firms widening at longer horizons. In addition, to zoom in on the displacement of future growth opportunities among innovative firms, we develop an innovation-based variant of the main IDE measure using patent data as a source of information about the growth opportunities of the displaced firm. Although our two measures are highly correlated across industries and over time, they have low intra-industry correlation. When applied to a subset of innovative firms, our two measures reveal distinct patterns of displacement, and both are significantly negatively associated with the displaced firm's future profit growth.

Thirdly, we demonstrate empirically that our IDE measure is not simply a proxy for the intra-industry product-market competition. In particular, our measure is not absorbed by a horizontal competition measure in the same profit growth regressions. We also show that our IDE measure is negatively associated with a number of other firm growth indicators, including growth of capital stock, employment, output, market share, and intangible capital. While the decline in output, employment, and market share is a typical outcome of increased horizontal product competition, our IDE measure also captures information related to the devaluation of intangible assets and capital. To better explain these observations, we discuss the concept of IDE as obsolescence risk using a "quality ladder" model and provide direct evidence based on the content of firms' annual reports that the perceived obsolescence risk increases when a firm faces higher levels of IDE.

We conduct a set of additional robustness checks to help validate our main technology-based IDE measure. First, we split the dataset by the focal firm's intangible capital, size, and profitability. Across all subsamples, we find a significantly negative association between our main displacement measure and the focal firm's profit growth. Moreover, we use a non-linear model coupled with the Debiased Machine Learning (DML) technique to rerun our analysis over the entire sample. This approach allows us to control for high-dimensional confounders. Using this method, we also confirm that the IDE measure is significantly negatively associated with profit growth, and this association is stronger at longer horizons.

We show that the IDE measure is useful as a predictor of firms' profit growth out of the sample. Specifically, we train and test 20 machine learning models to conduct out-of-sample forecasts of firms' profit growth using IDE and traditional firm characteristics. We used data up to 2012 for training and validation and data after 2012 as the out-of-sample testing set to evaluate our models. The best machine learning model can achieve a 12.7% out-of-sample R^2. Our IDE measure contributes significantly to this predictive power.

Finally, we offer a broadly-applicable method to expand our dataset using Large Language Models (LLMs). The generative ability of LLMs allows us to construct technology descriptions and the corresponding IDE values for firms and years where aggregated data sources such as 10-Ks do not exist. We show that the LLM-based IDE measure has similar properties to our original IDE measure: it is significantly negatively associated with firms' profit growth, particularly at longer horizons.

Our approach is rooted in the tradition of endogenous growth and creative destruction (Acemoglu et al., 2018), emphasizing the role of technological displacement in shaping firm dynamics and growth. Our paper aligns closely with macroeconomic research on technological innovation displacement and contributes to several literature strands at the intersection of innovation, firm performance, stock market outcomes, and machine learning. Traditional methods for identifying technology shocks include measuring technological change through Solow residuals and imposing long-run restrictions on vector auto-regressions (VARs). These approaches, however, are indirect and sensitive to specific assumptions. Our approach diverges by constructing direct measures of technological innovation and displacement using unstructured text data.

Previous studies, such as Shea (1998), measured technological innovation through patents and R&D spending but found weak links between these measures and total factor productivity (TFP). This weakness likely arises from assuming all patents have equal value, a notion disproved by Kortum and Lerner (1998). Additionally, fluctuations in patent counts often reflect regulatory changes rather than genuine innovation. R&D spending is still indirect and subject to efficiency variations over time (Kortum, 1993). Alexopoulos (2011) introduced an aggregate-level measure based on technology-related books, but it lacks firm-level precision. In contrast, our work introduces a text-based measure of innovation displacement, assessing how innovations by other firms impact a focal firm's future performance. This firm-level approach allows for granular analysis of reallocation and growth dynamics, revealing new insights into how external innovations have displacement effects on other firms' growth.

Our work also adds to the literature on the broader effects of technological innovation, particularly in balancing knowledge spillovers and business stealing. We show that similar innovations by major innovators can lead to significant displacements of the focal firm's current technologies, especially for firms less engaged in innovation (Kogan, Papanikolaou and Stoffman, 2020).

Finally, our paper makes a methodological contribution to the economics literature. A rapidly growing branch of the big-data literature in economics and finance uses natural language processing to quantify text (see e.g. Gentzkow, Kelly and Taddy (2019)).[2] Our paper introduces a new method to estimate the similarities between financial texts based on recent advances in computer science, allowing the use of text embeddings to represent the contextual meaning of texts.

The remainder of the paper is organized as follows. Section 2 presents our main technology-based measure. In Section 3, we study the association between innovation-induced displacement and firms' profit growth. Section 4 considers innovative versus non-innovative firms and introduces an alternative innovation-based measure of displacement exposure that solely relies on patents. Section 5 discusses an obsolescence-risk-driven mechanism of IDE that is different from product competition. Section 6 considers the robustness of our results by looking at different subsets of the data and different model specifications. Section 7 presents evidence that our IDE measure can be used to make effective forecasts of firms' profit growth in an out-of-sample setting. Section 8 introduces a method to expand the IDE dataset using LLMs. Section 9 concludes.

2 Data processing

To measure a firm's innovation displacement exposure (IDE), we construct a dataset of firm technologies and innovations spanning each year from 2005 to 2015.[3] We define a firm's innovations in year t by the patents it received from $t-4$ to t, and we define its current technologies as those employed by the firm over the past five years in its operations and research. Specifically, we obtain yearly technology summaries using 10-K files and then aggregate the summaries from $t-4$ to t to build a comprehensive technology

[2] A partial list of papers in this vein includes the work of Hansen, McMahon and Prat (2018), Chen, Wu and Yang (2019), Gentzkow, Kelly and Taddy (2019), Athey and Imbens (2019), Kelly et al. (2021), Erel et al. (2021), Fedyk et al. (2024).

[3] Our technology data are extracted from 10-Ks which begin in the year 2001; however, our displacement measure in each year is constructed with a look-back window of 5 years. Therefore, the earliest year we can compute displacement measures for is 2005.

stack for the firm.

First, we provide an overview of the algorithm we use to compute firm-level similarities, which is the key component in constructing IDE. Suppose we are computing similarities at the end of year 5, the inputs are the 10-K texts of the focal firm A and the patent texts of the innovative firm B from years 1 to 5. We first use GPT to summarize the text documents to get a refined more concise description of the operating technology stack of firm A and the recent innovation stack of firm B. Then, we use a text embedding model to compute numerical representations of the technology stack and innovation stack. Lastly, we compute the cosine similarity between the technology embedding of firm A and the innovation embedding of firm B. We provide a more detailed description in the following subsections.

2.1 Constructing representations of firm innovations

First, we generate a numerical representation of each firm's innovations in each year with a 5-year lookback window. We start by producing an innovations summary for each firm each year based on the patents issued to it in year t. We use GPT4o-mini to summarize firm i's patents in year t in at most 2000 tokens following the prompt:

> *You are an economist studying firms' innovation. Below is a collection of patent abstracts issued to a single firm within the same year. Your task is to read the abstracts very carefully and use only the provided information to summarize them in a structured and detailed manner as follows:*
>
> 1. ***General Themes and Common Topics:***
>   - *Identify and summarize the overarching themes or problems addressed by multiple patents.*
>   - *Highlight any shared goals or innovations that appear across the abstracts.*
> 2. ***Individual Patent Details:***
>   - *For each patent that does not fit into any of the common topics, provide a concise summary that captures its unique contributions, specific technological solutions, or methods described.*
>   - *Mention the targeted problem, proposed solution, and any notable technical features.*
> 3. ***Do not speculate:***

- *If something is not explicitly stated, omit it. We only want the information documented in these patents.*

Here are the patent abstracts:

{patent texts}

Next, we generate innovation summaries for firm i in each year $t \in \{t-4, \cdots, t-1\}$ and concatenate the summaries in these 5 years chronologically from the earliest to the latest. Finally, we extract a numerical embedding based on each concatenated innovation summary using a text embedding model.[4]

2.2 Constructing representations of firm technologies

The second key component of our methodology focuses on generating firm-level numerical representations of operating technologies. We rely on each company's 10-K filing as the primary source of information because it provides a comprehensive overview of operational processes, strategies, risks, and investments. However, due to the substantial length and complexity of these filings, we use GPT4o-mini to produce structured summaries that highlight only the most relevant technological details. In particular, we developed a specialized prompt to guide GPT4o-mini toward extracting concise yet detailed insights from the 10-Ks. Importantly, in both the summarization prompts for 10-K and patents, we ask GPT to not speculate and only extract information from the provided texts. Appendix B shows the exact prompt we use for summarizing 10-Ks.

Similar to constructing innovation embeddings, we obtain an at-most-2000-token summary for firm i for each year from $t-4$ through t, concatenate these summaries, and then utilize the same text-embedding model to create the technology embedding for firm i in year t.

2.3 Merging different sources to construct the main dataset

Lastly, we combine data from different sources to form our main dataset. We acquire the patent data from USPTO, the aggregate innovation value data based on Kogan et al. (2017), and firm characteristics from COMPUSTAT's global firm data set, including company profits, R&D expenses, capital expenses, employment, total asset, capital stock, market capitalization, 3-digit industry code, and the corresponding timestamps. These datasets are merged based on a unique firm identifier PERMNO and the year variable in

[4] We use text-embedding-3-large to generate 3072-dimensional embeddings.

each data set. Then, we merge this combined firm data with innovation and technology embeddings by the same identifiers.

Insert table A1 here.

Table A1 shows the summary statistics of our data from 2005 to 2015. To filter our sample for analysis, we drop data points with missing values in the independent variables we use: IDE, current profit, profitability, $\frac{\text{total sales}-\text{cost of goods sold}}{\text{total asset}}$, capital stock, employment, asset growth, $\frac{\text{total asset in the current year}}{\text{total asset in the previous year}}$, PERMNO, year, industry code, and industry category. Our filtered data set contains 28,075 data points (firm-year) at varying sizes and profit levels.[5] To interpret the size of our main results in later sections, it is worth noting that the average one-year profit growth in the filtered set (original) is 2.5% (1.9%) and the average seven-year profit growth is 19.8% (18.0%).

Moreover, we have 4,454 unique firms in our data set, and as shown in figure 1, the industries with the largest number of firms are Personal and Business Services and Business Equipment, which have about 600-800 firms each. These include personal and business software providers such as Apple and Microsoft. The smallest industry categories are Textile and Tobacco Products, which have fewer than 50 firms.

2.4 Computing the main displacement measure

We define Innovation Displacement Exposure (IDE) as the cumulative impact of innovations from other firms on the focal firm, capturing the aggregated displacement exposure due to external innovations. More specifically, we compute the value-weighted similarity between other firms' innovations compared to i's technologies used in daily operations and research.

$$IDE_{i,t}^{\text{Tech}} = \frac{\sum_j^{n_{t,\text{Innov}}} \cos(\mathbf{Innov}_{j,t}, \mathbf{Tech}_{i,t}) \cdot A_{j,t}^f}{\sum_j^{n_{t,\text{innov}}} \text{Mk Cap}_{j,t}}, \tag{1}$$

where $n_{t,\text{Innov}}$ is the number of innovative firms in year t, $A_{j,t}^f$ is the aggregate innovation value of firm j in year t computed by Kogan et al. (2017), $\mathbf{Innov}_{j,t}$ is the innovation embedding of innovative firm j in year t, $\mathbf{Tech}_{i,t}$ is the technology embedding of firm i in year t, and $\text{Mk Cap}_{j,t}$ is the market capitalization of innovative firm j in year t.

By construction, $IDE_{i,t}^{\text{Tech}}$ is higher when major innovators introduce advancements

[5] A more detailed discussion of our data construction procedure is outlined in Appendix O. This includes firms with no recorded profit in the next year; therefore, the sample size becomes smaller in our analysis.

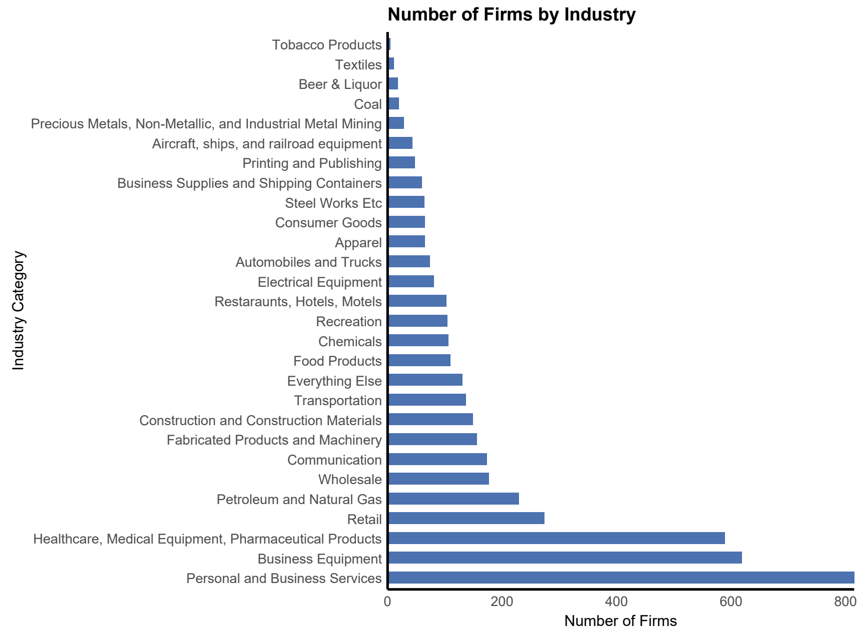


Figure 1: This table shows the number of firms in each industry category in our sample of firms from without any missing values in IDE, current profit, profitability, $\frac{\text{total sales}-\text{cost of goods sold}}{\text{total asset}}$, capital stock, employment, asset growth, $\frac{\text{total asset in the current year}}{\text{total asset in the previous year}}$, PERMNO, year, industry code, and industry category. Our sample is from 2005 to 2015.

that closely resemble the technologies employed by the focal firm. This measure captures the extent of innovation-driven displacement risk, extending beyond direct competitors and upstream suppliers to include broader technological disruptions. For instance, consider the recent decline in Nvidia's market valuation following the release of DeepSeek's reasoning language model, which demonstrated state-of-the-art performance while requiring significantly fewer GPUs for training. Although DeepSeek is not a direct competitor of Nvidia, it poses a competitive threat to some of Nvidia's largest clients. By reducing the demand for high-performance GPUs in large-scale AI training, this technological breakthrough effectively diminished the value of a key segment of Nvidia's technology stack.

3 Displacement and profit growth

We first study the association between innovation-induced displacement and firms' profit growth. In particular, when other firms develop innovations similar to the technologies used by the focal firm, the focal firm's operations are likely displaced. It either loses its competitive advantage or has to adopt these new technologies. In both cases, its future profit will be hindered. Therefore, we conjecture that a firm with a higher IDE value is expected to have lower profit growth. We test this hypothesis using the following 7 regressions ($k \in \{1, 2, 3, 4, 5, 6, 7\}$):

$$\Pi_{f,t+k} - \Pi_{f,t} = \alpha_1^k IDE_{f,t}^{\text{Tech}} + \alpha_2^k \log(AG)_{f,t} + \alpha_3^k P_{f,t} + \beta^k Z_{f,t} + \delta_{f,t}^k + \epsilon_{f,t}^k, \qquad (2)$$

where $\Pi_{f,t+k}$ and $\Pi_{f,t}$ are firm f's log profits in year $t+k$ and t, $IDE_{f,t}^{\text{Tech}}$ is the text-based displacement measure of firm f in year t, $\log(AG)_{f,t} = \log\left(\frac{AT_{f,t}}{AT_{f,t-1}}\right)$ is the log of asset growth of the firm f at year t ($AT_{f,t}$ is the total assets owned by firm f in year t), $P_{f,t}$ is the profitability of firm f in year t defined by $\frac{\text{total sales}-\text{cost of goods sold}}{\text{total asset}}$, $Z_{f,t}$ are the log of the profit, employment, and capital stock in year t of firm f, and $\delta_{f,t}^k$ are the fixed effect (year, industry code, and industry category interacted with year). All standard errors are double clustered at the firm-year level. All independent variables are standardized at the industry category and year level.

Insert table *A*2 here.

As shown in Table A2, the coefficients on profitability are significantly negative, indicating a mean-reversal trend: firms that are currently highly profitable tend to experience lower future profit growth. Additionally, the positive and significant coefficients on asset growth align with Fama and French (2015)'s finding that asset growth serves as a proxy for future growth opportunities. More importantly, the Innovation Displacement Exposure (IDE), our main displacement measure, is significantly and negatively associated with profit growth across all of the next 7 years.[6] This suggests that firms exposed to higher levels of innovation displacement are expected to experience a sustained decline in profit growth over at least the next 7 years. Furthermore, as shown in Figure 2, the magnitude of this effect increases from 0.9% in the next year to 4.7% in the seventh year. [7] This trend suggests that the association of displacement intensifies over time.

To formally test this observation of the worsening displacement effect over time, we construct a dataset where we concatenate all profit growth in the next 1 through 7 years into one column, concatenate all of the control variables used in equation 2, and include an additional variable corresponds to the number of years across which the profit growth

[6]This result that IDE is significantly negatively associated with the focal firm's profit growth is robust when we standardize IDE across the entire sample after demeaning by each year and industry category (as shown in table A3).

[7]Strictly speaking, the percent change refers to the change in log profit.

is measured. In particular, we run the regression:

$$\begin{aligned}\Pi_{f,t+k} - \Pi_{f,t} =& \alpha_1 IDE^{\text{Tech}}_{f,t} + \alpha_2 \log(AG)_{f,t} + \alpha_3 P_{f,t} \\ &+ \alpha_4 IDE^{\text{Tech}}_{f,t} k + \alpha_5 \log(AG)_{f,t} k + \alpha_6 P_{f,t} k \\ &+ \beta_1 Z_{f,t} + \beta_2 Z_{f,t} k + \beta_3 k + \delta_{f,t,k} + \epsilon_{f,t,k}, \end{aligned} \tag{3}$$

where α_1, α_2, and α_3 estimate the baseline association between displacement, asset growth, and profitability and profit growth. α_4, α_5, and α_6 estimate the change in magnitude of these associations as the forward-looking window increases by 1 year. The fixed effects include year, industry code, the interaction between industry category and year, as well as the interactions between the window size indicators (1, 2, $\cdots$, 7) and each of these fixed effects. We standardize all independent variables and triple-cluster standard errors at the firm-year-window size level.

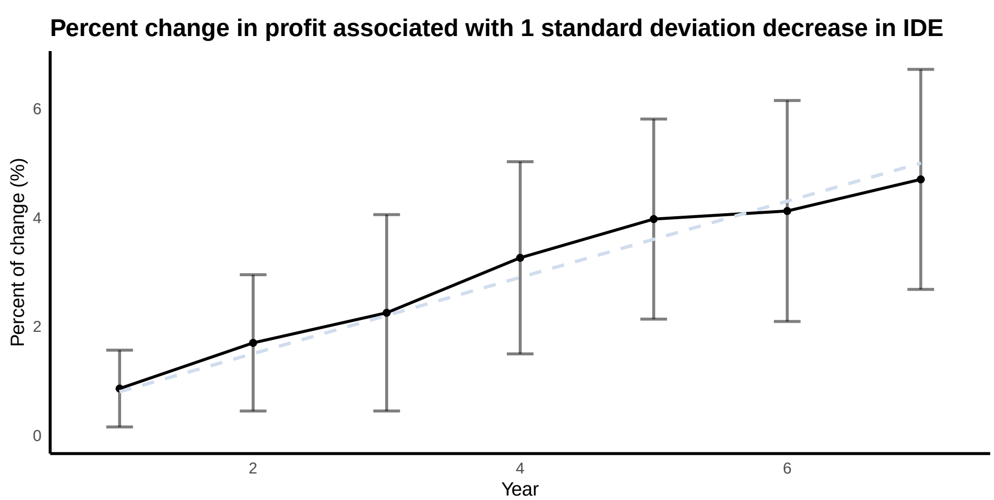


Figure 2: This figure shows the percent change of profit associated with one standard deviation **decrease** in displacement, estimated using equation 2. Note that displacement is negatively associated with profit growth. The blue line represents the widening of the gap estimated using equation 3.

Insert table *A4* here.

As shown in table A4, each additional standard error increase of IDE is associated with a 0.7% additional decrease in profit growth each year. As a sanity check, the association between IDE and profit growth in the first year is the baseline plus the additional association: 0.1% + 0.7% = 0.8% which is close to the results in table A2, and the magnitude in the subsequent years also match in the two tables. This 0.7% additional association is negatively significant. This confirms that the association between IDE and profit growth is increasing in magnitude.

Insert table A5 here.

The natural question to test next is whether the new displacement measure is a combination of traditional innovation-related firm characteristics and whether it captures additional information indicative of firms' profit dynamics. Moreover, once we remove the portion of IDE explained by these variables, are the residuals still significantly negatively associated with profit growth?

First, we analyze which firm characteristics are associated with our displacement measure using the following regression:

$$IDE^{\text{Tech}}_{f,t} = \alpha^{IDE^{\text{Tech}}} X_{f,t} + \delta^{IDE^{\text{Tech}}}_{f,t} + \epsilon^{IDE^{\text{Tech}}}_{f,t}, \tag{4}$$

where $X_{f,t}$ are vectors of firm characteristics including the firm's log capital stock, log employment, log profit, log asset growth, profitability, total asset, market cap, capital expenditures, and R&D expenditures.[8] $\delta^{\text{IDE}}_{f,t}$ are the fixed effects corresponding to each IDE value (year, industry code, and industry category interacted with year).

We find that firms with high IDE (exposed to large innovation displacements) tend to have higher capital stocks, higher profitability, lower asset growth, lower total assets, lower capital expenditure, higher research and development expenditures, and higher market capitalization. Nonetheless, these variables and the fixed effects can only explain about 15% of the variation in the displacement measure. This shows that our measure captures information that is not associated with traditional firm characteristics.

Insert table A6 here.

More importantly, when we reestimate the regression specified by equation 2 using the residual of IDE in equation 4, we still see a significantly negative association between the residuals and profit growth as shown in table A6. Each standard deviation increase in the residual of IDE is associated with a 0.4% decrease in profit by year 1 and 3.2% by year 7. This shows that the unexplained information captured by IDE is associated with negative profit growth. In other words, our main displacement measure (IDE) captures information indicative of firms' profit growth that is not included in any of the traditional innovation-related firm characteristics.

[8] We apply an asinh transformation to R&D and capital expenditures to stabilize the variance in these variables while allowing for negative values.

In the previous analysis, we have shown that IDE is significantly negatively associated with future profit growth even when only examining its residuals after regressing on firm characteristics; therefore, IDE is of clear interest to firm managers and investors prior to the realization of these displacement events. Hence, we study whether there are characteristics that indicate certain firms may have higher IDE values in the future. Since our IDE measure is computed based on innovation and technology descriptions with a 5-year lookback window, we start examining firm characteristics 5 years before the construction year of IDE and also consider a longer horizon from 10 years ago. More specifically, we consider the following variables: total assets, R&D expenditures, profitability, and capital expenditures (standardized in each year and industry category). One issue of this analysis is that characteristics such as capital expenditures and total assets are strongly correlated. Thus, we take the following steps to remove the collinearity of these variables: first, we run a Principal Component Analysis (PCA) on the 4 independent variables; then we remove the common variations in these variables using the first 2 principal components; lastly, we standardize the residuals and run a regression where the dependent variables are the IDE values T years in the future ($T \in \{5, 10\}$) and the independent variables are the residual total assets, R&D expenditures, profitability, and capital expenditures.[9]

Insert table *A7* here.

As shown in Table A7, the two variables that predict variations in the 5-year future IDE are capital expenditures and profitability. Firms with more capital expenditures are more likely to be exposed to higher displacement, while firms that are more profitable tend to be exposed to less. Intuitively, firms with more capital expenditures spend more on maintaining and operating their current technologies; on the other hand, firms that exhibit high profitability relative to their current assets are likely to possess unique assets that are difficult to replicate or have a stronger ability to deter competition and avoid displacement. Firms with high profitability shows a similar trend 10-years into the future. In addition, firms with more R&D expenditures now tend to be more prone to displacement 10-years in the future.

4 Innovative firms and innovation-based IDE

Our primary measure of IDE is based on the similarity between a focal firm's technologies and other firms' innovations, capturing the displacement of the focal firm's existing

[9]The first 2 principal components explain $78 - 79\%$ of variation in the data.

assets (operating technologies). However, firms with a substantial volume of innovations (innovative firms) may also face displacement from a different source—future growth opportunities. To explore this, we first conduct a subsample analysis to compare firms that are active innovators from those that are not. We find that both groups experience displacement of their existing assets. Building on this, we introduce an additional IDE measure focused solely on innovations, designed to capture the displacement of future growth opportunities specific to innovative firms.

4.1 Innovative versus non-innovative firms

Given the main result – IDE is negatively associated with profit growth – we study whether this association varies between two types of firms in the economy: innovative and non-innovative firms. We define a firm's innovation activity by counting the number of patents received by the firm within our patent data set spanning 1976 to 2022. A firm is considered innovative if it receives at least 10 patents during this interval.

Insert tables *A8* and *A9* here.

We begin our analysis by running the regressions specified in equation 2 on the two subsamples separately: one for innovative firms and one for non-innovative firms.[10] As shown in tables A8 and A9, in both subsamples, displacement is significantly negatively associated with firms' profit growth in the next 7 years. This shows our main result is robust for both innovative and non-innovative firms. In addition, as shown in figure 3, the difference in magnitude between the two subsamples is widening over time. In particular, one standard deviation increase in displacement corresponds to 1.1% decrease in profit growth in the first year within both subsamples. However, 7 years in the future, the same increase in displacement is associated with 4.7% decrease in profit growth among non-innovative firms, and the magnitude for innovative firms is at a similar 5.7%.

4.2 Patent-based displacement measure

We introduce an alternative innovation-based measure of displacement exposure that solely relies on patents' texts. This measure captures an additional form of innovation-induced asset displacement that innovative firms experience: the displacement of their recent innovations which we use to proxy for their growth opportunities. Specifically,

[10] We use the data from 2005 to 2015 for this analysis because we do not have data for non-innovative firms prior to 2005.

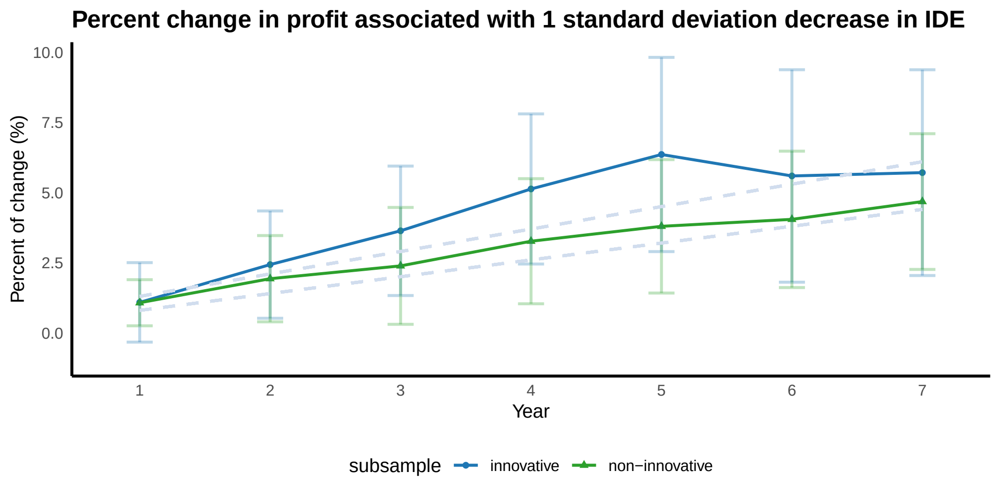


Figure 3: This figure shows the difference between innovative and non-innovative firms in terms of the percent change of profit associated with one standard deviation decrease in displacement, estimated using equation 2. Note that displacement is negatively associated with profit growth.

when a major innovator introduces a new innovation, the recent innovations of the focal firm, along with its current technologies, may lose their competitive edge as they no longer represent the state-of-the-art in their respective fields. Therefore, our technology and innovation-based displacement measures complement each other when studying the profit growth of innovative firms.

The innovation-based IDE measure is defined as

$$\mathrm{IDE}^{\mathrm{Innov}}_{i,t} = \frac{\sum_{j}^{n_{t,\mathrm{innov}}} \cos(\mathbf{Innov}_{j,t}, \mathbf{Innov}_{i,t}) \cdot A^{f}_{j,t}}{\sum_{j}^{n_{t,\mathrm{innov}}} \mathrm{Mk\ Cap}_{j,t}}, \tag{5}$$

where $\mathbf{Innov}_{j,t}$ and $\mathbf{Innov}_{i,t}$ are both innovation embeddings computed based on summaries of firms' recent patents.

First, we study the correlation between the two measures. We show that the innovation-based IDE measure is positively correlated with the main measure, but the correlation is mostly due to heterogeneity across years and industry categories.

Insert table *A*10 here.

More specifically, we compare the two measures using two types of correlations: Pearson correlation computes the correlation of the two continuous measures, and Kendall-Tau computes the correlation between the rank orderings (numbers in parentheses). As shown in table A10, the raw numbers of the main and alternative displacement measures are highly correlated at over 88% (67%); however, the correlation drops to a lower 25% (15%) when these measures are standardized within each year and industry category. This implies that within each industry category and year, the two measures are directionally aligned but capture different information.

Next, we study the association between this patent-based displacement measure and profit growth among firms with at least 10 patents – innovative firms.[11] In particular, we run a modified version of the main regression in equation 2 over the sample of innovative firms:

$$\Pi_{f,t+k} - \Pi_{f,t} = \alpha_1^k IDE_{f,t}^{\text{Tech}} + \alpha_2^k IDE_{f,t}^{\text{Innov}} + \alpha_3^k \log(AG)_{f,t} + \alpha_4^k P_{f,t} + \beta^k Z_{f,t} + \delta_{f,t}^k + \epsilon_{f,t}^k, \tag{6}$$

where $IDE_{f,t}^{\text{Tech}}$ is the standardized main displacement measure and $IDE_{f,t}^{\text{Innov}}$ is the standardized patent-based alternative measure.

Insert table A11 here.

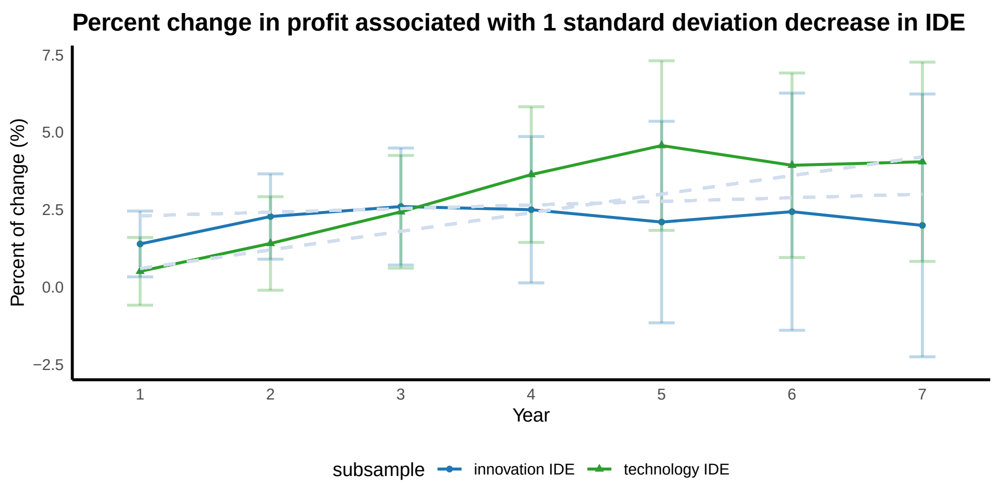


Figure 4: This figure shows the percent change of profit associated with one standard deviation **decrease** in our two displacement measures, respectively, estimated using equation 6. Note that displacement is negatively associated with profit growth.

As shown in Table A11, the alternative innovation-based IDE measure identifies significant and distinct displacement effects among innovative firms. Specifically, while the technology-based IDE has an incremental association with profit growth from year to

[11]Since this alternatively measure defined by equation 5 only depends on patents, we focus on the sample of innovative firms.

year: going from 0.5% in the next year to 4.0% 7 years in the future, the association between the innovation IDE measure and profit growth increases from 1.4% to 2.0%, and the coefficient on the innovation-based IDE measure is significant in the first 4 years. This shows that in the near future, innovative firms will suffer more from displacements of growth opportunities, while their profit growth is statistically more associated with displacements of assets in place in the long run.

5 Mechanism of IDE

In this section, we distinguish IDE from firm-level product competition. We propose a framework that views IDE as a risk of obsolescence. First, we demonstrate that IDE is associated with growth variables commonly linked to both product competition and innovation displacement, as well as those more directly related to obsolescence risk. Next, we present a theoretical motivation for this perspective and provide direct empirical evidence, drawn from 10-K filings, which shows a positive relationship between IDE and the focal firm's exposure to obsolescence risk.

5.1 IDE versus product competition

Our IDE captures the threat focal firms' face from major innovators' recent innovations that are similar to their operating and research technologies. Another type of competition is horizontal competition measured by product-product similarities – firms face greater threats if other firms are offering similar products. In this subsection, we show that our IDE measure captures information distinct from product competition.

First, we update the total similarity measure introduced in Hoberg and Phillips (2016) by utilizing the same language models employed in our IDE construction. This enhancement ensures that both measures are comparable on an equal footing, eliminating discrepancies arising from differences in text representation methodologies. To achieve this, we first instruct GPT4o to summarize each 10-K document and extract the relevant product descriptions.[12] We then apply OpenAI's text embedding model to generate numerical representations for each summary, aligning this approach with our IDE construction. Finally, in accordance with the total similarity framework, we compute each firm's total similarity score by aggregating the cosine similarities of its product embeddings with those of other firms within the same year. Formally, let $\text{emb}_{f,t}$ denote the product embed-

[12] The exact prompt is provided in Appendix B. Following the convention established by Hoberg and Phillips (2016), we direct GPT4o to focus specifically on the business section of 10-K filings.

ding of focal firm f in year t, and let $F_{-f,t}$ represent the set of all firms in year t excluding f. The total similarity for firm f in year t is then computed as

$$TS_{f,t} = \sum_{j \in F_{-f,t}} \cos(\text{emb}_{f,t}, \text{emb}_{j,t}). \tag{7}$$

Then, we run 7 regressions with both IDE and total similarity and include the same controls and fixed effects as our main regressions specified in equation 2.

$$\Pi_{f,t+k} - \Pi_{f,t} = \alpha_1^k IDE_{f,t}^{\text{Tech}} + \alpha_2^k TS_{f,t} + \alpha_3^k \log(AG)_{f,t} + \alpha_4^k P_{f,t} + \beta^k Z_{f,t} + \delta_{f,t}^k + \epsilon_{f,t}^k, \tag{8}$$

where $\Pi_{f,t+k}$ and $\Pi_{f,t}$ are firm f's log profits in year $t+k$ and t, $IDE_{f,t}^{\text{Tech}}$ is the text-based displacement measure of firm f in year t, $TS_{f,t}$ is the total similarity of firm f in year t, $\log(AG)_{f,t} = \log\left(\frac{AT_{f,t}}{AT_{f,t-1}}\right)$ is the log of asset growth of the firm f at year t ($AT_{f,t}$ is the total assets owned by firm f in year t), $P_{f,t}$ is the profitability of firm f in year t defined by $\frac{\text{total sales} - \text{cost of goods sold}}{\text{total asset}}$, $Z_{f,t}$ are the log of the profit, employment, and capital stock in year t of firm f, and $\delta_{f,t}^k$ are the fixed effect (year, industry code, and industry category interacted with year). All standard errors are double clustered at the firm-year level. All independent variables are standardized at the industry category and year level.

Insert table A12 here.

As shown in Table A12, both IDE and horizontal competition, our embedding-based measure of total similarity, exhibit a significant negative association with profit growth. A one standard deviation increase in IDE corresponds to a 0.8% decline in profit growth in the following year and a 4.5% decline by year 7. On the other hand, a one standard deviation increase in total similarity is associated with a 0.6% reduction in profit growth in the next year and a 3.1% reduction by year 7. These results indicate that our IDE measure captures distinct competitive forces beyond horizontal competition. In addition, our updated total similarity measure remains significantly negatively associated with profit growth, supporting its validity as a proxy for product competition.

5.2 Other firm growth outcomes

Beyond profit growth, we further examine the broader relationship between IDE and firm growth, comparing its associations with those of product competition. Specifically, we analyze how IDE relates to growth in capital stock, employment, output, intangible assets—defined by Eisfeldt, Kim and Papanikolaou (2020)—and market share, using a similar specification to Equation 2, where the dependent variable is replaced by these alternative growth measures over the next seven years. As expected for a competition measure, IDE is significantly negatively associated with market share, output, and employment growth, reinforcing its role in capturing competitive pressures. However, IDE differs from total similarity (horizontal product competition) in that it is also significantly negatively associated with growth in intangible assets and capital stock. This distinction suggests that IDE captures competitive forces beyond product market rivalry, potentially reflecting broader industry dynamics that influence firms' ability to invest in and accumulate intangible and physical capital.

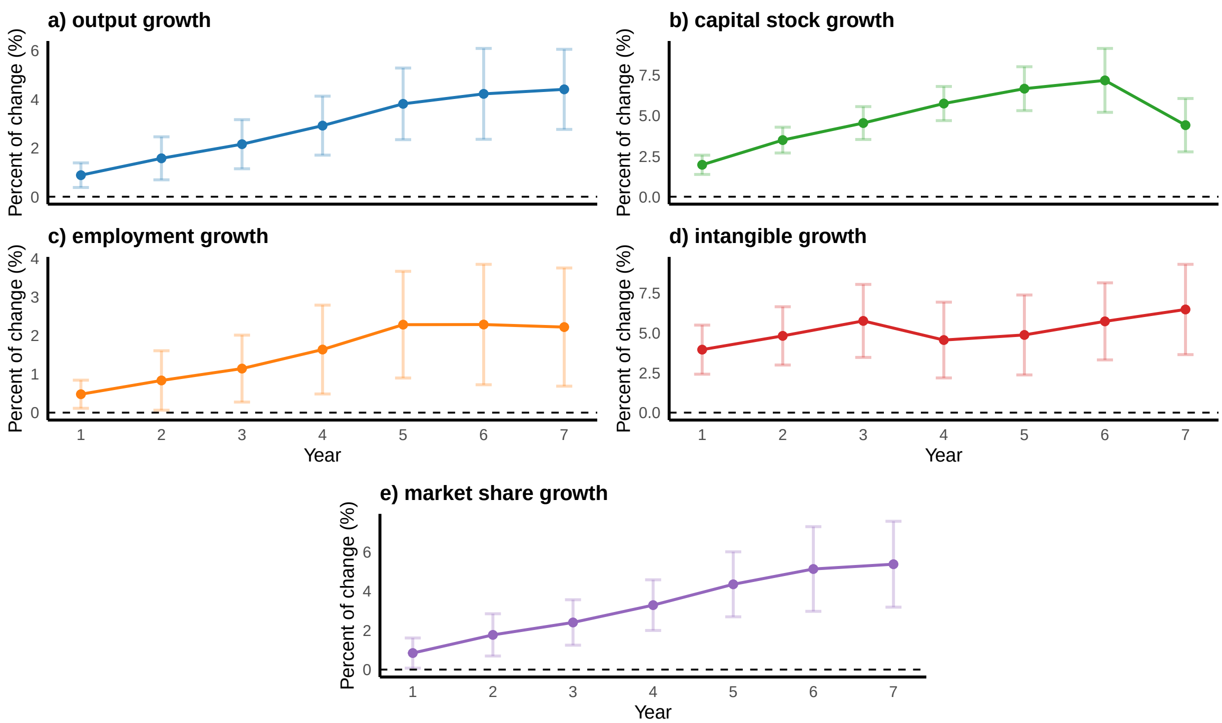


Figure 5: This figure shows the percent change of different firm growth outcomes associated with one standard deviation **decrease** in our technology-based displacement measures. These coefficients are estimated using a modified version of equation 2 where the independent variables are replaced with output growth, capital stock growth, employment growth, intangible asset growth, and market share growth, respectively.

Insert table *A*13 here.

As shown in table A13 and figure 5, IDE is significantly negatively associated with the growth of capital stock, employment, output, intangible asset, and market share in all of the next 7 years. Each standard deviation increase in IDE is associated with a 0.9%

decrease in output growth, 2% of capital stock growth, 0.4% of employment growth, 3.9% of intangible asset growth, and 0.8% of market share growth in the next year. By year 7, these gaps become 4.4%, 4.4%, 2.2%, 6.5%, and 5.4% respectively. This shows that our IDE measure captures displacement that is negatively indicative of the focal firm's growth along multiple dimensions, including profit, output, capital stock, employment, intangible asset, and market share. While the hindrance in market share, output, and employment is also expected with increases in horizontal product competition such as total similarity, the decline in intangible asset growth and capital stock growth is more evidence that our IDE captures additional information related to asset obsolescence risk.[13]

5.3 IDE as obsolescence risk

In subsection 5.2, we demonstrated that our IDE measure is significantly negatively associated with the growth of intangible assets (relative to book equity) and capital stock. This finding suggests that firms with higher IDE scores may be more exposed to obsolescence risk—that is, their operating technologies and current capital stocks are devalued by innovations from other firms. Here, we motivate this mechanism linking IDE and profit growth using a simple "quality-ladder" model inspired by Klette and Kortum (2004). The model captures obsolescence risk through an IDE parameter and illustrates how this risk reduces an focal firm (focal) firm's expected future profits. Specifically, when other firms' patents closely overlap with the focal firm's technology (or know-how), the risk of being technologically leapfrogged increases. Consequently, the focal firm must invest defensively to avoid displacement, and over multiple periods, this pressure diminishes its overall profit growth trajectory. Moreover, we provide direct empirical evidence linking the focal firm's current IDE with both its present and future exposure to obsolescence risk.

5.3.1 Theoretical motivation: a simple "quality ladder" model

Model setup: Consider a continuum of product lines indexed by $\ell \in [0,1]$. For each line ℓ, a single firm (the focal firm) owns the highest observed quality rung $q(\ell) \in \{0,1,2,\dots\}$. A larger $q(\ell)$ represents a higher-quality or more cost-efficient production technology.

[13] A more detailed discussion of the association between total similarity and the growth of firms' market share, output, employment, capital stock, and intangible assets are in Appendix N.

Each product variety ℓ is aggregated into a final output Y via a CES function:

$$Y = \left[\int_0^1 x(\ell)^{\frac{\varepsilon-1}{\varepsilon}} d\ell \right]^{\frac{\varepsilon}{\varepsilon-1}}, \quad \varepsilon > 1.$$

The quantity $x(\ell)$ is demanded at price $p(\ell)$. In standard CES fashion,

$$x(\ell) = \left(\frac{p(\ell)}{P} \right)^{-\varepsilon} Y,$$

where P is the price index. Each focal firm has a marginal cost $c_0/\Gamma^{q(\ell)}$ if it owns quality rung $q(\ell)$, with $\Gamma > 1$ capturing the cost (or efficiency) improvement from climbing the quality ladder. Under monopolistic competition with elasticity ε, one can show that for some constant $\kappa > 0$ (absorbing factors like c_0, μ, Y),

$$\pi(q) = \kappa \Gamma^{(\varepsilon-1)q}.$$

Hence, advancing from q to $q+1$ yields a proportional increase in per-period profit from a single product line.

R&D Competition and the IDE Parameter: In each period, two sources of R&D may act on line ℓ: the *focal firm* and the *outside competitors*. The focal firm invests R_I to improve $q(\ell) \to q(\ell) + 1$ with probability ηR_I for a given $\eta > 0$. Outside competitors invest R_C, and if they succeed, they leapfrog $q(\ell)$ to $q(\ell) + 1$ and *displace* the focal firm. Formally, the probability of outside success is

$$p_C = \lambda \text{IDE} \times R_C, \quad \lambda > 0. \tag{9}$$

Here, IDE quantifies the *overlap* or "displacement exposure" of external patents. A higher IDE implies the competitor's R&D is more effective at displacing the focal firm in that product line. Once a displacement event occurs, the focal firm loses ownership of the line, and hence loses future profits from it.[14]

Per-Line Value Function and Optimization: Let $V_I(q)$ denote the focal firm's *present value* of holding quality q on a single line ℓ. At the start of each period, the focal firm chooses R_I, observes whether it loses to an external innovator (with probability p_C), or successfully upgrades (with probability $p_I = \eta R_I$), or remains at q if neither event occurs. Let $0 < \beta <$

[14]For more details about (9), see Appendix E.1.

1 be the discount factor. The focal firm's value satisfies

$$V_I(q) = \max_{R_I \geq 0} \Big\{ \pi(q) - R_I + \beta \Big[(1-p_C) V_I(q) + p_I (1-p_C) \big(V_I(q+1) - V_I(q) \big) \Big] \Big\},$$

where $\pi(q)$ is the flow of profit from operating quality rung q, R_I is the R&D expenditure by the focal firm, $p_I = \eta R_I$ is the focal firm's chance of self-improvement, and $p_C = \lambda \text{IDE} R_C$ is the chance of competitor displacement. If displacement occurs, the focal firm's payoff for this line becomes zero.

Firm-Level Aggregation: A firm f may control a subset $\mathcal{L}_f \subset [0,1]$ of product lines. Each line $\ell \in \mathcal{L}_f$ is associated with a rung $q(\ell)$ and a value $V_I\big(q(\ell)\big)$. The total value of firm f is

$$\mathcal{V}_f = \int_{\ell \in \mathcal{L}_f} V_I\big(q(\ell)\big) d\ell.$$

When IDE_f is large, each line $\ell \in \mathcal{L}_f$ faces a higher probability p_C, and so $V_I(q(\ell))$ is lower. Summing across all lines therefore reduces the firm's aggregate value.

Future Profit Growth and Proof of Obsolescence Risk: We find that an increase in IDE (i.e., a higher hazard of external displacement) reduces the firm's expected future profit growth.[15]

Proposition 1. ***(Higher IDE Lowers Expected Future Profit Growth)*** *Suppose* IDE *increases while other parameters* $(\beta, \eta, \lambda, \Gamma, \dots)$ *remain fixed. Then, under the equilibrium policy* $\{R_I^*\}$ *that solves the focal firm's value function, the focal firm's expected profit over any finite horizon T strictly decreases. Equivalently, higher* IDE *reduces the growth path of total firm profit in expectation.*

Economic Conclusion: If IDE is high, outside R&D investments more effectively leapfrog the focal firm's product lines. The focal firm thus faces more frequent loss of its quality leadership and cannot credibly rely on multi-year profits from its current technology. In equilibrium, the firm responds by either investing more defensively (raising costs) or risking displacement. Both scenarios lower realized future profits. Consequently, higher obsolescence risk, quantified by IDE, predictably reduces the focal firm's expected future profits and profit growth.

This multi-product quality-ladder model demonstrates how overlapping external innovations, captured by a higher "IDE," increase the probability of displacement along each product line, thereby depressing the focal firm firm's overall future profit trajectory. These findings align with empirical results that link high innovation-overlap measures to

[15]The proof is in Appendix E.

systematically lower subsequent profits, validating the notion of "creative destruction" in a more granular, product-level setting.

5.3.2 Empirical evidence: Measuring obsolescence risk from 10-Ks

Our model and previous empirical findings shows that a higher IDE of the focal firm can lead to a decline in profit growth through an increased level of obsolescence risk. To provide direct empirical foundation for this mechanism, we now examine how firms' self-assessed obsolescence risk—as disclosed in their 10-K filings—is associated with their IDE scores. In doing so, we test whether a higher IDE is significantly positively associated with an increased exposure to obsolescence risk.

First, we instruct GPT4o to generate an annual summary of obsolescence risk discussions from each firm's 10-K filing. Next, we compute a numerical embedding for each summary using the same OpenAI embedding model employed in our primary analysis. We then apply a semantic axis approach, as described in Fedyk et al. (2024), to quantify the severity of the obsolescence risk discussed in each 10-K filing. Specifically, this method constructs a vector in the embedding space by leveraging the embeddings of a pair of contrasting sentences—one representing lower risk and the other higher risk—and projects each 10-K summary embedding onto this axis to estimate its corresponding riskiness score. We extend this method by extracting the first principal component from 10 similarly constructed risk axes rather than relying on a single axis. This enhancement helps mitigate noise arising from slight variations in the language used to describe high versus low risk. For example, one pair of sentences we employ is "The firm is exposed to significant technology obsolescence risk" (high risk) versus "The firm is exposed to minimal technology obsolescence risk" (low risk).[16]

After computing and standardizing the riskiness scores within each year and industry category, we regress the scores on IDE, along with other controls and fixed effects, as specified in Equation 2:

$$\text{risk}_{f,t+k} = \alpha_1^{r,k} IDE_{f,t}^{\text{Tech}} + \alpha_2^{r,k} \log(AG)_{f,t} + \alpha_3^{r,k} P_{f,t} + \beta^{r,k} Z_{f,t} + \delta_{f,t}^{r,k} + \epsilon_{f,t}^{r,k},$$

where $\text{risk}_{f,t+k}$ is firm f's obsolescence risk in year $t+k$ ($k \in \{0, 1, \cdots, 6\}$), $IDE_{f,t}^{\text{Tech}}$ is the text-based displacement measure of firm f in year t, $\log(AG)_{f,t} = \log\left(\frac{AT_{f,t}}{AT_{f,t-1}}\right)$ is the log of asset growth of the firm f at year t ($AT_{f,t}$ is the total assets owned by firm f in year t), $P_{f,t}$ is the profitability of firm f in year t defined by $\frac{\text{total sales} - \text{cost of goods sold}}{\text{total asset}}$, $Z_{f,t}$ are the

[16]The exact prompt used to extract the obsolescence risk discussions and all 10 pairs of sentences employed to measure obsolescence risk are presented in Appendix B.

log of the profit, employment, and capital stock in year t of firm f, and $\delta^{r,k}_{f,t}$ are the fixed effect (year, industry code, and industry category interacted with year). All standard errors are double clustered at the firm-year level. To control for language differences in different firms' reportings, we demean the obsolescence risk projection of each firm. All independent variables are standardized at the industry category and year level.

Insert table A16 here.

As shown in table A16, the focal firm's self-reflected obsolescence risk in the current year and the next 6 years are all significantly negatively associated with the current year's IDE. Each standard deviation increase in IDE is associated with a 0.023 standard deviation increase in obsolescence risk this year and 0.034 by year 6. This shows that our IDE measure captures significant information about the amount of obsolescence risk experienced by each firm.[17]

6 Robustness

In this section, we conduct some further robustness checks for our main result by exploring the negative association between IDE and profit growth across various data subsets, studying the association for data in different years, and testing an alternative model specification.

6.1 Subsample analyses

To explore how displacement affects different types of firms, we partition the data along three dimensions: (i) intangible capital, (ii) firm size, and (iii) profitability. In each case, we split firms at the median within their year–industry group, then estimate the regression in Equation 2 and 3 separately. The objective is to uncover whether displacement has distinct implications depending on a firm's resource profile (e.g., relative intangible assets, market capitalization) or competitiveness (e.g., profitability).

First, we examine intangible capital. Following Eisfeldt, Kim and Papanikolaou (2020), we define a firm's intangible capital as the ratio of intangible assets to book equity at year t, $G_{f,t} = \frac{\text{Int}_{f,t}}{\text{BE}_{f,t}}$. In each year–industry group, firms at or above the median $G_{f,t}$ are clas-

[17]For robustness, we add the obsolescence risk of the focal firm from 5 years ago as an additional control variable. We observe the IDE is still significantly positively associated with the current and next year's self-expressed obsolescence risk, conditioning on previously reported obsolescence risk levels (as shown in table A17).

sified as high intangible capital, and those below are considered low intangible capital. Next, we split the data by firm size, measured by market capitalization within each year–industry group. Firms at or above the median market capitalization are considered large; those below the median are small. We then consider profitability, defined as gross profit over total assets (i.e., $\frac{\text{sale}_{f,t} - \text{cogs}_{f,t}}{\text{at}_{f,t}}$). Firms in the lower half of the profitability distribution within each year–industry group form the low-profitability subsample; the remaining firms form the high-profitability subsample.

Overall, as shown in figure 6, these subsample analyses show that IDE is negatively associated with profit growth in all subsamples, and the associations are especially significant in later years. In addition, the gap between high and low IDE firms widens in all subsamples. This shows that our main result: IDE is negatively associated with profit growth with the gap widening each year is not driven by a particular type of firms but applies generally in a variety of subsamples.

6.2 IDE and profit growth in different years

A potential issue is that our approach relies on embeddings generated by a model trained on a single snapshot in time, even though word and sentence representations may evolve. Since our IDE measure is based on technology and innovation embeddings, its informativeness could vary across different periods. In particular, a document from an earlier year might be misrepresented by a model calibrated to current language usage.

We estimate the association between IDE and profit growth separately for each year and find no clear trend in the magnitude of this relationship. In other words, the informativeness of our IDE measure remains stable over time, neither systematically increasing nor decreasing. To demonstrate this, we modify the specification presented in equation 2 as follows:

$$\Pi_{f,t+k} - \Pi_{f,t} = \sum_{\text{Year} \in \{2005,\cdots,2015\}} \alpha^k_{1,\text{Year}} IDE^{\text{Tech}}_{f,t} * \mathbb{1}(t = \text{Year}) \tag{10}$$

$$+ \alpha^k_2 \log(AG)_{f,t} + \alpha^k_3 P_{f,t} + \beta^k Z_{f,t} + \delta^k_{f,t} + \epsilon^k_{f,t}. \tag{11}$$

Figure 7 illustrates that there is no systematic time trend in the magnitude of the association between IDE and profit growth. Specifically, with the exception of 2010, IDE is negatively associated with the focal firm's profit growth (with estimates above the zero line) beyond year 2. The most pronounced associations occur in 2013, where the negative effect is strongest, and in 2010, where the association is positive until year 6. Notably, the beginning year 2005 and ending year 2015 lie near the median of the distribution. These findings indicate that (i) even when estimated separately for each year, IDE is negatively associated with profit growth in all but one year, and (ii) the informativeness of our IDE measure, as reflected by the magnitude of the associations, does not exhibit a clear time trend.

6.3 Non-linear confounders and debiased machine learning

Another potential concern for our results is that the estimates are based on a linear confounding specification, which may not accurately represent the underlying relationship between displacement and profit growth, confounded by firm characteristics and year. In this subsection, we analyze the entire sample using a non-linear model of the confounding relationship leveraging the debiased machine learning technique developed in Chernozhukov et al. (2018). Debiased machine learning allows us to estimate the parameters of interest (association between IDE and profit growth) while controlling for high-dimensional confounders.

As shown in figure 8, we utilize debiased machine learning in the following steps:

1. Randomly split the dataset into 2 equal-sized subsets D_1 and D_2.
2. Train an XGBoost model $g_T^{D_1}$ using D_1 including all of the control variables and fixed effects in equation 2 to predict IDE (our main displacement measure).
3. Train another XGBoost model $g_O^{D_1}$ using D_1 including all of the control variables and fixed effects in equation 2 to predict the profit growth (we do this 7 times because we consider the profit growth over each of the next 7 years).
4. Generate predictions of IDE and profit growth in D_2.
5. Repeat steps 2-4. Use D_2 as the training data and generate predictions in D_1. The new models are $g_T^{D_2}$ and $g_O^{D_2}$.
6. Combine all of the out-of-sample predictions to form a new dataset.
7. Regress the residuals of profit growth on the residuals of IDE:

$$(\Pi_{f,t+k} - \Pi_{f,t+k}) - \left(\widehat{\Pi_{f,t+k} - \Pi_{f,t}}\right) = \alpha_k + \tau_k^{IDE}(IDE_{f,t} - \widehat{IDE_{f,t}}) + \epsilon_{f,t+k}, \quad (12)$$

where τ_k^{IDE} is the estimated effect of displacement on profit growth.

Insert tables A14 and A15 here.

As shown in table A14, when controlling for all of the confounders (including the continuous variables and fixed effects in equation 2), the negative effect of displacement on profit growth is significant over all of the next 7 years. In particular, as shown by figure 9, the difference is widening each year from 0.8% in the first year to 3.2% in the 7th year. This increase in the effect size per year is a significant 0.5% as shown by table A15.[18] This demonstrates that our main results discussed in section 3 are unlikely a coincidence due to model misspecification.

7 IDE for out-of-sample prediction

In the previous sections, we have shown that IDE is significantly associated with firms' profit growth, and this significance is robust to subsetting by firm types and changing the underlying model specification. Now, we show that IDE has a significant out-of-sample predictive power of firms' profit growth. More specifically, the question we address in this section is whether our IDE measure can significantly improve the performance of predictive machine learning models.

To assess this, we partition our dataset based on a timestamp. Specifically, we train and validate 20 predictive machine learning models using data up to 2012, while testing their performance on data in years after. The objective of these models is to predict each firm's profit growth in the next 5 years. Our modeling approach includes a diverse set of algorithm classes: tree-based methods, boosting, neural networks, and weighted ensembles of these models.

Tree-based methods construct individual decision trees as weak learners and aggregate them to form the final prediction. If we think of each weak learner as a human forecaster, this tree-based approach is similar to recruiting multiple independent forecasters and using the wisdom of the crowd to make the final predictions. Gradient boosting also uses multiple learners, but these learners are built sequentially to gradually improve the performance. This process is analogous to using multiple human forecasters but the next forecaster observes the mistakes made by the previous ones and can try to address them. For a discussion on gradient boosting techniques, see Appendix M.

Neural networks address this problem more directly: each using only one model that predicts profit growth based on each firm's characteristics. These models fit nonlinear

[18] This growth is shown using the following regression analogous to equation 3:

$$\left(\Pi_{f,t+k} - \Pi_{f,t+k}\right) - \left(\widehat{\Pi_{f,t+k} - \Pi_{f,t}}\right) = \alpha + \beta_1 k + \beta_2 (IDE_{f,t} - \widehat{IDE_{f,t}}) + \tau^{IDE} k (IDE_{f,t} - \widehat{IDE_{f,t}}) + \epsilon_{f,t,k}$$

.

functions using the training sample, freeze the parameters, and apply the same functions in the test set to make predictions. For details on neural networks and their application, please refer to Appendix K.

The weighted ensemble approach combines predictions from multiple models (e.g., boosting, neural networks) using a stacking technique across N levels. This method involves training a meta-model that uses both the predictions from the base models and the raw input features to produce a final prediction. The stacking process is iteratively applied for N levels. For example, when $N = 1$, let $X_{n,d}$ be the $n \times d$ matrix of raw inputs and $Y_{n,M}$ be the $n \times M$ matrix of predictions from the ensemble's M base models. The ensemble then maps the concatenated matrix of $X_{n,d}$ and $Y_{n,M}$ to a set of n final predictions.

Given a forward-looking interval $T = 5$, and a model m, we optimize the mean-squared error (MSE) over all training samples. The input variables are $X_{f,t}$: the IDE values, capital stock, employment, current profit, profitability, asset growth, and the fixed effects year, industry code, and industry category. All continuous independent variables are standardized within the year and industry category. Then we choose the model that has the highest R^2 in the validation set and study it further using the test set.

Insert table *A*18 here.

Table A18 shows that the weighted ensemble model with 3 layers of stacking performs the best on the validation set. Therefore, we focus on the test performance of this model. We observe that the test set R^2 of this model is 12.7%. This R^2 is significantly larger than 0, indicating that this model has systematic out-of-sample predictive power.

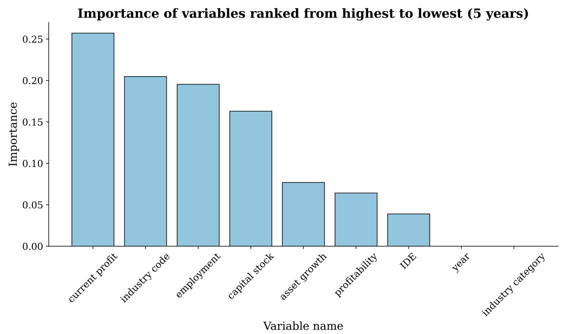
The more important question is to examine how much our IDE measure contributes to this learning and whether this increment is statistically significant. We measure the importance of each variable in the best out-of-sample (test set) predictor of profit growth. The importance of each variable is measured using a permutation-based ablation test. In particular, we take a 5000-data point random sample from the testing data, and for each independent variable, we randomly shuffle the order of its values in this subset. Then, we compute the drop in the prediction R^2 when the perturbed data is fed into the trained model. We repeat this process 5 times with different subsamples. We compute the R^2 of these 5 subsamples and compute the means and standard errors of these perturbed R^2s. For each variable, the mean drop in R^2 compared to the original data is used as its raw importance value. We then scale the importance of all variables so they sum up to 1, which are the values shown in figure 10. Then, we perform a t-test on the means and standard errors of the full test data and the perturbed samples to compute the statistical significance of including each variable in the best predictive model. Intuitively, this method allows us to measure the performance loss when one variable is lost (completely noisy) and evaluate the statistical significance using the accuracies before and after random shuffling.

As shown in figure 10, current profit and firm ID are the most important variables since the profit growth of firms likely depends on their identities and their current profit level.[19] Moreover, we observe that our IDE measure is significantly important at the 1% level and is ranked in the middle of all variables. This shows that displacement events caused by other firms' innovations have a significant role in the long-term growth of the focal firm.[20]

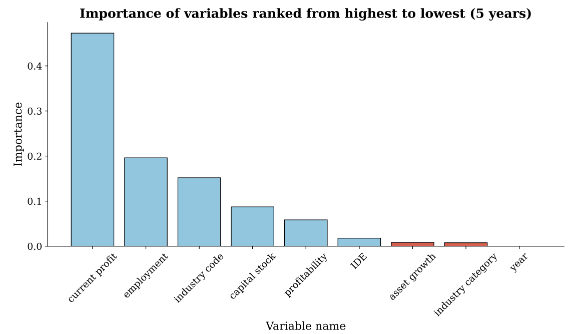
In addition, we test our best fully predictive model on a few subsamples. We confirm that this result – IDE is statistically significantly predictive of profit growth – is robust to sample selection. In particular, we run this analysis on firms with high asset growth, firms with low asset growth, firms with low profitability, and firms with high profitability each year and industry category. The ablation tests are run on random subsets of 2000 test data points because the overall test sample sizes are halved in each subsample. The importance values are shown in figure 11.

[19]We include the firm's ID as a categorical input so the model can use the all of the data points in the training set (past data) when generating a profit growth prediction.

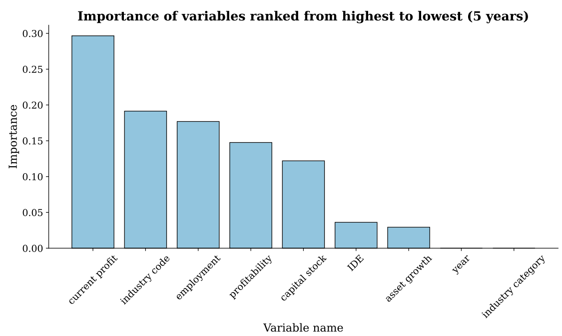
[20]It is expected that the year variable is not important in forecasting profit growth because the years in the testing sample (2013 to 2015) never appeared in the training set. Therefore, the trained models cannot use a categorical variable that takes unseen values to make predictions.



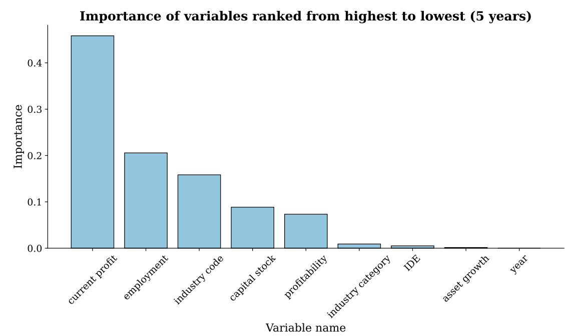
(a) High Profitability



(b) Low Profitability



(c) High Asset Growth



(d) Low Asset Growth

Figure 11: These figures show the feature importance of the best predictive model when tested on a variety of subsamples. The importance values are computed using a permutation-based ablation test. We draw 5 random 2000-data-point samples from the test set and, for each variable, shuffle its values within this subset. We then measure the decrease in prediction R^2 when the perturbed data is input into the trained model. For each variable, the mean R^2 drop serves as the raw importance value, which is subsequently scaled so that all importance values sum to one.

8 Expanding the IDE dataset using LLMs

Our main technology-based IDE measure utilizes 10-K text data which has been available since 2001 for firms reporting in the United States. The main difficulty of expanding this dataset is extracting the hidden technology-related information for firm-year pairs without 10-Ks. We leverage the vast knowledge base of Large Language Models (GPT4o) to generate technology stacks of this complementary dataset.

8.1 Constructing LLM-based IDE

To enhance the generative capacity of the model, we employ the larger GPT4o model rather than GPT4o-mini for this task.

Our first step is to construct technology descriptions using GPT4o. For each firm i in year t, we generate a summary of the technologies it utilized by applying the following prompt:

> *Instructions: You are an economist analyzing companies' technologies. In three paragraphs, describe the technological stack of name of i in the year t. Focus on technologies that the company was using in its day-to-day operations and research. Be as specific as possible and give a lot of details. For each technology, mention whether or not it was considered legacy in the field in t and whether or not there were disruptive technologies that could replace it. If there was a risk of disruption, mention companies and technologies that were threatening to the given technology. Do not use information that became public after the year t.*

We repeat this process for each year from $t - 4$ through t, concatenate the resulting summaries, and apply the same text-embedding model used for the 10-K-based IDE measure to construct the technology embedding for firm i in year t.

The next step is to construct the firm's innovation stack for year t. To ensure consistency across descriptions, we again rely on GPT4o for generating summaries. Given the higher computational cost of using GPT4o, we introduce a sampling procedure when selecting patents for inclusion. Specifically, we proceed as follows: i) If a firm i has at most 10 patents in year t, we take the abstracts of all patents issued in t. ii) If a firm i has more than 10 patents but at most 100, we select a random subset of 10 patents issued in t. iii) If a firm i has more than 100 patents, we take a random 10% subset of all patents issued in t.

This approach balances computational efficiency with the need for representative innovation descriptions while maintaining comparability with the LLM-generated technology descriptions.

After selecting the set of patents, we use GPT4o to summarize firm i's patents in year t following the prompt:

> *You are an economist studying firms' innovations. Summarize the following patent abstracts. First, give a summary of the common topics covered by many abstracts, then focus on individual patents.*
>
> *{Selected patent texts.}*

Next, we generate innovation summaries for firm i in each year $t \in \{t-4, \cdots, t-1\}$ and concatenate the summaries in these 5 years chronologically from the earliest to the latest to form the innovation stack of firm i in year t. Finally, we extract a numerical embedding based on each concatenated innovation summary using a text embedding model.

Finally, we use the resulting technology and innovation embeddings and firm-level innovation values to compute IDE using the formula defined by equation 1.

8.2 LLM-based IDE versus profit growth

Similar to our main analysis in section 3, we test whether this LLM-based IDE can capture information related to firms' profit growth. For consistency, we run the same specifications defined in equations 2 and 3.

Insert tables *A*19 and *A*20 here.

As shown by tables A19 and A20, the LLM-based IDE is significantly negatively associated with the focal firm's profit growth in all of the next 7 years. More specifically, each standard deviation increase in IDE is associated with a 1.5% decrease in profit growth by the next year and this gap grows to 6.6% by the end of year 7. The widening of this gap is statistically significant at 0.9% per additional year. This shows that the expanded IDE dataset can also capture information related to profit growth.

9 Conclusion

We develop a text-based measure of Innovation Displacement Exposure (IDE) and study its association with firms' growth. Our measure is constructed using the text of patents and 10-Ks, and it can be applied to all firm-year pairs with a recorded 10-K file. We find that the focal firm's IDE is significantly negatively associated with its profit growth in the next 7 years with worsening magnitudes each additional year. Moreover, we demonstrate that this negative innovation displacement is prevalent among different types of firms and under different types of model specifications. Our IDE measure is also significantly negatively associated with other firm growth variables such as capital stock, employment, output, market share, and intangible assets. We further delve into one mechanism behind this negative association between IDE and profit growth using a simple "quality ladder" model and show that one motivating reason is the focal firms' exposure to technological obsolescence risk. In addition, we apply our main measure in two practical extensions.

First, we build 20 fully predictive forecasting models that take the current firm characteristics of each focal firm including IDE, and predict this firm's profit growth in the future. We find that IDE can significantly improve the performance of the best predictive model and is ranked around the middle of all variables in terms of predictive importance. In addition, we show in Appendix G that an equal-weight long-short portfolio sorting on IDE can achieve significant α. However, since our data is limited in time range, we do not consider it as part of our main findings and leave additional risk analyses for future research. Lastly, we provide one way to use LLMs such as GPT4o to expand our dataset by generating technology descriptions of firm-year pairs using the model's knowledge base instead of a well-organized data source such as 10-Ks. We show that this LLM-based IDE measure is also significantly negatively associated with the focal firm's profit growth and the gap widens over time. One caveat of this method is the potential of look-ahead biases that exist in GPT4o-generated technology descriptions. We use two approaches to alleviate this concern. First, we use SEC 10-K files to show that GPT4o is aware of the temporal ordering of technologies used by these firms. In addition, we use the technology descriptions generated by GPT4o to show that it is aware of the ordering of technologies that are mentioned in its own descriptions. We believe similar approaches can be used to evaluate the extent of look-ahead bias in other applications where Large Language Models are used to generate augmentative information based on their knowledge.

Our study demonstrates the utility of text embedding models in extracting numerical representations from patent texts and technology descriptions. The performance of our measure based on OpenAI's text-embedding-3-large suggests that it effectively captures relationships between technologies and innovations, encompassing both similarities and distinctions. As advancements in Large Language Models continue, future research may uncover even greater potential in applying these embeddings to analyze technological change.

This paper expands the agenda for examining firms' exposure to displacement caused by external innovations. A key area for future study is identifying the sources of variation in the relationship between innovation displacement exposure and profit growth. Additionally, developing forward-looking measures to predict potential displacement events could provide practical insights for firms and researchers alike. The IDE measure introduced here also raises questions about its implications for asset pricing, such as its role in constructing risk factors. Furthermore, while we outline initial approaches for addressing look-ahead bias in LLM-based analyses, generalizing these methods to broader applications and refining systematic evaluation frameworks remain important directions for future work. Together, these efforts may deepen our understanding of how innovation influences firm outcomes and market dynamics.

References

Acemoglu, Daron, Ufuk Akcigit, Harun Alp, Nicholas Bloom, and William Kerr. 2018. "Innovation, reallocation, and growth." *American Economic Review*, 108(11): 3450–3491.

Aghion, Philippe, and Peter Howitt. 1992. "Growth Through Creative Destruction." *Econometrica*, 60(2): 323–51.

Alexopoulos, Michelle. 2011. "Read all about it!! What happens following a technology shock?" *American Economic Review*, 101(4): 1144–1179.

Athey, Susan, and Guido W Imbens. 2019. "Machine Learning Methods that Economists Should Know About." *Annual Review of Economics*, 11: 685–725.

Chen, Mark A, Qinxi Wu, and Baozhong Yang. 2019. "How valuable is FinTech innovation?" *The Review of Financial Studies*, 32(5): 2062–2106.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal*, 21(1): C1–C68.

Eisfeldt, Andrea L, Edward Kim, and Dimitris Papanikolaou. 2020. "Intangible value." National Bureau of Economic Research.

Erel, Isil, Léa H Stern, Chenhao Tan, and Michael S Weisbach. 2021. "Selecting directors using machine learning." *Review of Financial Studies*, 34(7): 3226–3264.

Erickson, Nick, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. "Autogluon-tabular: Robust and accurate automl for structured data." *arXiv preprint arXiv:2003.06505*.

Fama, Eugene F, and Kenneth R French. 2015. "A five-factor asset pricing model." *Journal of Financial Economics*, 116(1): 1–22.

Fedyk, Anastassia, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier. 2024. "ChatGPT and Perception Biases in Investments: An Experimental Study." *Available at SSRN 4787249*.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as data." *Journal of Economic Literature*, 57(3): 535–74.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and deliberation within the FOMC: a computational linguistics approach." *Quarterly Journal of Economics*, 133(2): 801–870.

Hoberg, Gerard, and Gordon Phillips. 2016. "Text-based network industries and endogenous product differentiation." *Journal of Political Economy*, 124(5): 1423–1465.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. "Measuring technological innovation over the long run." *American Economic Review: Insights*, 3(3): 303–20.

Klette, Tor Jakob, and Samuel Kortum. 2004. "Innovating firms and aggregate innovation." *Journal of Political Economy*, 112(5): 986–1018.

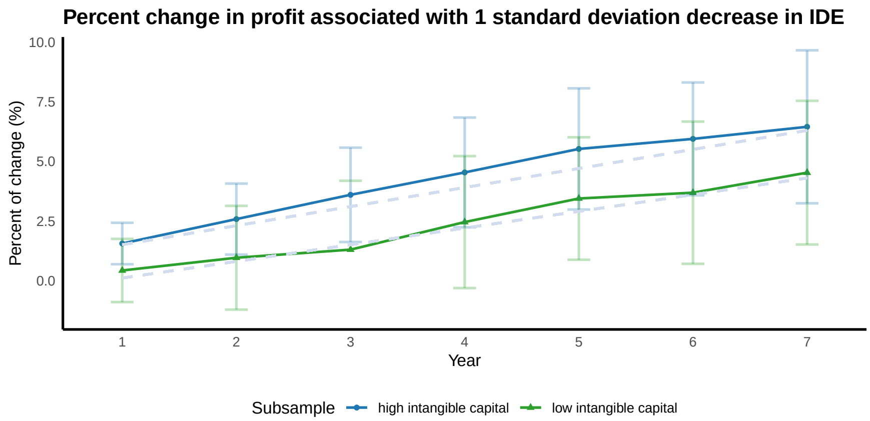
Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman. 2017. "Technological innovation, resource allocation, and growth." *The Quarterly Journal of Economics*, 132(2): 665–712.

Kogan, Leonid, Dimitris Papanikolaou, and Noah Stoffman. 2020. "Left behind: Creative destruction, inequality, and the stock market." *Journal of Political Economy*, 128(3): 855–906.

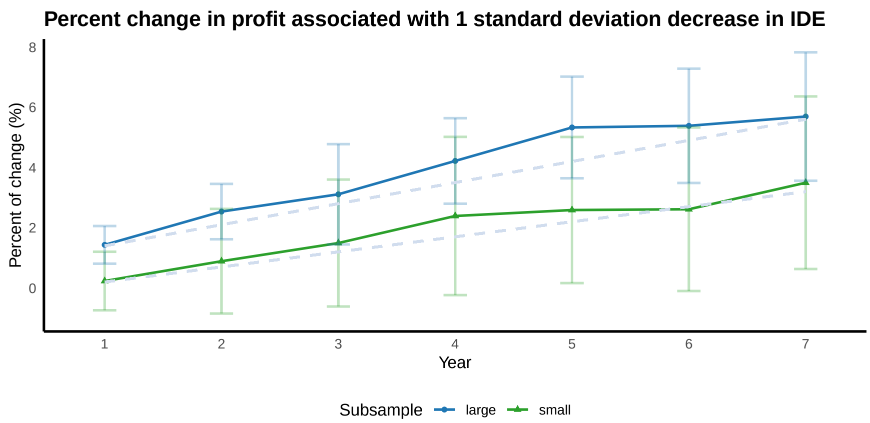
Kortum, Samuel. 1993. "Equilibrium r&d and the patent–r&d ratio: Us evidence." *The American Economic Review*, 83(2): 450–457.

Kortum, Samuel, and Josh Lerner. 1998. "Stronger protection or technological revolution: what is behind the recent surge in patenting?" Vol. 48, 247–304, Elsevier.

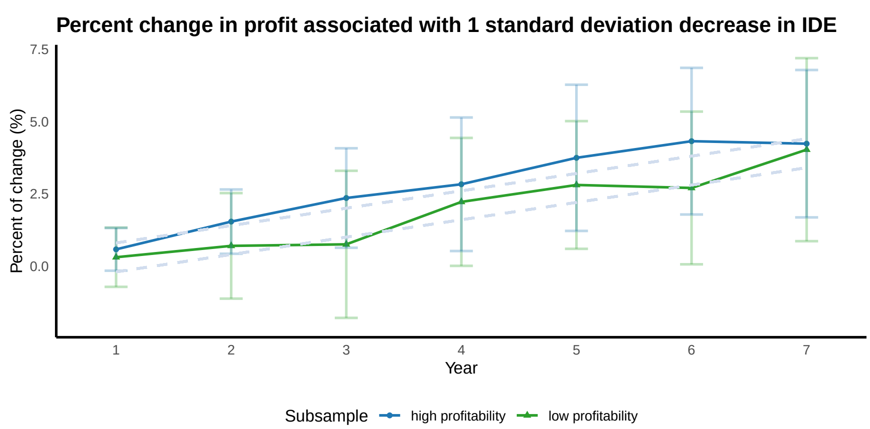
Shea, John. 1998. "What do technology shocks do?" *NBER macroeconomics annual*, 13: 275–310.



(a) Low versus high relative intangible capital.



(b) Large versus small market capitalization.



(c) Low versus high profitability.

Figure 6: Subsample analysis of various factors on firm performance. Each subfigure highlights a different factor: (a) Relative intangible capital, (b) Firm Size, and (c) Profitability. The coefficients are estimated using equation 2, and the subsamples are divided by the median in each industry category and year.

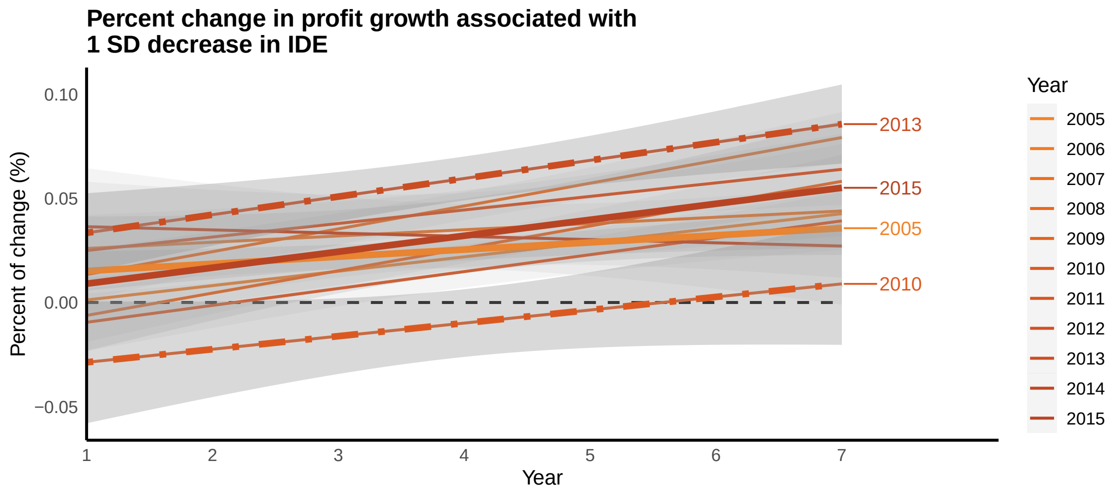


Figure 7: This figure shows the association between IDE and the focal firm's profit growth over the next 7 years, where the coefficients are estimated separately for each year. The estimation specification is shown in equation 10. The lines in the figure are linear regression lines fitted across the association between IDE and profit growth in the next 7 years. The horizontal dash line signals profit growth of 0%.

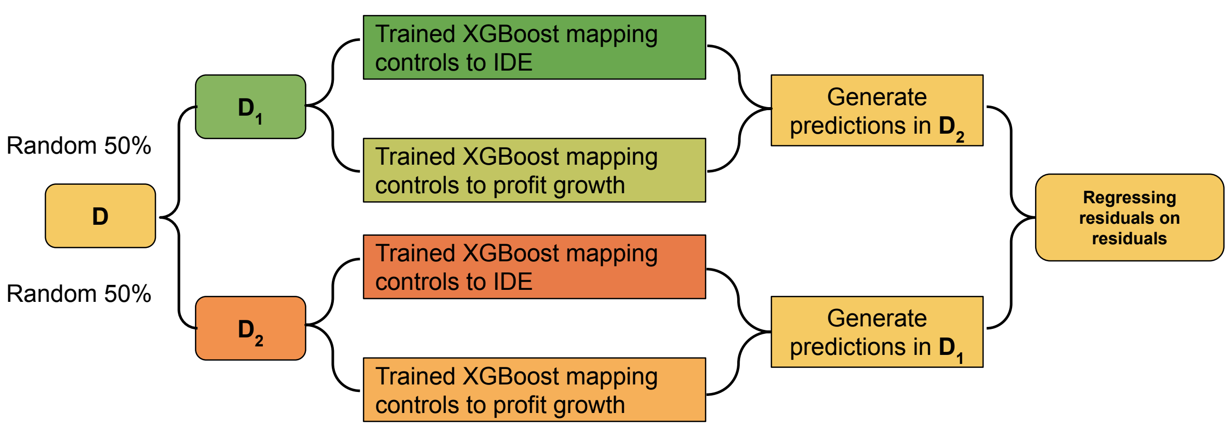


Figure 8: The diagram of debiased machine learning applied to identify the effect of displacement (IDE) on profit growth

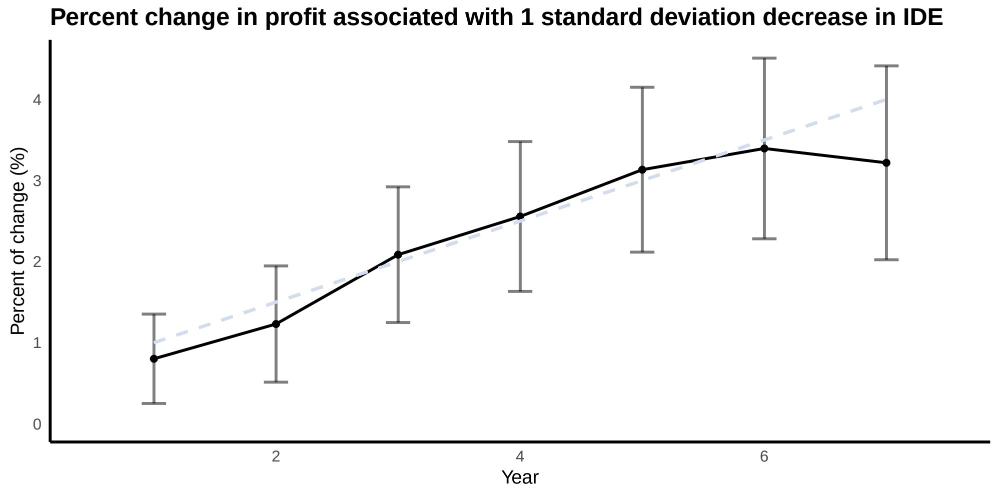


Figure 9: This figure shows the percent change of profit associated with one standard deviation **decrease** in displacement (using the Debiased Machine Learning approach). Note that displacement is negatively associated with profit growth. The blue line represents the widening of the gap.

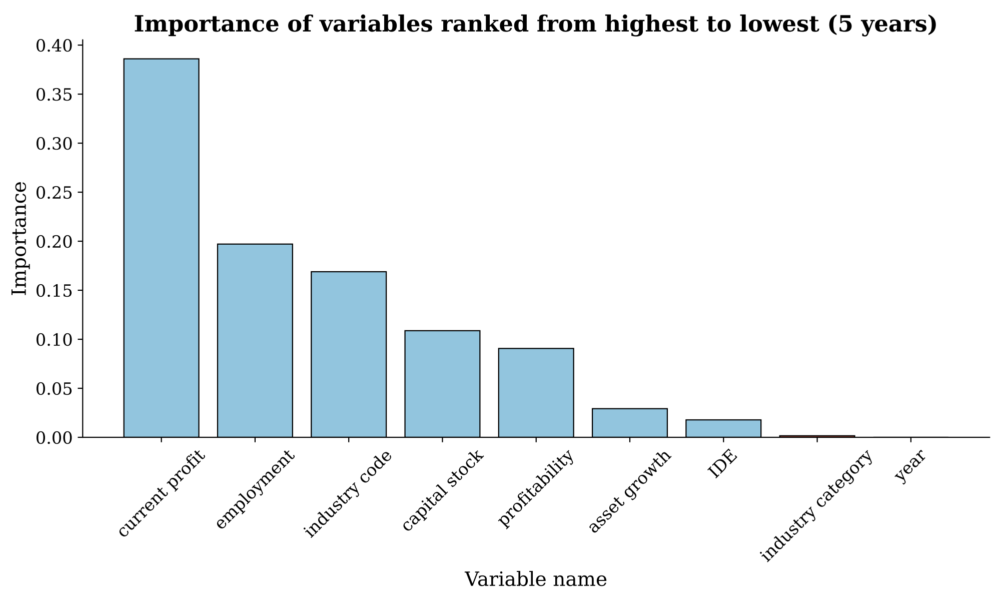


Figure 10: This figure shows the importance of different variables in the best predictive model of profit growth in the next 5 years. The importance values are scaled to sum to 1, and the ones colored blue are significant at the 5% level. The importance values are computed using a permutation-based ablation test. We draw 5 random 5000-data-point samples from the test set and, for each variable, shuffle its values within this subset. We then measure the decrease in prediction R^2 when the perturbed data is input into the trained model. For each variable, the mean R^2 drop serves as the raw importance value, which is subsequently scaled so that all importance values sum to one.

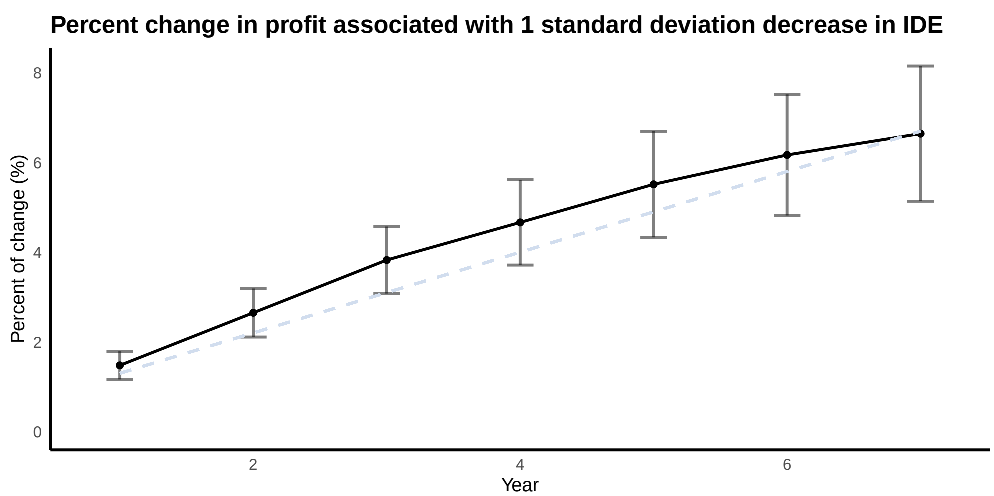


Figure 12: This figure shows the percent change of profit associated with one standard deviation **decrease** in displacement (LLM-based) estimated using equation 2. Note that displacement is negatively associated with profit growth. The blue line represents the widening of the gap estimated using equation 3.

Online Appendix

1. Appendix A includes all of the tables used in this paper.
2. Appendix B shows the exact prompt we use to extract technology summaries and discussions about obsolescence risk from 10-K files. We also show the sentence pairs we use the construct the semantic axis for measuring obsolescence risk.
3. Appendix C shows an example of a technology summary of Apple.
4. Appendix D shows an example of an innovation summary.
5. Appendix E includes a model where we discuss the different between IDE and market fluidity threats.
6. Appendix F shows a comparison between the industry-level KPSS measure developed in Kogan et al. (2017) and our IDE measure.
7. Appendix G shows the return-generating potential of an IDE-based long-short portfolio.
8. Appendix H introduces a modified IDE measure where similarities scale nonlinearly with respect to cosine.
9. Appendix I discusses more details about the construction of technology and innovation embeddings using an LLM (GPT4o).
10. Appendix J includes details about validity checks against look-ahead bias in GPT4o.
11. Appendix K discusses the basic architecture of neural networks and the GPT family of models.
12. Appendix L discusses the details of debiased machine learning.
13. Appendix M introduces gradient boosting and the XGBoost model.
14. Appendix N discussion the association between our updated total similarity measure and firms' growth in market share, output, employment, capital stock, and intangible assets.
15. Appendix O outlines the exact steps we take to construct our main dataset.

A Tables

Variable	Observations	Mean	Standard deviation	Min	25th percentile	Median	75th percentile	Max
one-year profit growth	29716	0.019	0.432	-6.989	-0.086	0.034	1.5000e-01	4.992000e+00
seven-year profit growth	19942	0.180	0.828	-8.529	-0.197	0.190	5.7800e-01	6.268000e+00
market capitalization	31965	4686250.018	19690451.824	202.650	123117.160	552129.500	2.2786e+06	6.431201e+08
log of asset growth	32193	0.067	0.276	-4.549	-0.042	0.045	1.4800e-01	4.059000e+00
profitability	32193	0.369	0.271	0.000	0.198	0.314	4.6900e-01	1.294400e+01
log of employment	31415	0.726	2.127	-6.908	-0.753	0.762	2.1970e+00	7.741000e+00
log of capital stock	31995	0.896	2.471	-10.845	-0.851	0.902	2.5900e+00	8.504000e+00
log of profit	32193	5.308	2.099	-5.459	3.917	5.315	6.6740e+00	1.166400e+01

TABLE A1: Summary Statistics of our dataset from 2005 to 2015 without filtering. One and seven year profit growth are defined as the log profit 1 and 7 years in the future minus the log profit of this year. Asset growth is computed as the ratio between the total asset in year t and $t-1$, and profitability is defined as total sales minus cost of goods sold divided by total asset.

	Dependent variable:						
	$\Pi_{t+1} - \Pi_t$	$\Pi_{t+2} - \Pi_t$	$\Pi_{t+3} - \Pi_t$	$\Pi_{t+4} - \Pi_t$	$\Pi_{t+5} - \Pi_t$	$\Pi_{t+6} - \Pi_t$	$\Pi_{t+7} - \Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	-0.009**	-0.017**	-0.022**	-0.033***	-0.040***	-0.041***	-0.047***
	(0.004)	(0.006)	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)
profitability t	-0.018**	-0.030***	-0.040***	-0.050***	-0.058***	-0.064***	-0.068***
	(0.006)	(0.008)	(0.009)	(0.010)	(0.013)	(0.014)	(0.017)
log asset growth	0.064***	0.079***	0.091***	0.091***	0.096***	0.104***	0.111***
	(0.005)	(0.008)	(0.007)	(0.009)	(0.011)	(0.010)	(0.011)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Year×*IndCategory*	✓	✓	✓	✓	✓	✓	✓
Observations	25,810	23,908	22,306	20,875	19,591	18,439	17,351
R^2	0.125	0.138	0.133	0.132	0.142	0.149	0.160
Adjusted R^2	0.107	0.118	0.112	0.109	0.119	0.124	0.133
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE A2: This table shows the association between IDE and profit growth. We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total asset in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total asset.

	Dependent variable:						
	$\Pi_{t+1} - \Pi_t$	$\Pi_{t+2} - \Pi_t$	$\Pi_{t+3} - \Pi_t$	$\Pi_{t+4} - \Pi_t$	$\Pi_{t+5} - \Pi_t$	$\Pi_{t+6} - \Pi_t$	$\Pi_{t+7} - \Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	-0.008**	-0.015**	-0.023**	-0.033***	-0.041***	-0.044***	-0.047***
	(0.003)	(0.005)	(0.008)	(0.008)	(0.008)	(0.009)	(0.010)
profitability t	-0.003	-0.013	-0.022	-0.030*	-0.034*	-0.039*	-0.045*
	(0.009)	(0.009)	(0.012)	(0.015)	(0.018)	(0.021)	(0.021)
log asset growth	0.066***	0.083***	0.096***	0.097***	0.104***	0.111***	0.119***
	(0.004)	(0.007)	(0.007)	(0.010)	(0.012)	(0.010)	(0.013)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Year×*IndCategory*	✓	✓	✓	✓	✓	✓	✓
Observations	25,958	24,044	22,426	20,985	19,690	18,528	17,434
R^2	0.128	0.140	0.136	0.133	0.143	0.150	0.162
Adjusted R^2	0.109	0.120	0.115	0.110	0.119	0.125	0.136
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE A3: This table shows the association between IDE and profit growth. We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized across the entire sample after demeaning by each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total asset in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total asset.

	Dependent variable:	
	Baseline	Additional percent change in profit per year
	(1)	(2)
IDE	-0.001	-0.007***
	(0.004)	(0.001)
Year		✓
Ind Code		✓
Year × *IndCategory*		✓
Observations		148,280
R^2		0.142
Adjusted R^2		0.117
Note:		*p<0.1; **p<0.05; ***p<0.01

TABLE A4: This table shows the association between IDE and profit growth over time (stacking the data from years 1 to 7). We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. We also control for the interaction between the number of years foward and these independent variables. In addition, we use fixed effects of year, industry code, year cross industry category, and the number of years forward interacted with all of them. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

	Dependent variable:
	IDE
Log capital stock	0.593***
	(0.045)
Log employment	-0.008
	(0.039)
Log profit	0.013
	(0.041)
Profitability	0.082***
	(0.012)
Log asset growth	-0.051***
	(0.009)
Total assets	-0.059**
	(0.022)
Capital expenditures	-0.459***
	(0.041)
Research and development expenditures	0.063**
	(0.024)
Market cap	0.044*
	(0.023)
Year	✓
Ind Code	✓
Year×*IndCategory*	✓
Observations	17,675
R^2	0.153
Adjusted R^2	0.128
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

TABLE A5: This table shows the association between firm characteristics and our displacement measure. The data spans from 2005 to 2015. We use fixed effects for year, industry code, and year interacted with industry category. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. An asinh transformation is applied to capital and R&D expenditures to stablize the variance and allow for non-positive values.

	Dependent variable:						
	$\Pi_{t+1} - \Pi_t$	$\Pi_{t+2} - \Pi_t$	$\Pi_{t+3} - \Pi_t$	$\Pi_{t+4} - \Pi_t$	$\Pi_{t+5} - \Pi_t$	$\Pi_{t+6} - \Pi_t$	$\Pi_{t+7} - \Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	-0.004	-0.013*	-0.019*	-0.026**	-0.032***	-0.032**	-0.032**
	(0.004)	(0.006)	(0.009)	(0.009)	(0.010)	(0.011)	(0.012)
profitability t	-0.019**	-0.033***	-0.040***	-0.045***	-0.055***	-0.058***	-0.064***
	(0.006)	(0.008)	(0.009)	(0.011)	(0.014)	(0.017)	(0.019)
log asset growth	0.056***	0.076***	0.096***	0.098***	0.107***	0.117***	0.120***
	(0.005)	(0.009)	(0.009)	(0.011)	(0.013)	(0.012)	(0.013)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Year×*IndCategory*	✓	✓	✓	✓	✓	✓	✓
Observations	16,365	15,144	14,112	13,190	12,364	11,619	10,948
R^2	0.107	0.129	0.134	0.132	0.147	0.158	0.165
Adjusted R^2	0.080	0.100	0.103	0.099	0.112	0.122	0.126
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE A6: This table shows the association between IDE and profit growth after we remove the part of IDE that can be explained by traditional firm characteristics. We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

	Dependent variable:	
	IDE	
	$t+5$	$t+10$
Capital expenditure	0.260***	0.002
	(0.059)	(0.043)
R&D expenditure	-0.123	0.104**
	(0.071)	(0.045)
Asset total	0.009	0.564*
	(0.049)	(0.271)
Profitability	-0.322***	-0.541*
	(0.082)	(0.272)
Year	✓	✓
Ind Code	✓	✓
Year×Ind Category	✓	✓
Observations	14,796	11,583
R^2	0.160	0.175
Adjusted R^2	0.131	0.140
Note:	*p<0.1; **p<0.05; ***p<0.01	

TABLE A7: This table shows the association between future IDE and firm characteristics after we remove the first two principal components from the firm characteristics. We use fixed effects for year, industry code, and year interacted with industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level.

	Dependent variable:						
	$\Pi_{t+1} - \Pi_t$	$\Pi_{t+2} - \Pi_t$	$\Pi_{t+3} - \Pi_t$	$\Pi_{t+4} - \Pi_t$	$\Pi_{t+5} - \Pi_t$	$\Pi_{t+6} - \Pi_t$	$\Pi_{t+7} - \Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	-0.011	-0.024**	-0.036**	-0.051***	-0.064***	-0.056**	-0.057**
	(0.007)	(0.010)	(0.012)	(0.013)	(0.017)	(0.019)	(0.018)
profitability t	-0.021*	-0.041***	-0.042**	-0.040*	-0.031	-0.026	-0.022
	(0.010)	(0.011)	(0.014)	(0.018)	(0.022)	(0.025)	(0.028)
log asset growth	0.052***	0.076***	0.098***	0.102***	0.116***	0.125***	0.134***
	(0.009)	(0.013)	(0.014)	(0.015)	(0.017)	(0.018)	(0.018)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Year×*IndCategory*	✓	✓	✓	✓	✓	✓	✓
Observations	9,368	8,816	8,313	7,862	7,449	7,072	6,707
R^2	0.134	0.168	0.173	0.167	0.184	0.189	0.195
Adjusted R^2	0.089	0.122	0.124	0.116	0.131	0.134	0.137
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE A8: This table shows the association between IDE and profit growth among innovative firms (those who received at least 10 patents over our time interval). We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

	Dependent variable:						
	$\Pi_{t+1}-\Pi_t$	$\Pi_{t+2}-\Pi_t$	$\Pi_{t+3}-\Pi_t$	$\Pi_{t+4}-\Pi_t$	$\Pi_{t+5}-\Pi_t$	$\Pi_{t+6}-\Pi_t$	$\Pi_{t+7}-\Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	-0.011**	-0.019**	-0.024**	-0.033**	-0.038***	-0.040***	-0.047***
	(0.004)	(0.008)	(0.010)	(0.011)	(0.012)	(0.012)	(0.012)
profitability t	-0.017**	-0.024**	-0.037***	-0.052***	-0.069***	-0.082***	-0.094***
	(0.006)	(0.010)	(0.010)	(0.011)	(0.013)	(0.016)	(0.020)
log asset growth	0.069***	0.079***	0.084***	0.082***	0.081***	0.086***	0.093***
	(0.006)	(0.007)	(0.008)	(0.011)	(0.012)	(0.012)	(0.014)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Year$\times$*IndCategory*	✓	✓	✓	✓	✓	✓	✓
Observations	16,442	15,092	13,993	13,013	12,142	11,367	10,644
R^2	0.142	0.152	0.148	0.153	0.165	0.177	0.194
Adjusted R^2	0.113	0.121	0.115	0.118	0.127	0.137	0.153
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE A9: This table shows the association between IDE and profit growth among non-innovative firms (those who received fewer than 10 patents over our time interval). We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

	raw measures	standardized measures
Pearson Correlation	0.88^{***}	0.25^{***}
Kendall-Tau correlation	0.67^{***}	0.15^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01

TABLE A10: This table shows the correlation between the two measures: one based on patents and technologies, and an alternative purely based on patents. The left column shows the Pearson and Kendall-Tau correlation between the two measures as their raw values, and the right column shows the correlation between the two measures once standardized in each year and industry category. The sample is for innovative firms where the alternative patent-based measure is well-defined.

	Dependent variable:						
	$\Pi_{t+1} - \Pi_t$	$\Pi_{t+2} - \Pi_t$	$\Pi_{t+3} - \Pi_t$	$\Pi_{t+4} - \Pi_t$	$\Pi_{t+5} - \Pi_t$	$\Pi_{t+6} - \Pi_t$	$\Pi_{t+7} - \Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Technology IDE	-0.005	-0.014*	-0.024**	-0.036***	-0.046***	-0.039**	-0.040**
	(0.005)	(0.008)	(0.009)	(0.011)	(0.014)	(0.015)	(0.016)
Innovation IDE	-0.014**	-0.023***	-0.026**	-0.025*	-0.021	-0.024	-0.020
	(0.005)	(0.007)	(0.009)	(0.012)	(0.016)	(0.019)	(0.021)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Year$\times$*IndCategory*	✓	✓	✓	✓	✓	✓	✓
Observations	9,352	8,800	8,297	7,846	7,433	7,056	6,691
R^2	0.134	0.167	0.172	0.166	0.182	0.188	0.193
Adjusted R^2	0.090	0.123	0.125	0.115	0.130	0.134	0.136
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE A11: This table shows the association between the original and alternative IDE measures and profit growth. We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

	Dependent variable:						
	$\Pi_{t+1} - \Pi_t$	$\Pi_{t+2} - \Pi_t$	$\Pi_{t+3} - \Pi_t$	$\Pi_{t+4} - \Pi_t$	$\Pi_{t+5} - \Pi_t$	$\Pi_{t+6} - \Pi_t$	$\Pi_{t+7} - \Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	-0.008**	-0.016**	-0.021**	-0.030***	-0.037***	-0.039***	-0.045***
	(0.003)	(0.006)	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)
total similarity	-0.006	-0.017***	-0.021***	-0.025***	-0.029***	-0.030***	-0.031***
	(0.004)	(0.005)	(0.006)	(0.007)	(0.006)	(0.007)	(0.008)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Year×*IndCategory*	✓	✓	✓	✓	✓	✓	✓
Observations	25,568	23,692	22,113	20,703	19,434	18,294	17,216
R^2	0.126	0.138	0.134	0.133	0.143	0.149	0.161
Adjusted R^2	0.107	0.119	0.113	0.110	0.119	0.124	0.134
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE A12: This table shows the association between IDE and profit growth and profit growth versus total similarity (horizontal product competition) versus profit growth. We control for the log of asset growth, current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and total similarity is computed using equation 7.

Output						
(1)	(2)	(3)	(4)	(5)	(6)	(7)
-0.009***	-0.016***	-0.022***	-0.029***	-0.038***	-0.042***	-0.044***
(0.003)	(0.004)	(0.005)	(0.006)	(0.007)	(0.009)	(0.008)

Capital stock						
(1)	(2)	(3)	(4)	(5)	(6)	(7)
-0.020***	-0.035***	-0.045***	-0.057***	-0.067***	-0.072***	-0.044***
(0.003)	(0.004)	(0.005)	(0.005)	(0.007)	(0.010)	(0.008)

Employment						
(1)	(2)	(3)	(4)	(5)	(6)	(7)
-0.005**	-0.008*	-0.011**	-0.016**	-0.023***	-0.023**	-0.022**
(0.002)	(0.004)	(0.004)	(0.006)	(0.007)	(0.008)	(0.008)

Intangible asset						
(1)	(2)	(3)	(4)	(5)	(6)	(7)
-0.039***	-0.048***	-0.057***	-0.045***	-0.049***	-0.057***	-0.065***
(0.008)	(0.009)	(0.011)	(0.012)	(0.013)	(0.012)	(0.014)

Market share						
(1)	(2)	(3)	(4)	(5)	(6)	(7)
-0.008*	-0.018***	-0.024***	-0.033***	-0.043***	-0.051***	-0.054***
(0.004)	(0.005)	(0.006)	(0.006)	(0.008)	(0.011)	(0.011)

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE A13: This table shows the association between IDE and firm growth outcomes other than profit growth. In particular, we run the regression specified in equation 2 while replacing profit growth with growth in capital stock, output, employment, intangible assets ($\frac{\text{intangible asset}}{\text{book equity}}$), and market share. We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

	Dependent variable:						
	$\Pi_{t+1} - \Pi_t$	$\Pi_{t+2} - \Pi_t$	$\Pi_{t+3} - \Pi_t$	$\Pi_{t+4} - \Pi_t$	$\Pi_{t+5} - \Pi_t$	$\Pi_{t+6} - \Pi_t$	$\Pi_{t+7} - \Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	-0.008***	-0.012***	-0.021***	-0.026***	-0.031***	-0.034***	-0.032***
	(0.003)	(0.004)	(0.004)	(0.005)	(0.005)	(0.006)	(0.006)
Observations	17,350	17,350	17,350	17,350	17,350	17,350	17,350
R^2	0.0005	0.001	0.001	0.002	0.002	0.002	0.002
Adjusted R^2	0.0004	0.001	0.001	0.002	0.002	0.002	0.002
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE A14: This table shows the association between IDE and profit growth identified by an XGBoost-based debiased machine learning algorithm. The data spans from 2005 to 2015. We remove the following confounders from both profit growth and IDE: log of profit, employment, and capital stock, firm and industry-level innovation values, industry code, industry category, and year. All independent variables are standardized within each industry category and year. All fixed effects are included in the XGBoost model.

	Dependent variable:	
	Baseline	Additional percent change in profit per year
	(1)	(2)
IDE	-0.005	-0.005***
	(0.004)	(0.001)
Observations	121,450	
R^2	0.002	
Adjusted R^2	0.002	
Note:		*p<0.1; **p<0.05; ***p<0.01

TABLE A15: This table shows the growth of the effect of IDE on profit growth over time using the debiased machine learning approach and XGBoost. The data spans from 2005 to 2015. We remove the following confounders from both profit growth and IDE: log of profit, employment, and capital stock, firm and industry-level innovation values, industry code, industry category, and year. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level.

	Dependent variable:						
	$risk_t$	$risk_{t+1}$	$risk_{t+2}$	$risk_{t+3}$	$risk_{t+4}$	$risk_{t+5}$	$risk_{t+6}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	0.023***	0.020***	0.023**	0.028**	0.036*	0.034**	0.034*
	(0.006)	(0.006)	(0.008)	(0.008)	(0.016)	(0.012)	(0.013)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Year×*IndCategory*	✓	✓	✓	✓	✓	✓	✓
Observations	27,631	23,572	19,835	16,586	13,702	11,051	8,715
R^2	0.009	0.014	0.017	0.024	0.030	0.036	0.042
Adjusted R^2	-0.010	-0.008	-0.007	-0.004	-0.001	0.001	0.0005
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE A16: This table shows the association between IDE and obsolescence risk. We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total asset in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total asset.

	Dependent variable:						
	risk_t	risk_{t+1}	risk_{t+2}	risk_{t+3}	risk_{t+4}	risk_{t+5}	risk_{t+6}
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	0.044**	0.060**	0.036	0.026	0.039	0.043	0.012
	(0.018)	(0.019)	(0.022)	(0.019)	(0.021)	(0.037)	(0.035)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Year×*IndCategory*	✓	✓	✓	✓	✓	✓	✓
Observations	17,532	14,841	12,283	9,976	7,944	6,104	4,474
R^2	0.041	0.059	0.061	0.072	0.090	0.093	0.107
Adjusted R^2	0.013	0.028	0.026	0.032	0.044	0.037	0.038

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE A17: This table shows the association between IDE and obsolescence risk. We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, capital stock, and the self-expressed obsolescence risk from 5 years ago. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total asset in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total asset.

Model	Score Test	Score Val
WeightedEnsemble_L2	0.147784	0.363597
LightGBMLarge_BAG_L1	0.133747	0.321098
NeuralNetFastAI_BAG_L1	0.129864	0.276889
LightGBM_BAG_L2	0.128839	0.363449
CatBoost_BAG_L2	0.128357	0.372942
WeightedEnsemble_L3	0.126655	0.381668
ExtraTreesMSE_BAG_L2	0.125240	0.370709
XGBoost_BAG_L2	0.121152	0.358142
LightGBMXT_BAG_L1	0.119270	0.255578
LightGBMLarge_BAG_L2	0.119132	0.351741
NeuralNetFastAI_BAG_L2	0.118969	0.351699
LightGBM_BAG_L1	0.109428	0.320986
NeuralNetTorch_BAG_L2	0.107529	0.371292
CatBoost_BAG_L1	0.106495	0.246066
LightGBMXT_BAG_L2	0.106198	0.363117
RandomForestMSE_BAG_L2	0.098096	0.360273
XGBoost_BAG_L1	0.095149	0.258406
RandomForestMSE_BAG_L1	0.057193	0.232832
NeuralNetTorch_BAG_L1	0.046464	0.254914
ExtraTreesMSE_BAG_L1	0.040546	0.263911
KNeighborsUnif_BAG_L1	-0.069596	-0.005157
KNeighborsDist_BAG_L1	-0.071794	0.011462

TABLE A18: Performance of various machine learning models for predicting firms' 5-year profit growth. Training and validation use data up to 2012, and the test set is 2013–2015. The scores represent the corresponding R^2 values. The input variables are log of current profit, employment, capital stock, and asset growth, and profitability, IDE, industry code, industry category, and year. All continuous input variables are standardized in each year and industry category.

	Dependent variable:						
	$\Pi_{t+1}-\Pi_t$	$\Pi_{t+2}-\Pi_t$	$\Pi_{t+3}-\Pi_t$	$\Pi_{t+4}-\Pi_t$	$\Pi_{t+5}-\Pi_t$	$\Pi_{t+6}-\Pi_t$	$\Pi_{t+7}-\Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	-0.015***	-0.027***	-0.038***	-0.047***	-0.055***	-0.062***	-0.066***
	(0.002)	(0.003)	(0.004)	(0.005)	(0.006)	(0.007)	(0.008)
profitability t	-0.012***	-0.027***	-0.040***	-0.049***	-0.057***	-0.060***	-0.065***
	(0.003)	(0.004)	(0.005)	(0.006)	(0.007)	(0.008)	(0.009)
log asset growth	0.073***	0.086***	0.096***	0.097***	0.100***	0.103***	0.103***
	(0.003)	(0.005)	(0.006)	(0.007)	(0.008)	(0.008)	(0.008)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Year×*IndCategory*	✓	✓	✓	✓	✓	✓	✓
Observations	118,542	108,628	99,837	92,033	85,137	79,008	73,460
R^2	0.101	0.119	0.129	0.134	0.144	0.147	0.152
Adjusted R^2	0.092	0.109	0.118	0.123	0.131	0.133	0.137
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE A19: This table shows the association between the LLM-based IDE and profit growth. We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 1980 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

	Dependent variable:	
	Baseline	Additional percent change in profit per year
	(1)	(2)
IDE	-0.004**	-0.009***
	(0.002)	(0.001)
Year		✓
Ind Code		✓
Year×*IndCategory*		✓
Observations		656,645
R^2		0.132
Adjusted R^2		0.119
Note:		*p<0.1; **p<0.05; ***p<0.01

TABLE A20: This table shows the association between the LLM-based IDE and profit growth over time (stacking the data from years 1 to 7). We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. We also control for the interaction between the number of years foward and these independent variables. In addition, we use fixed effects of year, industry code, year cross industry category, and the number of years forward interacted with all of them. The data spans from 1980 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

B 10-K prompts

B.1 Prompt to extract technology descriptions

You are an economist tasked with analyzing a firm's 10-K filing to extract and summarize all explicit mentions of technologies (hardware, software, platforms, processes, R&D projects, etc.) described in the text. Your goal is to identify how these technologies are used, their strategic significance, and any associated risks or partnerships. Follow these detailed instructions:

***Objective**: Thoroughly identify and summarize all explicit mentions of technologies described in the firm's 10-K filing. The focus is on documenting how these technologies are utilized, their strategic significance, and any associated risks or partnerships.*

***Scope**: Focus on the following sections of the 10-K: 1. Business 2. Risk Factors 3. Properties 4. Management's Discussion and Analysis (MD&A) 5. Notes to Financial Statements 6. Exhibits 7. Any sections specifically about R&D*

Extraction Guidelines

*1. **Identify All Technology Mentions**:*

- *Capture every explicit mention of technologies, tools, systems, platforms, hardware, software, proprietary processes, or innovation initiatives.*
- *Include technologies used for daily operations (e.g., manufacturing, logistics, retail, customer engagement) and R&D activities.*

*2. **Detailed Information for Each Technology**: For every identified technology, extract and summarize the following:*

- ***Name or Description**:*
  - *The exact name or description of the technology as stated in the text. If the name is not provided, include a brief descriptive phrase.*
- ***Function/Use Case**:*
  - *Detailed explanation of how the technology is used in daily operations or R&D, including specific applications (e.g., automation in production, data analysis in R&D, customer experience optimization).*

- ***Key Features or Capabilities***:
  - *Note any features, functionalities, or technical capabilities explicitly mentioned (e.g., AI-powered, cloud-based, automated).*
- ***Mentioned Partnerships or Suppliers***:
  - *List any disclosed third-party providers, collaborators, or suppliers associated with the technology.*
  - *Describe their role (e.g., technology provider, implementation partner).*
- ***Strategic Importance or Business Rationale***:
  - *Summarize any direct statements regarding why the technology is significant to the company's strategy, competitive positioning, or operational efficiency.*
  - *Note any goals or objectives linked to the technology's adoption (e.g., reducing costs, accelerating innovation, enhancing customer satisfaction).*
- ***Challenges or Risks***:
  - *Document any risks, vulnerabilities, or challenges associated with the technology, as explicitly stated in the text.*
- ***Historical Context or Evolution***:
  - *If mentioned, include details about when and why the company adopted the technology, how it has evolved, or its integration into the company's broader operations over time.*
- ***Impact on Financial Performance***:
  - *Extract any explicit references to the impact of the technology on the company's revenues, costs, or profitability.*
- ***Environmental, Social, and Governance (ESG) Factors***:
  - *Include any mentions of how the technology aligns with the company's ESG initiatives or sustainability goals.*

3. ***Section and Page References***:
   - *Always include the section (e.g., Business, MD&A) and page number where the information appears.*
   - *If the same technology is discussed across multiple sections, include a summary of insights from each relevant section.*

4. ***Highlight Relationships Between Technologies***:

   - *Note if the filing mentions interactions or integrations between technologies (e.g., how software platforms interact with hardware systems or proprietary tools).*

5. ***Do Not Speculate***:

   - *Focus only on details explicitly stated in the 10-K. Avoid inferring functionality, partnerships, or strategic importance not directly mentioned.*

Summary Requirements

After extracting details for each technology, conclude with:

1. *A **summary paragraph** synthesizing how these technologies collectively support the company's operations, R&D efforts, and strategic objectives as described in the 10-K.*
2. *Any recurring themes or notable patterns (e.g., focus on automation, emphasis on partnerships, reliance on proprietary tools).*

B.2 Prompt to extract discussions of obsolescence risk

You are an economist tasked with analyzing a firm's 10-K filing to extract concise information about the severity of obsolescence risk—the risk that the firm's products, services, or technologies may become outdated due to competitor innovations or emerging technologies, causing shifts in market dynamics.

Objective: *Identify and quote any explicit statements in the 10-K indicating how serious, imminent, or significant the firm considers its obsolescence risk to be. These statements should reflect the firm's own wording about the level of concern, threat, or potential impact severity.*

Relevant 10-K Sections:

- *Risk Factors*
- *Business Description*
- *Management's Discussion and Analysis (MD&A)*

- *R&D or Innovation-Related Disclosures*
- *Quantitative and Qualitative Disclosures About Market Risk*
- *Notes to Financial Statements*

Extraction Guidelines

1. ***Identify Mentions of Obsolescence Risk Severity***:
   - *Search for any direct language in the 10-K that explicitly describes how severe, critical, or urgent the obsolescence risk is (e.g., "could seriously harm our business," "poses a significant challenge").*
2. ***Use Exact Quotations***:
   - *Provide the specific sentence(s) from the 10-K that discuss the level of concern regarding obsolescence risk. Do not paraphrase; use the original wording.*
3. ***Section and Page References***:
   - *Note the specific section (e.g., Risk Factors, MD&A) and page number for each quoted sentence.*
4. ***Concise Summaries***:
   - *You may add a very short summary in your own words only if it helps clarify the quoted statement. However, keep direct quotes intact and clearly identified.*
5. ***Avoid Speculation***:
   - *Only present what is explicitly stated in the 10-K. Do not infer or interpret risk severity beyond the firm's own words.*

Summary Requirements

1. ***Overall Severity Profile***:
   - *Conclude with a brief paragraph synthesizing how severe the firm perceives its obsolescence risk to be, based on the exact language quoted.*

2. ***Notable Patterns or Themes***:

- *If multiple sections emphasize similar language indicating a high or escalating concern about obsolescence, highlight that with direct quotes where appropriate.*

B.3 Sentence pairs used to compute the obsolescence risk axis

Pair 1:

- ***Low Risk:*** *"The firm is exposed to minimal technology obsolescence risk"*
- ***High Risk:*** *"The firm is exposed to significant technology obsolescence risk"*

Pair 2:

- ***Low Risk:*** *"The company faces a low likelihood of technology becoming obsolete"*
- ***High Risk:*** *"The company faces a high likelihood of technology becoming obsolete"*

Pair 3:

- ***Low Risk:*** *"The firm encounters limited risks of technology obsolescence"*
- ***High Risk:*** *"The firm encounters severe risks of technology obsolescence"*

Pair 4:

- ***Low Risk:*** *"The company is subject to low technology obsolescence concerns"*
- ***High Risk:*** *"The company is subject to high technology obsolescence concerns"*

Pair 5:

- ***Low Risk:*** *"The firm has a minor risk of its technology becoming outdated"*
- ***High Risk:*** *"The firm has a major risk of its technology becoming outdated"*

Pair 6:

- ***Low Risk:*** *"The company experiences low exposure to risks of outdated technology"*
- ***High Risk:*** *"The company experiences high exposure to risks of outdated technology"*

Pair 7:

- ***Low Risk:*** *"The firm's technology obsolescence risk is low"*
- ***High Risk:*** *"The firm's technology obsolescence risk is high"*

Pair 8:

- ***Low Risk:*** *"The company faces negligible risk of technology obsolescence"*
- ***High Risk:*** *"The company faces considerable risk of technology obsolescence"*

Pair 9:

- ***Low Risk:*** *"The technology obsolescence risk for the firm is minimal"*
- ***High Risk:*** *"The technology obsolescence risk for the firm is substantial"*

Pair 10:

- ***Low Risk:*** *"The firm is at a low risk of technological obsolescence"*
- ***High Risk:*** *"The firm is at a high risk of technological obsolescence"*

C Example: 10-K technology summary

Technology Mentions in Apple Inc.'s 10-K Filing

1. Business Section

Technologies Identified: ***Macintosh Computers***: *Personal computing devices designed for various markets.* ***iPhone, iPad, iPod, Apple TV***: *Mobile communication and media devices.* ***Mac OS X and iOS Operating Systems***: *Proprietary operating systems for Mac and mobile devices.* ***iTunes Store***: *Digital content distribution platform including App Store and iBookstore.* ***Xserve***: *Server hardware for enterprise solutions.* ***iLife and iWork Software Suites***: *Consumer-oriented software applications for media management and productivity.*

Function/Use Case: *The technologies are used for personal computing, mobile communication, media consumption, and digital content distribution. The operating systems provide the foundation for the hardware, while software applications enhance user experience.*

Strategic Importance: *These technologies are central to Apple's strategy of providing integrated solutions that enhance user experience and drive customer loyalty.*

***Partnerships**: Collaborations with third-party developers for applications and content on the iTunes Store.*

***Risks**: Competitive pressures in the technology market, particularly from Windows-based systems and other mobile device manufacturers.*

2. Risk Factors Section

***Technologies Identified**: **Microprocessors, NAND Flash Memory, DRAM, LCDs**: Key components sourced from single or limited suppliers.*

***Challenges/Risks**: Supply chain risks due to reliance on single-source suppliers for critical components, which could lead to production delays or increased costs.*

***Strategic Importance**: The ability to secure these components at favorable prices is crucial for maintaining competitive pricing and product availability.*

3. Management's Discussion and Analysis (MD&A)

***Technologies Identified**: **Research and Development (R&D)**: Investment in new technologies and product enhancements.*

***Function/Use Case**: R&D is focused on developing innovative products and technologies to maintain competitive advantage.*

***Strategic Importance**: Continuous investment in R&D is deemed critical for future growth and market leadership.*

***Risks**: The fast-paced nature of technological change requires ongoing innovation, which can be resource-intensive.*

4. Notes to Financial Statements

***Technologies Identified**: **Patents and Intellectual Property**: Extensive portfolio of patents related to hardware and software.*

***Strategic Importance**: Protecting intellectual property is vital for maintaining competitive advantage and preventing infringement.*

***Risks**: Potential for litigation related to patent infringement, which could result in significant financial liabilities.*

5. Exhibits

***Technologies Identified**: **AppleCare**: Support and warranty services for hardware products.*

***Function/Use Case**: Provides customers with extended service options and support.*

Strategic Importance: Enhances customer satisfaction and loyalty through reliable support services.

6. R&D Section

Technologies Identified: **Investment in New Technologies**: Focus on developing new products and enhancing existing ones.

Function/Use Case: R&D efforts are aimed at expanding product offerings and improving technology integration.

Strategic Importance: Essential for sustaining innovation and meeting evolving consumer demands.

Risks: High costs associated with R&D and the risk of not achieving desired outcomes.

Summary

Apple Inc. leverages a diverse array of technologies, including proprietary hardware, operating systems, and software applications, to create a seamless and integrated user experience. The strategic significance of these technologies lies in their ability to enhance customer satisfaction, drive innovation, and maintain competitive advantage in a rapidly evolving market. However, the company faces several risks, including supply chain vulnerabilities, competitive pressures, and potential legal challenges related to intellectual property. Continuous investment in R&D is critical for Apple to sustain its market leadership and adapt to technological advancements.

Recurring Themes

Focus on Integration: Apple emphasizes the seamless integration of hardware and software to enhance user experience. **R&D Investment**: A strong commitment to R&D is evident, highlighting the importance of innovation in maintaining competitive advantage. **Supply Chain Risks**: Reliance on specific suppliers for key components poses significant risks to operational stability. **Intellectual Property Protection**: The importance of patents and trademarks in safeguarding technological innovations is a recurring theme.

D Example: Innovation summary

1. General Themes and Common Topics: *The patents issued to the firm primarily focus on enhancing data management and processing systems. The overarching themes include:*

- ***Data Reconstruction and Storage Efficiency****: The first patent addresses the challenge of reconstructing corrupted data from storage units without performance degradation, emphasizing efficient data retrieval and redundancy.*
- ***Digital Media Processing and Editing****: The second patent revolves around improving digital video delivery systems, particularly in parsing, seeking, and editing video content, which enhances user experience in media consumption.*
- ***Database Query Optimization****: The third patent introduces a method for optimizing database queries, focusing on efficient record retrieval and logical evaluation, which is crucial for improving database performance.*
- ***Content Analysis and Knowledge Representation****: The fourth patent deals with content processing systems that analyze discourse, identify themes, and manage knowledge representation, highlighting advancements in natural language processing and knowledge management.*

2. Individual Patent Details:

- ***Patent 1****:*
  - ***Targeted Problem****: Corrupted data reconstruction in storage systems.*
  - ***Proposed Solution****: A method that allows for real-time reconstruction of corrupted data without performance degradation by using a parity block system.*
  - ***Notable Technical Features****: Data is read in stripes, and the reconstruction is achieved through XOR operations on received data blocks, ensuring consistent data delivery rates.*
- ***Patent 2****:*
  - ***Targeted Problem****: Inefficiencies in digital video delivery and editing.*
  - ***Proposed Solution****: A system that parses digital representations of audio-visual works to create tag files for efficient seeking and editing.*

- **Notable Technical Features**: *The system allows for fast and slow forward/rewind operations based on tag file information, and a video editor that generates new video files from existing ones based on specified parameters.*

- ***Patent 3***:
  - ***Targeted Problem***: *Inefficient query processing in database systems.*
  - ***Proposed Solution***: *A computerized query optimizer that converts queries into optimized Boolean trees for efficient record evaluation.*
  - ***Notable Technical Features***: *The use of disjoint semi-open ranges for record selection and the ability to skip over intervals of records based on logical conditions.*

- ***Patent 4***:
  - ***Targeted Problem***: *Understanding and processing input discourse for knowledge representation.*
  - ***Proposed Solution***: *A content processing system that utilizes a lexicon and a knowledge catalog to analyze discourse and identify key themes.*
  - ***Notable Technical Features***: *The system includes a theme parsing system, a knowledge catalog processor, and a content indexing processor that dynamically expands ontologies based on discourse analysis.*

E More details about the model

E.1 Microfondation for $p_C(\ell) = \lambda(\text{IDE})R_C$,

In the main text, the probability of successful "leapfrogging" by an outside competitor is modeled as

$$p_C(\ell) = \lambda(\text{IDE})R_C,$$

where IDE is a single scalar meant to capture *how well competitor patents target* the focal firm's specific technology line ℓ. This section clarifies why such a term naturally appears and how one can *microfound* it beyond a purely "reduced form" approach.

Consider a single product line ℓ whose current quality rung is $q(\ell)$. Suppose:

- A competitor j invests R_C in R&D *aimed at* improving or displacing the focal firm's technology on line ℓ.

- The *effectiveness* of R_C depends on the *technological overlap* between competitor j's existing patent portfolio (or knowledge base) and the focal firm's line ℓ. Denote this overlap by $\Omega(\ell, j)$, where larger $\Omega(\ell, j)$ means the competitor is more likely to develop a quality-improving innovation specifically relevant to line ℓ.

Proof of proposition 2

Proposition 2. ***(Higher IDE Lowers Expected Future Profit Growth)*** *Suppose* IDE *increases while other parameters* $(\beta, \eta, \lambda, \Gamma, \dots)$ *remain fixed. Then, under the equilibrium policy* $\{R_I^*\}$ *that solves the focal firm's value function, the focal firm's expected profit over any finite horizon* T *strictly decreases. Equivalently, higher* IDE *reduces the growth path of total firm profit in expectation.*

Proof. Consider a single line with current quality rung q. Let $V_I(q; \text{IDE})$ denote the focal firm's value for this line when the external displacement probability is $p_C = \lambda(\text{IDE})R_C$. For each period, the flow of profit $\pi(q)$ remains the same, but the chance of losing the line to an outside competitor each period is strictly increasing in IDE. By standard arguments for dynamic programming under uncertainty, a higher probability of losing the asset sooner reduces the expected duration over which the focal firm can collect $\pi(q)$ (or any future upgraded profit $\pi(q+1)$). Hence $V_I(q; \text{IDE}_1) > V_I(q; \text{IDE}_2)$ whenever $\text{IDE}_1 < \text{IDE}_2$. This dominance holds for all q.

Next, suppose the firm holds a collection $\mathcal{L}_f$ of lines. The total focal firm value is

$$\mathcal{V}_f(\text{IDE}) = \int_{\ell \in \mathcal{L}_f} V_I\big(q(\ell); \text{IDE}\big) d\ell.$$

If IDE rises, each $V_I\big(q(\ell); \text{IDE}\big)$ decreases strictly. Thus $\mathcal{V}_f(\text{IDE})$ decreases. Because the firm's flow of profits in future periods derive from these lines, the *expected* sum of future profits also declines in IDE. Over any finite horizon T, the present value of total profits declines strictly as IDE increases, which means the firm's expected profit growth path is lower. This completes the proof. □

Overlap and Displacement Probability

If the competitor invests R_C, we posit that the *instantaneous* probability of successfully leapfrogging the focal firm on line ℓ is

$$p_C(\ell, j) = \lambda \Omega(\ell, j) R_C, \quad \lambda > 0.$$

This expression states that the competitor's R&D has an incremental chance of success proportional to how effectively competitor j's knowledge base is *targeted* to line ℓ. A natural interpretation is that each unit of R_C is more potent (i.e., yields a higher chance of quality improvement) when $\Omega(\ell, j)$ is large.

Definition of $\Omega(\ell, j)$. One can define $\Omega(\ell, j)$ by comparing the *textual* (or technological) similarity of competitor j's latest patents to the *technology descriptor* of line ℓ. In a data-driven model, we might write

$$\Omega(\ell, j) = \text{Similarity}\big(\text{PatentsOf}(j), \text{TechDescriptor}(\ell)\big),$$

where Similarity is computed via embeddings or other NLP-based measures of how competitor j's newly patented innovations overlap the focal firm's line.

From $\Omega(\ell, j)$ to a Firm-Level IDE

In practice, a focal focal firm firm f might face multiple competitor firms $j = 1, 2, \ldots, J$ over the same line ℓ. One can define the *composite* outside displacement hazard as

$$p_C(\ell) = \sum_{j=1}^{J} \lambda \Omega(\ell, j) R_{C,j},$$

where $R_{C,j}$ is competitor j's R&D on line ℓ. In some cases, the $\sum_j \Omega(\ell, j)$ term is simplified into a single measure of *aggregate* threat,

$$\text{IDE}_\ell = \sum_{j=1}^{J} \Omega(\ell, j) \left(\frac{R_{C,j}}{R_C^{\text{tot}}} \right), \quad R_C^{\text{tot}} = \sum_{j=1}^{J} R_{C,j}.$$

Then $p_C(\ell) = \lambda \text{IDE}_\ell \times R_C^{\text{tot}}$. This matches the earlier reduced-form expression where IDE_ℓ is a scalar summarizing the *relevance* of external innovators' patent portfolios to line ℓ.

Impact on Firm's Expected Profit

Once IDE_ℓ is defined, the probability of losing line ℓ in any given period is $\lambda(\text{IDE}_\ell) R_C^{\text{tot}}$. A *larger* IDE_ℓ implies that, for the same competitor R&D investment R_C^{tot}, the *effective* success rate in displacing the focal firm is higher. Consequently, the focal firm's expected duration of leadership on line ℓ is *shorter*, lowering the net present value it can extract from rung $q(\ell)$.

Therefore, although in the main text the relationship $p_C(\ell) = \lambda(\mathrm{IDE})R_C$ might appear reduced-form, it can be justified by explicitly modeling the *technological overlap* $\Omega(\ell, j)$ between the competitor's patent portfolio and the focal firm's specific line. Aggregating over multiple competitors yields an overall *IDE* measure. Hence, "IDE" is not merely a scaled hazard but a microfounded reflection of the competitor's *targeted R&D effectiveness*, itself increasing with textual or technological *similarity* of new patents to the focal firm's existing technology.

Proof. We consider a monopolistically competitive market in partial equilibrium. Each product line ℓ is produced with a *quality index* $q \in \{0, 1, 2, \dots\}$. The marginal cost of production at quality rung q is

$$c(q) = \frac{c_0}{\Gamma^q}, \quad \Gamma > 1,$$

for some constant $c_0 > 0$. The firm sets a constant markup factor

$$p(q) = \mu c(q), \quad \mu = \frac{\varepsilon}{\varepsilon - 1}, \quad \varepsilon > 1,$$

due to the constant elasticity of substitution ε. The price then becomes

$$p(q) = \mu \times \frac{c_0}{\Gamma^q}.$$

The (per-period) *flow profit* is the product of the per-unit markup $\left[p(q) - c(q)\right]$ and the quantity $x(q)$ demanded.

Derivation of $\pi(q)$

Step 1: Express $\left[p(q) - c(q)\right]$

$$p(q) - c(q) = \mu \frac{c_0}{\Gamma^q} - \frac{c_0}{\Gamma^q} = (\mu - 1)\frac{c_0}{\Gamma^q}.$$

Since $\mu = \varepsilon/(\varepsilon - 1)$, we have

$$\mu - 1 = \frac{\varepsilon}{\varepsilon - 1} - 1 = \frac{1}{\varepsilon - 1}.$$

Thus

$$p(q) - c(q) = \frac{c_0}{(\varepsilon - 1)\Gamma^q}.$$

Step 2: Express the Demand $x(q)$

Assume a CES aggregator for final consumption with elasticity ε. The representative consumer's demand function for this variety (in partial equilibrium) is

$$x(q) = \left(\frac{p(q)}{P}\right)^{-\varepsilon} Y,$$

where P is the CES price index, and Y is total expenditure on the differentiated goods. Substituting $p(q) = \mu c_0/\Gamma^q$ into the demand function yields

$$x(q) = \left(\frac{\mu c_0/\Gamma^q}{P}\right)^{-\varepsilon} Y = \left(\frac{P}{\mu c_0}\right)^{\varepsilon} \Gamma^{\varepsilon q} Y.$$

Define

$$K \equiv \left(\frac{P}{\mu c_0}\right)^{\varepsilon} Y,$$

which is a positive constant (holding P, μ, c_0, Y fixed). Then

$$x(q) = K\Gamma^{\varepsilon q}.$$

Step 3: Flow Profit $\pi(q)$

By definition,

$$\pi(q) = \Big[p(q) - c(q)\Big] x(q).$$

Using the results above:

$$\pi(q) = \left(\frac{c_0}{(\varepsilon - 1)\Gamma^q}\right) \left(K\Gamma^{\varepsilon q}\right) = \frac{c_0 K}{\varepsilon - 1}\Gamma^{\varepsilon q - q} = \frac{c_0 K}{\varepsilon - 1}\Gamma^{(\varepsilon - 1)q}.$$

Let

$$\kappa \equiv \frac{c_0 K}{\varepsilon - 1},$$

which is constant with respect to q. Hence we obtain the functional form

$$\pi(q) = \kappa\Gamma^{(\varepsilon-1)q}.$$

This completes the derivation.

F Industry-level Kogan et al. (2017) measure versus IDE

In this Appendix section, we compare the amount of displacement information captured by the industry-level KPSS measure and our new IDE measure. Since the industry-level KPSS values are proportional to the aggregated innovation value of the other firms in the same industry, they do not vary much within each industry code. Hence, these are almost industry-level measures of displacement. Therefore, we repeat the main regressions defined by equation 2 but only standardize the independent variables at the year level and drop the year cross industry category fixed effects.

As shown in table F1, the firm-level KPSS measure is positively associated with profit growth in all of the next 7 years. In particular, each standard deviation increase in the firm-level KPSS corresponds to a 2.1% increase in profit growth by year 1 and 7.3% by year 7. In addition, the industry-level KPSS measure is also significantly associated with profit growth. By year 1, each standard deviation increase in the industry-level KPSS for each firm corresponds to a 0.8% decrease in profit growth, and by year seven, the association is a negative 4.4%. Next, we restrict our sample to 2005 to 2015 of which we can compute our technology-based IDE measure and redo the analysis while adding IDE as another independent variable. As shown by table F2, when adding our IDE measure to the regression, the size of the association between the IDE and profit growth is similar to our main analysis: each standard deviation increase in IDE is associated with a 1.0% decrease in profit growth by the next year and 5.3% by year 7. On the other hand, the association with respect to the industry-level KPSS measure becomes insignificant and economically small.

	Dependent variable:						
	$\Pi_{t+1} - \Pi_t$	$\Pi_{t+2} - \Pi_t$	$\Pi_{t+3} - \Pi_t$	$\Pi_{t+4} - \Pi_t$	$\Pi_{t+5} - \Pi_t$	$\Pi_{t+6} - \Pi_t$	$\Pi_{t+7} - \Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Industry-level KPSS (other firms)	-0.008**	-0.018***	-0.021***	-0.026***	-0.034***	-0.039***	-0.044***
	(0.004)	(0.006)	(0.007)	(0.009)	(0.010)	(0.012)	(0.014)
Firm-level KPSS (focal firm)	0.021***	0.035***	0.045***	0.053***	0.061***	0.068***	0.073***
	(0.002)	(0.003)	(0.004)	(0.005)	(0.005)	(0.005)	(0.006)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Observations	145,853	131,840	119,412	108,286	98,368	89,527	81,577
R^2	0.075	0.086	0.093	0.096	0.102	0.108	0.112
Adjusted R^2	0.073	0.084	0.090	0.093	0.099	0.105	0.109
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE F1: This table shows the association between firm- and industry-level KPSS and profit growth. We control for the log asset growth and profitability among other variables – log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 1976 to 2022. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

G Application in asset pricing

In this section, we rank firms by their IDE values to construct an equal-weight long-short portfolio. First, we show that an equal-weight portfolio that goes long on the top 10% of firms with the highest IDE values and shorts the bottom 10% can achieve significant α in the next year. In addition, we examine the extent to which the correlation between the returns of the long and short sides of the portfolio is driven by intertemporal variations in risk factors, such as Fama-French factors and momentum.

G.1 IDE-based long-short portfolio

First, we analyze the equal-weight long-short portfolio based on the IDE measure. Since the IDE measures are calculated annually, we create a portfolio at the end of each year after computing this year's IDE values. We long the top 10% of firms with the highest IDE values and short the bottom 10% (equal-weight). This portfolio is rebalanced monthly, but the firms on the long and short sides are selected annually.

Insert tables *G*1 and *G*2 here.

As shown in tables G1 and G2, the sample of firms on the long side of the portfolios (in the sorting year) have significantly lower profit growth both over the next year and 7 years (a sanity check for our main result), larger market capitalization, lower asset growth, higher profitability, higher employment, higher capital stock, and more profit. Overall, we see that firms that are more exposed to displacements of their asset in place – the ones longed in our portfolio – are generally larger firms that are more profitable.

Figure G1 shows the P&L of this portfolio and the long and short sides separately. We observe that expectedly both the long and short sides had significant drops from 2008 to 2009 (the Global Financial Crisis). However, the overall long-short portfolio was profitable during this time period because the short side lost more than the long side. This shows that the IDE-based portfolio can gain profits during recessions. This demonstrates that firms with low IDE have very low returns during recessions. Economically, a firm whose main technologies are not close to major innovators devalues quickly when the overall economy is weak.

Overall, as shown by figure G2. The short side of the IDE portfolio has an annualized return of $\frac{0.49}{11} = 4.5\%$ below the S&P 500 index. Similarly, the annualized return of the long side is $\frac{0.50}{11} = 4.5\%$ above. This shows that roughly an equal amount of profit is made from the long and short sides of the portfolio.

	Dependent variable:						
	$\Pi_{t+1} - \Pi_t$	$\Pi_{t+2} - \Pi_t$	$\Pi_{t+3} - \Pi_t$	$\Pi_{t+4} - \Pi_t$	$\Pi_{t+5} - \Pi_t$	$\Pi_{t+6} - \Pi_t$	$\Pi_{t+7} - \Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	-0.010**	-0.019**	-0.026**	-0.038***	-0.046***	-0.048***	-0.053***
	(0.004)	(0.007)	(0.010)	(0.009)	(0.010)	(0.011)	(0.012)
Industry-level KPSS (other firms)	0.003	0.003	0.012	0.013	0.025	0.020	0.001
	(0.009)	(0.013)	(0.015)	(0.020)	(0.022)	(0.022)	(0.024)
Firm-level KPSS (focal firm)	0.024***	0.037***	0.046***	0.054***	0.064***	0.079***	0.088***
	(0.004)	(0.006)	(0.007)	(0.012)	(0.013)	(0.012)	(0.013)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Observations	25,958	24,044	22,426	20,985	19,690	18,528	17,434
R^2	0.082	0.093	0.097	0.100	0.110	0.121	0.132
Adjusted R^2	0.073	0.084	0.087	0.090	0.100	0.110	0.121
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE F2: This table compares the association between IDE and profit growth versus industry-level KPSS and profit growth. We control for the log asset growth and profitability among other variables – log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

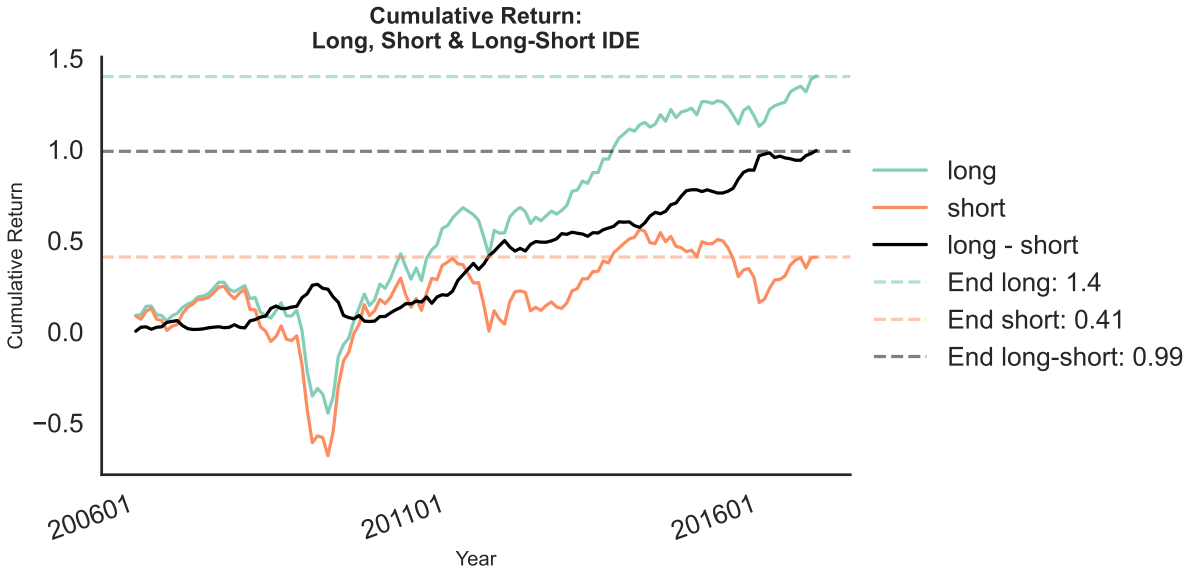


Figure G1: This figure shows the cumulative (monthly) return of the long and short sides and the overall long-short IDE portfolio. The long-short portfolio is equal-weight and rebalanced monthly while keeping the selection of firms constant each year. The long-side is consisted of firms in each industry category and year with the highest 10% of IDE, and the short-side is consisted of the lowest 10%.

Next, we compare the IDE portfolio return with well-studied factor models such as CAPM, Fama-French 3, Fama-French 5, and Fama-French 5 with momentum. We compare the monthly portfolio returns against these models.[21] We compute the α of the IDE portfolio using the following regression

$$r_t - rf_t = \alpha + \beta \cdot X_t + \epsilon_t, \tag{13}$$

where r_t is the return of the IDE portfolio in month t, rf_t is the monthly risk-free rate, and X_t are the factors used in each model in month t.

Insert table G3 here.

As shown in Table G3, the monthly α values vary across different models: CAPM yields a monthly α of 0.74% (annualized to 8.92%), FF3 yields 0.76% (annualized to 9.13%), FF5 yields 0.53% (annualized to 6.41%), and FF5 with momentum yields 0.53% (annualized to 6.38%). This shows that the IDE-based long-short portfolio generates positively signif-

[21]If we use annual factor data, we only have 11 data points in our regression. Therefore, we analyze monthly data to have a meaningful number of observations.

icant returns that cannot be explained by these asset pricing models.

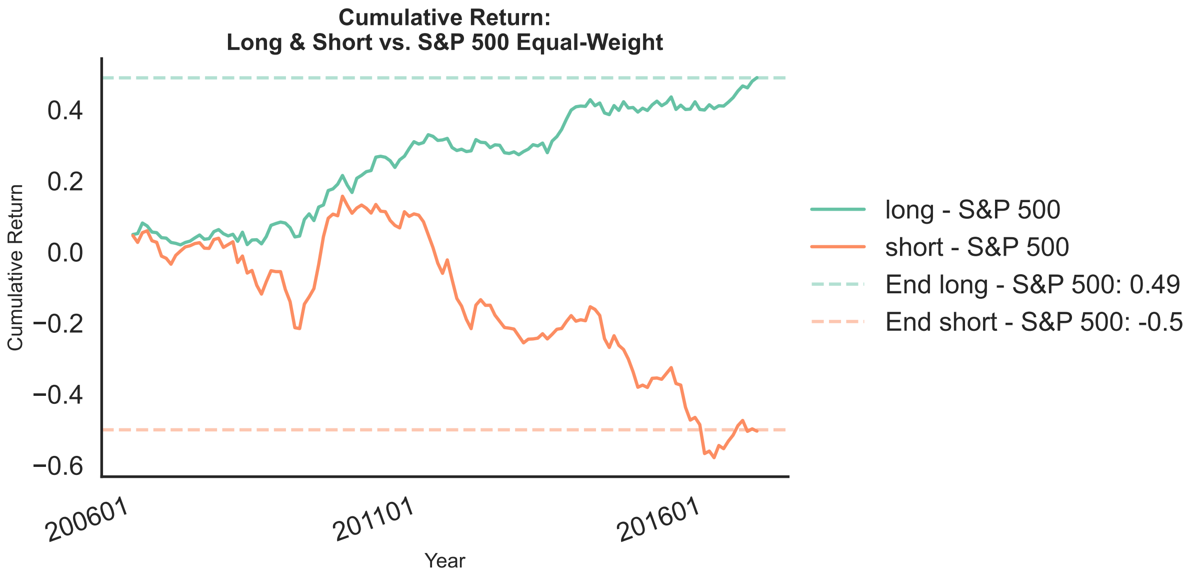


Figure G2: This figure shows the cumulative (monthly) return of the long and short sides of the IDE equal-weight portfolio relative to the equal-weight S&P500 index (without dividends). The long-short portfolio is equal-weight and rebalanced monthly while keeping the selection of firms constant each year. The long-side is consisted of firms in each industry category and year with the highest 10% of IDE, and the short-side is consisted of the lowest 10%.

G.2 Understanding the correlation between the long-side and short-side

Figure G1 shows a strong correlation between the long and short returns. In this subsection, we decompose this high correlation using traditional risk factors. To compute the amount of correlation due to the inter-temporal variation of these factors, we first establish a baseline correlation between raw returns in the long and short sides at 0.95. Then, for each set of factors, we compute their contribution to the baseline correlation in the following way:

$$\Delta\rho = 0.95 - \text{corr}(r^{\text{adj}}_{\text{long}}, r^{\text{adj}}_{\text{short}}), \tag{14}$$

where $r^{\text{adj}}_{\text{long}}$ and $r^{\text{adj}}_{\text{short}}$ are the residual of the IDE portfolio returns after regressing on the set of factors similar to equation 13.

We compute $\Delta\rho$ for CAPM, FF3, FF5, and FF5 with momentum. CAPM (the expected market return factor) contributes to roughly 0.20 in correlation. FF3 (small minus big and value minus growth) explains another 0.10 in correlation. FF5 (robust minus weak and conservative minus aggressive) does not further reduce the correlation. Lastly, momentum contributes to another 0.10 in correlation.

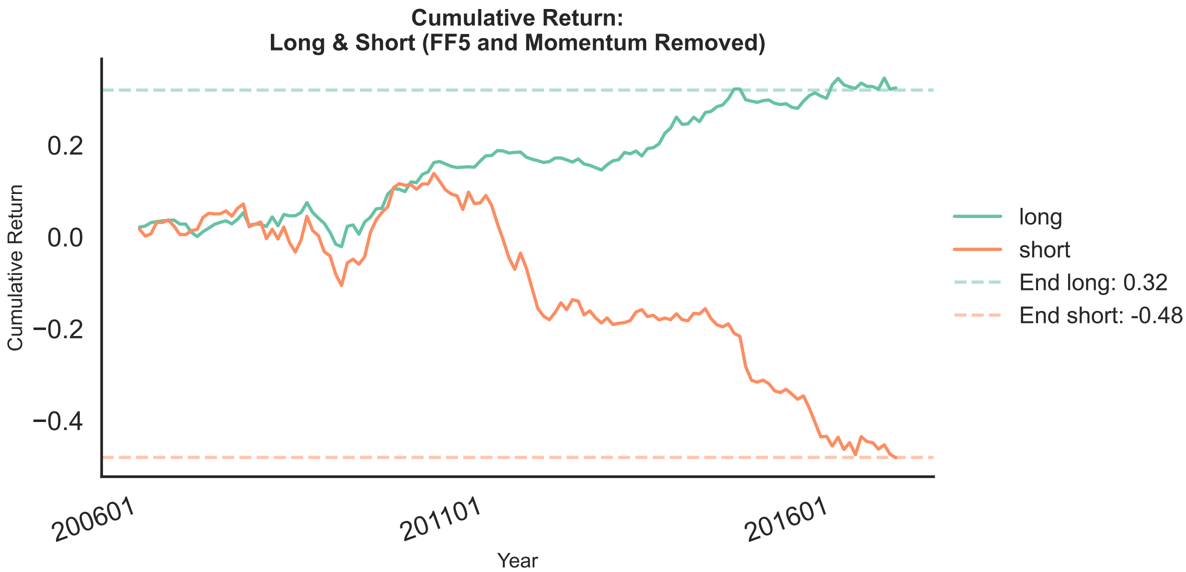


Figure G3: This figure shows the cumulative (monthly) return of the long and short sides of the IDE portfolio when Fama-French 5 factors and momentum are removed. The long-short portfolio is equal-weight and rebalanced monthly while keeping the selection of firms constant each year. The long-side is consisted of firms in each industry category and year with the highest 10% of IDE, and the short-side is consisted of the lowest 10%.

Overall, the biggest decrease in correlation happens when including the market return, small minus big, value minus growth, and momentum also contribute to large decreases in correlation. After all five Fama French factors and momentum are removed from the long and short sides, the correlation drops from 0.95 to 0.57. Figure G3 shows the P&L of the long and short sides of the IDE-based portfolio after Fama-French 5 and momentum are removed from the returns.

Variable	N	Mean	SD	SE	Min	P25	Median	P75	Max
one-year profit growth	2883	0.017	0.356	0.007	-4.828	-0.080	0.026	0.120	4.643
seven-year profit growth	2031	0.154	0.676	0.015	-4.711	-0.147	0.192	0.515	6.232
market capitalization	3069	7917491.833	30270291.674	546409.399	700.615	183685.320	761674.920	3178831.000	442093617
log of asset growth	3083	0.050	0.247	0.004	-1.918	-0.039	0.041	0.125	3.596
profitability	3083	0.393	0.235	0.004	0.000	0.255	0.351	0.479	3.430
log of employment	3047	0.873	2.001	0.036	-6.215	-0.639	1.030	2.198	6.096
log of capital stock	3077	0.947	2.344	0.042	-10.845	-0.752	0.992	2.516	8.391
log of profit	3083	5.487	2.015	0.036	-4.705	4.133	5.474	6.820	11.324

TABLE G1: This table shows the summary statistics of firms on the long-side of the portfolio. The long-short portfolio is equal-weight and rebalanced monthly while keeping the selection of firms constant each year. The long-side is consisted of firms in each industry category and year with the highest 10% of IDE, and the short-side is consisted of the lowest 10%.

Variable	N	Mean	SD	SE	Min	P25	Median	P75	Max
one-year profit growth	2787	0.056	0.526	0.010	-5.720	-0.080	0.066	0.224	4.605000e+00
seven-year profit growth	1542	0.354	1.005	0.026	-4.307	-0.162	0.324	0.848	5.495000e+00
market cap	3061	2523784.381	18767483.333	339214.473	1189.552	83561.340	346063.900	1223399.120	6.431201e+08
log of asset growth	3083	0.128	0.369	0.007	-4.549	-0.041	0.071	0.243	3.360000e+00
profitability	3083	0.363	0.347	0.006	0.000	0.175	0.292	0.482	1.294400e+01
log of employment	2965	-0.069	1.980	0.036	-6.215	-1.431	-0.152	1.281	5.886000e+00
log of capital stock	3046	-0.063	2.314	0.042	-7.703	-1.737	-0.132	1.663	6.053000e+00
log of profit	3083	4.529	1.937	0.035	-3.063	3.295	4.568	5.726	1.117200e+01

TABLE G2: This table shows the summary statistics of firms on the short-side of the portfolio. The long-short portfolio is equal-weight and rebalanced monthly while keeping the selection of firms constant each year. The long-side is consisted of firms in each industry category and year with the highest 10% of IDE, and the short-side is consisted of the lowest 10%.

H Nonlinear IDE

Our definition of IDE in equation 1 assumes that the magnitude of exposure scales linearly with the cosine similarity between firm j's innovations and firm i's technologies. In this section, we relax this assumption and develop a measure where the cosine similarity enters the function nonlinearly. While potentially less economically interpretable, this measure is more strongly predictive of profit growth. To construct this measure, we evaluate over 20 machine learning models and select the one with the best performance on testing data. These models use modified IDE values with varying degrees of nonlinearity as inputs, and predict the next year's profit growth as targets.[22] The resulting predictions form our new IDE measure, which scales nonlinearly with respect to the similarities.

We randomly split 60% of the data for training and validation, and the other 40% is used for testing. As shown by table H1, the neural network implementation using FastAI achieves the highest testing performance in terms of mean-squared-error. Therefore, we use the predictions of this model as our nonlinear IDE measure.[23]

We evaluate the informativeness of this measure using the 40% testing data. More specifically, we run the regression in our main study specified by equation 2 where the negative of the nonlinear IDEs (model predictions) are used to replace the original IDEs in the regression. We use the negative of the predictions because these model predictions are meant to be positively correlated with profit growth; however, our IDE measure is intended to have a negative association with profit growth. As a sanity check, we observe that our main IDE measure is significantly negatively associated with the model predictions. Comparing the results with the new nonlinear measure shown in table H2 and the original main result in table A2, we observe that the nonlinear measure is significantly negatively associated with profit growth similar to our main measure: each standard deviation increase in the nonlinear IDE measure is associated with a 0.4% decrease in profit growth in year 1 and 4.0% by year 7.

[22] These modified IDEs mostly follow the definition in equation 1 except the cosine similarity term is replaced with cos, $\cos^2$, $\cos^3$, $\cos^{\frac{1}{3}}$, and $\cos^{\frac{1}{5}}$.

[23] Similar to section 7, we compute the importance of each input in this best model. We see that all versions of IDE (cosine, squared cosine, cubic cosine, cubic root of cosine, fifth root of cosine) are significantly important, and cubic of cosine is the most important one.

Variable	Estimate	s.e.	t value	Pr(>|t|)
CAPM	0.7437	0.185	4.017	< 0.001 ***
Fama-French 3	0.7593	0.185	4.111	< 0.001 ***
Fama-French 5	0.5251	0.179	2.926	0.003 ***
Fama-French 5 and momentum	0.5232	0.172	3.041	0.002 ***
number of observations		132		

TABLE G3: This table shows the α of the IDE long-short portfolio after controlling for traditional asset pricing factors. The return data is from 2006 to 2016. The long-short portfolio is equal-weight and rebalanced monthly while keeping the selection of firms constant each year. The long-side is consisted of firms in each industry category and year with the highest 10% of IDE, and the short-side is consisted of the lowest 10%.

Model	Score Test	Score Val
WeightedEnsemble_L3	-0.153465	-0.153842
LightGBMXT_BAG_L1	-0.153483	-0.154181
WeightedEnsemble_L2	-0.153485	-0.153996
NeuralNetTorch_BAG_L1	-0.153499	-0.154048
NeuralNetFastAI_BAG_L1	-0.153507	-0.154048
XGBoost_BAG_L2	-0.153508	-0.154341
NeuralNetTorch_BAG_L2	-0.153538	-0.154047
LightGBM_BAG_L1	-0.153541	-0.154328
CatBoost_BAG_L2	-0.153556	-0.154222
CatBoost_BAG_L1	-0.153569	-0.154196
LightGBMXT_BAG_L2	-0.153575	-0.154208
LightGBMLarge_BAG_L2	-0.153578	-0.154362
LightGBMLarge_BAG_L1	-0.153581	-0.154379
XGBoost_BAG_L1	-0.153581	-0.154357
LightGBM_BAG_L2	-0.153594	-0.154289
NeuralNetFastAI_BAG_L2	-0.153751	-0.154610
RandomForestMSE_BAG_L2	-0.156167	-0.160802
ExtraTreesMSE_BAG_L2	-0.156188	-0.164847
RandomForestMSE_BAG_L1	-0.175173	-0.176518
ExtraTreesMSE_BAG_L1	-0.181487	-0.183492
KNeighborsUnif_BAG_L1	-0.185870	-0.185160
KNeighborsDist_BAG_L1	-0.192314	-0.192117

TABLE H1: Performance of machine learning models using nonlinear similarities to construct an alternative IDE measure that predicts next year's profit growth. The performance metric is the negative MSE. The input variables are IDE, the square of IDE, the cube of IDE, the cubic root of IDE, and the fifth root of IDE. The target value is next year's profit growth.

I More details on constructing the LLM-based IDE measure

To measure a firm's innovation displacement exposure, we construct a dataset of firm technologies and innovations spanning each year from 1980 to 2015.[24] We define a firm's innovations in year t by the patents it received from $t-4$ to t, and we define its current technologies as those employed by the firm over the past five years in its operations and research. Specifically, we first obtain yearly technology summaries using GPT4o and then aggregate the summaries from $t-4$ to t to build a comprehensive technology stack for the firm.

I.1 Constructing representations of firm innovations

First, we generate a numerical representation of each firm's innovations in each year with a 5-year lookback window. We start by producing an innovations summary for each firm each year based on the patents issued to it in year t. The primary challenge arises from the fact that some firms receive over a hundred patents annually, making it computationally impractical to generate a numerical vector that captures all the information from each patent. To address this issue, we employ a procedure designed to select a representative sample of patents, while accounting for the variability in innovation capacity across different firms: i) If a firm i has at most 10 patents in year t, we take the abstracts of all of its patents issued in t; 2) If a firm i has more than 10 patents but at most 100, we take a random subset of 10 patents issued in t; 3) If a firm i has more than 100 patents, we take a random 10% subset of all of the patents issued in t.

After selecting the set of patents, we use GPT4o to summarize firm i's patents in year t in at most 500 tokens following the prompt:

> *You are an economist studying firms' innovations. Summarize the following patent abstracts. First, give a summary of the common topics covered by many abstracts, then focus on individual patents.*
>
> *{Selected patent texts.}*

Next, we generate innovation summaries for firm i in each year $t \in \{t-4, \cdots, t-1\}$ and concatenate the summaries in these 5 years chronologically from the earliest to the

[24] Our patent data begins in the year 1976; however, our displacement measure in each year is constructed with a look-back window of 5 years. Therefore, the earliest year we can compute displacement measures for is 1980.

	Dependent variable:						
	$\Pi_{t+1} - \Pi_t$	$\Pi_{t+2} - \Pi_t$	$\Pi_{t+3} - \Pi_t$	$\Pi_{t+4} - \Pi_t$	$\Pi_{t+5} - \Pi_t$	$\Pi_{t+6} - \Pi_t$	$\Pi_{t+7} - \Pi_t$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IDE	-0.004	-0.007	-0.016	-0.026*	-0.038**	-0.040***	-0.040**
	(0.004)	(0.006)	(0.011)	(0.012)	(0.012)	(0.011)	(0.013)
profitability t	-0.015**	-0.040***	-0.048***	-0.047***	-0.060***	-0.056***	-0.057**
	(0.006)	(0.009)	(0.012)	(0.014)	(0.016)	(0.016)	(0.020)
log asset growth	0.067***	0.074***	0.083***	0.087***	0.082***	0.082***	0.092***
	(0.007)	(0.012)	(0.012)	(0.016)	(0.018)	(0.022)	(0.024)
Year	✓	✓	✓	✓	✓	✓	✓
Ind Code	✓	✓	✓	✓	✓	✓	✓
Year$\times$*IndCategory*	✓	✓	✓	✓	✓	✓	✓
Observations	10,117	9,383	8,737	8,194	7,683	7,226	6,795
R^2	0.150	0.163	0.159	0.152	0.154	0.168	0.170
Adjusted R^2	0.106	0.116	0.108	0.097	0.095	0.107	0.106
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE H2: This table shows the association between nonlinear IDE and profit growth. We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data a test set spans from 2005 to 2015 (a random 40% subset of the entire dataset). All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

latest. Finally, we extract a numerical embedding based on each concatenated innovation summary using a text embedding model.[25]

I.2 Constructing representations of firm technologies

The second key component of our measures involves creating firm-level numerical representations of technologies. First, we generate a summary for each firm annually. For a given firm i in year t, we apply the following prompt to obtain a summary of i's technologies in t:

> *Instructions: You are an economist analyzing companies' technologies. In three paragraphs, describe the technological stack of name of i in the year t. Focus on technologies that the company was using in its day-to-day operations and research. Be as specific as possible and give a lot of details. For each technology, mention whether or not it was considered legacy in the field in t and whether or not there were disruptive technologies that could replace it. If there was a risk of disruption, mention companies and technologies that were threatening to the given technology. Do not use information that became public after the year t.*

Similar to constructing innovation embeddings, we obtain a 500-token summary for firm i for each year from $t-4$ through t, concatenate these summaries, and then utilize the same text-embedding model to create the technology embedding for firm i in year t.

J Validity checks against the potential look-ahead bias of GPT4o

One of the components of our LLM-based displacement measure is a set of descriptions of the technologies used by each firm each year. These technologies are generated by GPT4o following a prompt that specifies the firm and year (the exact prompt is shown in Appendix section I). A potential concern of this approach is the look-ahead bias in the GPT4o-generated summaries. In this subsection, we take two approaches show GPT4o is aware of the temporal ordering of technologies.

[25]We use text-embedding-3-large to generate 3072-dimensional embeddings.

J.1 10-K-based ordering of technologies

We use the SEC 10-K filings as a data source of technology descriptions with a definitive time stamp. We first randomly select 500 firm-year pairs in our dataset. Then, we extract the 10-K filing for each firm in the corresponding year. Some firms are foreign issuers and do not file for 10-K. Therefore, after removing these firms, we have a set of 304 firm-year pairs. For each pair, we also download the 10-K filing of the firm from 3 years ago. Then, we pass each 10-K file to GPT4o and ask it to summarize the technologies based on each document:

> *This file is a 10-K form summarizing the performance of a firm in a year. Please read this website and summarize the technologies used by this firm this year in its day to day operations and research. Use no more than 500 tokens. Do not include any information about the year. Be as specific as possible and give a lot of details. For each technology, mention whether or not it was considered legacy in the field in t and whether or not there were disruptive technologies that could replace it. If there was a risk of disruption, mention companies and technologies that were threatening to the given technology: [Downloaded file]*

Then, each pair of summaries from the same firm 3 years apart is passed to GPT4o, and we ask GPT4o to determine which one is from an earlier year. The order of the two documents is randomized: half of the time, the summary of the earlier document appears first in the prompt, and the other times, the later one appears first. We use the prompt:

> *Given these two summaries above, tell me which one is from an earlier year: [newline] First summary: [first summary] [newline] Second summary: [second summary].*

Lastly, we compare the GPT4o ordering and the correct ordering.

As an example, take the firm Apple and the year 2010, we first download the 10-K filings of Apple in 2010 and 2007. Then, we ask GPT4o to create a short summary for each 10-K document focusing on the technologies described in the corresponding filing. Next, we pass these two summaries to GPT4o in a randomized order. For example, if we randomly pick the summary based on the 2010 filing as the first one in the prompt. Our query to GPT4o would be

> *Given these two summaries above, tell me which one is from an earlier year: [newline] First summary: [Apple 2010 technology summary] [newline] Second summary: [Apple 2007 technology summary].*

Then, we check if the response is that the second summary is from an earlier year.

We apply this process to the entire sample of firm-year pairs, and we observe that GPT4o gives the correct time ordering of the two technology summaries over 90% of the time. More specifically, the accuracy is balanced regardless of whether or not the earlier summary appears first in the prompt. When the earlier document appears first, GPT4o's accuracy is 91%, and when the later one appears first, GPT4o's accuracy is 92%. The two numbers are not statistically different.

J.2 GPT4o description-based ordering of technologies

Next, we test whether GPT4o can correctly order technologies based on the summaries we use to construct our displacement measure (discussed in Appendix section I).

We conduct this test at two levels of varying difficulty. First, we select technology summaries from the same firm 5 years apart. Then, to increase the difficulty, we choose summaries that are 3 years apart. In each test, we randomly select 500 firm-year pairs first and then find the corresponding earlier documents. We drop the ones that do not have a match (the firms were not included 5 or 3 years ago).

In particular, given two summaries, we use the following prompt:

> *You are an economist reading descriptions of technologies used by firm in daily operations and research. You will read two summaries and tell me which one is from an earlier year. Respond with 1 if the earlier one is the first summary and 0 if the earlier one is the second summary. Do not include any words or symbols, just use an integer:*
>
> *[newline] First summary: <summary 1>*
>
> *[newline] Second summary:<summary 2>*

Similar to the exercise we did with 10-K files, the order of the first and second summaries is randomized: half of the time, the earlier summaries appear first, and in the other cases, the later summaries appear first. Overall, we observe that when the 2 documents are 5 years apart, GPT4o correctly orders them 77% of the time. In particular, when the earlier summary appears first, GPT4o correctly identifies it as the earlier one 99% of the time. When the separation is only 3 years, GPT4o achieves a 67% accuracy and 95% when the earlier summary appears first. Note that in both the 5-year and 3-year tests, the technology summaries are similar, so this task is challenging even for humans.

These exercises show that GPT4o understands the time ordering of technologies at a high accuracy.

K Neural Networks and GPT

K.1 Introduction to Neural Networks

We begin by exploring a basic neural network: a linear neural network with a single hidden layer. This network is characterized by three dimensions: input dimension $\dim_{in}$, hidden dimension $\dim_h$, and output dimension $\dim_{out}$.

This one-hidden-layer network involves two mappings. The first mapping translates the input space into the hidden space. Formally, if the input data is denoted as X, then the hidden space H is given by

$$H = f_1(X) = XW_1 + B_1,$$

where W_1 and B_1 are trainable matrices of parameters.

The second mapping converts the hidden space H into the output space Y:

$$Y = f_2(H) = f_2(f_1(X)) = (XW_1 + B_1)W_2 + B_2,$$

$$= XW_1W_2 + B_1W_2 + B_2,$$

where W_2 and B_2 are also trainable matrices of parameters.

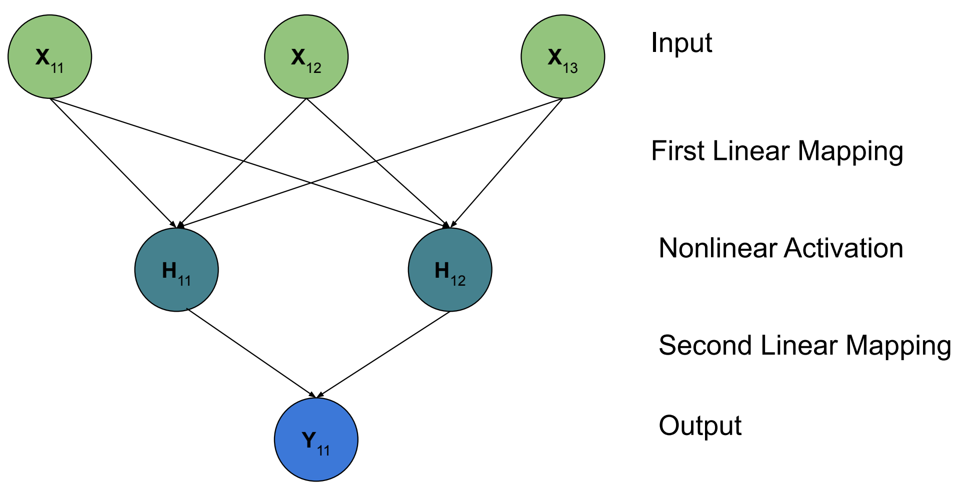


Figure K1: Diagram of a one-hidden-layer neural network with a nonlinear activation function. The top green circles represent the inputs, which are three-dimensional vectors in this figure. The turquoise circles in the middle represent the nodes in the hidden layer, and the blue circle at the bottom represents the output.

As illustrated in Figure K1, this simple linear network can be extended by introducing a nonlinear activation function $g(\cdot)$. When the activation function is applied to the first mapping, the output of the first mapping (the hidden space) becomes

$$H = g(f_1(X)) = g(XW_1 + B_1).$$

In a similar manner, applying a nonlinear activation function after any mapping adds nonlinearity to the linear transformation, enabling neural networks to more effectively model complex relationships between the input X and output Y.

K.2 Introduction to GPT-4

In this section, we delve into the key components of GPT-4. We begin by discussing the decoder architecture, which forms the foundation of models such as those in the GPT family. Following this, we explain the self-attention mechanism that empowers GPT-4 to consider the context when generating new text. We then explore the pre-training process for GPT-4's base model, and conclude by highlighting the advancements that GPT-4 (and 3.5) brings over its predecessors.

K.2.1 Decoder

The original Transformer architecture was conceived for tasks like machine translation, utilizing both an encoder and a decoder. In a typical transformer-based translation model, the encoder generates a numerical representation of the input up to token $t + 1$, where the $(t + 1)$th token is the next to be translated. The decoder then uses this encoder output along with a numerical representation of the t translated words to predict the translation of the $(t + 1)$th word. However, GPT employs a modified version of the Transformer architecture that includes only the decoder component. This design centers exclusively on generating a numerical representation of the already generated text and predicting the next token, which is ideal for tasks such as text completion and generation.

The decoder in GPT models comprises four essential components: positional encoding, self-attention layers, position-wise feedforward networks, and layer normalization with residual connections. We will briefly discuss each of these components before focusing on the self-attention layer, which is the core mechanism driving the model's capabilities.

In the Generative Pre-trained Transformer (GPT) model architecture, positional encoding is vital for providing information about the sequence position of tokens. The Transformer architecture does not inherently account for the order of tokens, so positional encoding ensures that the model can distinguish tokens based on their positions within a sequence.

Positional encoding involves adding fixed-length vectors to the input token embeddings before feeding them into the model. These positional embeddings encode positional information relative to other tokens in the sequence. GPT uses sinusoidal functions to generate these embeddings:

$$\begin{aligned} \mathrm{PE}_{(pos,2i)} &= \sin\left(\frac{pos}{10000^{2i/d_{\mathrm{model}}}}\right) \\ \mathrm{PE}_{(pos,2i+1)} &= \cos\left(\frac{pos}{10000^{2i/d_{\mathrm{model}}}}\right), \end{aligned}$$

where $\mathrm{PE}_{(pos,2i)}$ are the positional embeddings for tokens at even positions, $\mathrm{PE}_{(pos,2i+1)}$ are the positional embeddings for tokens at odd positions, and d_{model} is the dimensionality of the embeddings (1,536 dimensions for GPT-4).

These positional encoding vectors are combined with the input embeddings of tokens, injecting positional information into the model's representation of the input. By incorporating positional encoding, GPT ensures that the model can effectively capture the sequence structure and relationships between tokens based on their positions.

Additionally, the decoder is composed of multiple layers of self-attention mechanisms. Each layer processes the input sequence independently and captures dependencies within the sequence. The self-attention mechanism allows the model to assign different importance to each token based on its relevance to other tokens in the sequence, enabling the model to understand context and generate coherent and contextually appropriate text.

After the self-attention layers, each position in the sequence passes through a position-wise feedforward neural network. This network consists of fully connected layers with non-linear activation functions, enabling the model to capture complex patterns and interactions in the data. The position-wise feedforward networks refine the token representations, incorporating both local and global context information.

To further stabilize training and improve the flow of gradients, GPT incorporates layer normalization and residual connections after each self-attention layer and position-wise feedforward network. Layer normalization normalizes the activations within each layer, reducing internal covariate shifts and enhancing training stability. Residual connections allow gradients to bypass certain layers, mitigating the vanishing or exploding gradient problem that can arise in deep neural networks.

K.2.2 Self-attention

The purpose of the self-attention mechanism is to generate a numerical embedding for each piece of text while accounting for the contextual relationships among all tokens in the text. Specifically, the raw input to the attention mechanism is a piece of text T. This text is divided into sub-word tokens through a process called tokenization. The tokenization process relies on a predefined set of tokens that can be combined to represent a vast number of unique words. For example, the prefix "un" is a token in many models because it negates the meaning of many other sub-word tokens, such as "happy." Other tokens may represent short and common words like "and."

After tokenization, each token is assigned an initial embedding that combines the token's semantic meaning and its position within the text. This results in a set of initial embeddings:

$$\text{EMB}_0 = [[\text{BOS}], t_1, \ldots, t_N, [\text{EOS}]],$$

where t_i represents the embedding for a token i, "[BOS]" (beginning of sentence) denotes the start of the sentence, and "[EOS]" (end of sentence) signifies the sentence's conclusion.

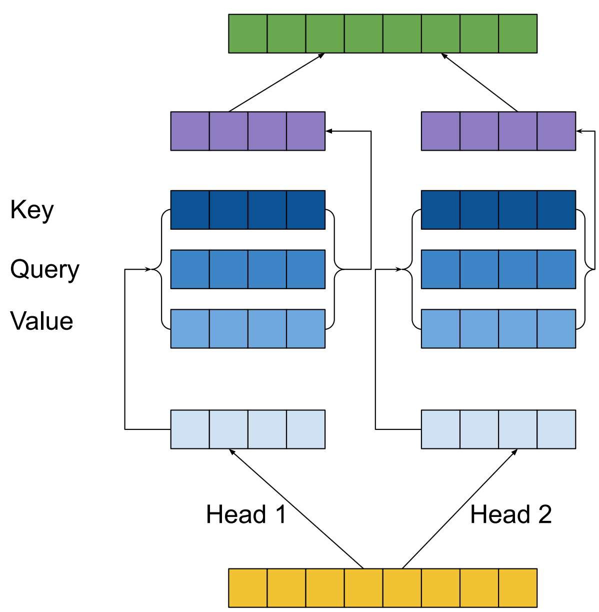


Figure K2: Diagram of a multi-head attention layer

As shown in Figure K2, a multi-head self-attention layer processes an embedding as input and produces a refined embedding as output. The input embedding is passed through three parallel linear mappings to form three matrices: the key matrix, the query matrix, and the value matrix.

$$Q = \mathrm{EMB}_0 W^Q,$$
$$K = \mathrm{EMB}_0 W^K,$$
$$V = \mathrm{EMB}_0 W^V,$$

where Q, K, and V are trainable parameter matrices. For each query, a cosine similarity score is calculated between the query and all keys, including the query itself. The value of the token is then represented as a weighted combination of all the token values in the text, with the weights determined by the cosine similarities between queries and keys. Mathematically, this is represented as:

$$\mathrm{Attention}(\mathrm{EMB}_0) = \mathrm{softmax}(QK^T)V.$$

In the GPT model, the attention mechanism is typically computed as:

$$\mathrm{Attention}(\mathrm{EMB}_0) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where d_k is a scaling factor equal to the number of columns in K. To enhance the model's capacity to capture various aspects of the input, the embeddings EMB_0 are often divided into multiple sub-vectors of equal size. Attention is computed for each sub-vector independently, and the results are concatenated to produce a multi-head attention output for the input embedding. This attention mechanism is typically repeated multiple times, with the output of the $(i-1)$th attention layer being normalized and combined with its input to serve as the input for the ith attention layer. In GPT-4, each embedding has 1,536 dimensions.

K.2.3 Pre-training

GPT is trained using the autoregressive language modeling task, which involves predicting the next token in a sequence based on its preceding context. This task is mathematically formulated as maximizing the log-likelihood of observing the next token x_{i+1} given the preceding tokens $x_1, x_2, \ldots, x_i$ and the model parameters θ. The objective can be expressed as:

$$\mathcal{L}_{\mathrm{pretrain}}(\theta) = \sum_{i=1}^{n-1} \log P(x_{i+1}|x_1, x_2, \ldots, x_i; \theta),$$

where $\mathcal{L}_{\text{pretrain}}(\theta)$ represents the training objective, and θ denotes the model's parameters.

The goal of this pre-training task is to enable the model to learn the intricate patterns and dependencies present in natural language. By predicting the next token based on its context, GPT captures both syntactic and semantic structures, allowing it to generate text that adheres to grammatical rules and maintains coherence. The autoregressive nature of the training process further encourages the model to understand long-range dependencies, enabling it to consider information across a wide span of tokens.

Through backpropagation and gradient descent, the model gradually adjusts its parameters to minimize the negative log-likelihood of the next token, improving its ability to capture nuanced linguistic patterns and generate coherent, contextually appropriate text. The following list includes some of the data sources used to pretrain the base model of GPT-4:

1. **Common Crawl**: A vast dataset consisting of web pages collected from across the Internet, offering a broad variety of text data.

2. **Wikipedia**: Articles from Wikipedia in various languages and domains, providing structured and comprehensive information on a wide range of topics.

3. **BooksCorpus**: A collection of books spanning different genres and authors, allowing the model to learn from literary works and fictional narratives.

K.2.4 Reinforcement Learning with Human Feedback (RLHF)

GPT-4 significantly enhances its text generation capabilities through a process known as reinforcement learning with human feedback (RLHF). This method involves training the model not only through traditional supervised learning but also by incorporating direct feedback from human evaluators to fine-tune its behavior.

In the RLHF framework, GPT-4 generates text samples based on its current model, and these samples are then evaluated by human judges or annotators. The human feedback provided acts as a reward signal, guiding the model towards generating higher-quality text over time.

Formally, consider a set S representing all possible text samples that GPT-4 can generate. The model produces text samples according to its current policy π_θ, which is parameterized by θ. Each generated sample $s \in S$ is then evaluated by human judges, resulting in a feedback signal $r(s)$. This feedback signal $r(s)$ reflects the quality and desirability of the generated text, as determined by human evaluators.

The objective of GPT-4 in this framework is to learn an optimal policy π_θ that maximizes the expected cumulative reward across the distribution of text samples. Mathematically, this can be represented as the following optimization problem:

$$\max_\theta \mathbb{E}_{s\sim\pi_\theta}[r(s)],$$

where $\mathbb{E}_{s\sim\pi_\theta}[r(s)]$ denotes the expected reward across the distribution of text samples generated by the model.

To achieve this, GPT-4 employs policy gradient methods, a class of algorithms in reinforcement learning that adjust the model's parameters θ based on the gradient of the expected reward. The model iteratively updates its policy to increase the likelihood of generating text that receives positive feedback from human judges.

The RLHF process involves multiple iterations of text generation, evaluation, and feedback, which gradually refine the model's behavior. By directly incorporating human judgments into the training loop, GPT-4 learns to align its text generation more closely with human preferences, producing text that is not only syntactically correct but also contextually relevant, coherent, and aligned with desired outcomes.

Moreover, RLHF enables GPT-4 to address complex generation tasks where the quality of output cannot be easily quantified by automated metrics alone. By leveraging human expertise, the model can better understand and adapt to nuanced requirements, leading to more effective and reliable text generation.

L Debiased machine learning

In this section, we discuss the debiased machine learning in more detail. First, we explain the overview and then we discuss more details about the cross-fitting technique and Neymann orthogonality condition.

L.1 Overview

The debiased machine learning method proposed in Chernozhukov et al. (2018) is designed to address the challenges posed by high-dimensional nuisance parameters in the estimation of low-dimensional parameters of interest. In our study, we control for continuous confounders such as current profit and current employment and we also control for over 1000 discrete variables covering inter-industrial and inter-temporal variations in the data. Traditional approaches often suffer from biases introduced by regularization techniques necessary to handle high-dimensional data, leading to inconsistent estimates. To mitigate this issue, the method leverages cross-fitting. Cross-fitting is a technique that further reduces bias by splitting the data into multiple folds. The nuisance parameters are estimated in one part of the data and then used to estimate the parameter of interest in another part. This process is repeated across different folds, and the results are averaged to obtain the final estimate. In our study, we do cross-fitting in two folds as discussed in section 6.3.

In our application, we aim to estimate the association between IDE and profit growth values over each of the next 7 years while controlling for a large set of confounders X (including profit, employment, year, industry, etc). In this debiased machine learning approach, we model IDE and profit growth as non-parametric functions of the confounders.

$$IDE_{f,t} = g_T(x_{f,t}) + \epsilon^T_{f,t} \tag{15}$$

$$\Pi_{f,t+1} - \Pi_{f,t} = g^1_O(x_{f,t}) + \epsilon^{1,O}_{f,t} \tag{16}$$

$$\vdots$$

$$\Pi_{f,t+7} - \Pi_{f,t} = g^7_O(x_{f,t}) + \epsilon^{7,O}_{f,t} \tag{17}$$

There are 7 parameters of interest denoted by τ^{IDE}_k:

$$\epsilon^{1,O}_{f,t} = \tau^{\text{IDE}}_1 \epsilon^T_{f,t} + \epsilon_{1,f,t} \tag{18}$$

$$\vdots$$

$$\epsilon^{7,O}_{f,t} = \tau^{\text{IDE}}_7 \epsilon^T_{f,t} + \epsilon_{7,f,t} \tag{19}$$

where $\epsilon^{k,O}_{f,t}$ and $\epsilon^T_{f,t}$ are the residuals estimated in system 15, and $\epsilon_{k,f,t}$ are the unexplained variations in 18. The intuition of this debiased machine learning method can be explained in two steps. First, we remove any information in the treatment and outcome variables that are associated with the set of high-dimensional confounders. Then, the residuals capture variations in treatment and outcome that do not depend on any of the confounders. Therefore, in the second stage, when we run a regression of the residuals of the outcome variables on the residuals of the treatment, the resulting coefficient only captures changes in the outcome that are associated with changes in the treatment variable.

L.2 Implementing cross-fitting

Cross-fitting is a technique employed in debiased machine learning to reduce bias in the estimation of parameters. The core idea is that when we use machine learning models (g^T and g_O^k in system 15) to fit the entire sample, these highly complex models may lead to overfitting of the data instead of estimating the true underlying relationships.

1. **Data Splitting:** The dataset is randomly divided into two disjoint subsets: *Training Set 1* and *Training Set 2*. Each subset is used for different purposes in the estimation process to avoid overfitting and to mitigate the influence of bias introduced by the machine learning methods used to estimate the nuisance parameters.

2. **Nuisance Parameter Estimation:**

   - *Training Set 1:* The nuisance parameters, denoted by η, are estimated using the observations in Training Set 1. These nuisance parameter estimates, $\hat{\eta}_1$, are then applied to the observations in Training Set 2.
   - *Training Set 2:* Similarly, the nuisance parameters are estimated using Training Set 2, resulting in $\hat{\eta}_2$, which are then applied to the observations in Training Set 1.
   - **Treatment parameter Estimation:** Pool together the predictions on the two sets. Compute the residuals like the system of equations specified in 15, and run another set of regressions with the residuals like system 18.

L.3 Neymann Orthogonality

The Neymann orthogonality condition ensures that the moment conditions used for estimating the parameter of interest are locally insensitive to the estimation errors of the nuisance parameters. Let θ_0 be the low-dimensional parameter of interest and η_0 represent the high-dimensional nuisance parameters. The orthogonality condition requires that the score function $\psi(W;\theta,\eta)$ satisfies:

$$\begin{aligned} \mathbb{E}[\psi(W;\theta_0,\eta_0)] &= 0, \\ \partial_\eta \mathbb{E}[\psi(W;\theta_0,\eta_0)] &= 0, \end{aligned}$$

where W denotes the observed data, and ∂_η is the derivative with respect to η. This condition implies that small deviations in the estimation of the nuisance parameter η do not affect the score function's expectation, making the estimator robust to errors in η.

In the context of debiased machine learning (DML), the orthogonality condition is used to construct a score function that reduces sensitivity to the regularization bias from machine learning methods. Suppose the moment equation is defined as:

$$\psi(W;\theta,\eta) = (Y - D\theta - g(X))(D - m(X)),$$

where $g(X) = \mathbb{E}[Y|X]$ and $m(X) = \mathbb{E}[D|X]$ are functions estimated using machine learning techniques. The DML estimator leverages the orthogonality condition by approximately removing the effect of regularization bias from the nuisance parameters, ensuring that:

$$\mathbb{E}[(D - m(X))(g(X) - \hat{g}(X))] \approx 0,$$

where $\hat{g}(X)$ is an estimator of $g(X)$.

As shown in Chernozhukov et al. (2018), the bias term in this score function can be written as

$$\mathbb{E}\left(V^2\right)^{-1} \frac{1}{n} \sum_{i \in I} (\hat{m}_0(X_i) - m_0(X_i))(\hat{g}_0(X_i) - g_0(X_i))$$

where $V = D - \hat{m}_0(X)$, n is the number of testing samples, and I is the set of testing samples.

Therefore, we need the machine learning estimators to have unbiased errors which can be satisfied by many machine learning models.[26]

M Intoduction of Gradient Boosting

M.1 Gradient Boosted trees

Gradient Boosted Trees (GBT) is an ensemble learning technique that combines the predictions of multiple weak learners, decision trees, to form a strong predictive model. The idea behind GBT is to iteratively add trees to the ensemble, each tree correcting the errors made by the previous ones. The model starts with an initial tree, and subsequent trees are trained on the residual errors (the difference between the actual target values and the current model predictions). This process is known as gradient descent in function space because each tree is added in a direction that minimizes the loss function, typically a mean

[26] Near orthogonality is satisfied in our application: the estimation errors in m and g are weakly correlated (always lower than 6%), and the estimation errors are all insignificant from 0 except for the profit growth over the next year. Nonetheless, the average error for next year's profit growth is small in magnitude at 0.4%.

squared error for regression tasks or a log-loss for classification tasks.

Mathematically, we can express the model's prediction after m trees as:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \eta f_m(\mathbf{x}_i) \tag{20}$$

where $\hat{y}_i^{(m)}$ is the prediction for instance i after m trees, $\hat{y}_i^{(m-1)}$ is the prediction after $m-1$ trees, η is the learning rate, and $f_m(\mathbf{x}_i)$ is the prediction of the m-th tree. The objective is to minimize the loss function $L(y_i, \hat{y}_i^{(m)})$, which can be approximated by using a first-order Taylor expansion around the current prediction:

$$L(y_i, \hat{y}_i^{(m)}) \approx L(y_i, \hat{y}_i^{(m-1)}) + \frac{\partial L(y_i, \hat{y}_i^{(m-1)})}{\partial \hat{y}_i^{(m-1)}} \cdot f_m(\mathbf{x}_i) \tag{21}$$

Here, $\frac{\partial L(y_i, \hat{y}_i^{(m-1)})}{\partial \hat{y}_i^{(m-1)}}$ represents the gradient of the loss function with respect to the previous prediction, guiding the new tree to correct the errors of the ensemble model.

M.2 XGBoost: Advanced Gradient Boosted Trees

M.2.1 Intuition of XGBoost

XGBoost (eXtreme Gradient Boosting) is a highly efficient and scalable implementation of Gradient Boosted Trees, widely used in machine learning competitions and industry applications due to its performance and speed. XGBoost builds upon the fundamental concept of Gradient Boosted Trees (GBT), where multiple decision trees are sequentially added to correct the residual errors of previous models. However, XGBoost introduces various optimizations and enhancements, making it faster and more effective than standard GBT.

Similar to a standard Gradient Boosted Tree, the intuition behind XGBoost is that each new tree added to the ensemble helps to correct the mistakes of the existing trees by focusing on the most challenging instances. XGBoost incorporates both first-order and second-order (Hessian) information, which allows it to perform more precise updates in minimizing the loss function. This approach enables XGBoost to converge faster and achieve better performance with fewer trees.

M.2.2 Mathematical Formulation

As discussed in equation 20, the model's prediction after adding the m-th tree can be mathematically expressed as:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \eta f_m(\mathbf{x}_i)$$

XGBoost optimizes an objective function that includes a loss term $L(y_i, \hat{y}_i^{(m)})$ and a regularization term $\Omega(f_m)$, which controls the complexity of the model to prevent overfitting:

$$\text{Objective} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(m)}) + \sum_{m=1}^{M} \Omega(f_m) \tag{22}$$

The loss function can be approximated using a second-order Taylor expansion around the current predictions:

$$L(y_i, \hat{y}_i^{(m)}) \approx L(y_i, \hat{y}_i^{(m-1)}) + g_i f_m(\mathbf{x}_i) + \frac{1}{2} h_i f_m(\mathbf{x}_i)^2 \tag{23}$$

where:

- $g_i = \frac{\partial L(y_i, \hat{y}_i^{(m-1)})}{\partial \hat{y}_i^{(m-1)}}$ is the gradient of the loss function.
- $h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(m-1)})}{\partial \hat{y}_i^{(m-1)2}}$ is the Hessian (second derivative), which provides information on the curvature of the loss function.[27]

M.2.3 Controlling Overfitting with Early Stopping

We use a technique known as early stopping to prevent overfitting when training the XGBoost models, especially when training models with a large number of trees. Overfitting occurs when the model learns noise and patterns specific to the training data rather than general trends, resulting in poor performance on unseen data.

Early stopping monitors the model's performance on a validation set during training and stops the training process if the performance does not improve for a specified

[27] For a more detailed discussion for the implementation of these models see Erickson et al. (2020).

number of consecutive iterations (trees). This is done by specifying a patience parameter, commonly referred to as 'early_stopping_rounds'. If the evaluation metric does not improve after the given number of rounds, training halts, and the model reverts to the best iteration observed. In our analysis, we set this hyperparameter as 5 rounds.

Mathematically, let $M(t)$ be the evaluation metric on the validation set at iteration t. Early stopping stops training at iteration t^* if:

$$M(t) \geq M(t^*) \quad \text{for all} \quad t^* \leq t < t^* + \text{early\_stopping\_rounds} \tag{24}$$

where t^* is the best iteration so far. This process helps to prevent adding unnecessary complexity to the model and ensures that the training process stops when the validation performance starts to degrade, effectively controlling overfitting.

By using early stopping, XGBoost finds the optimal number of boosting rounds automatically, enhancing generalization and improving model performance on new data.

N Total similarity and firm growth

In this section, we discuss the association between our total similarity measure and firms' growth in market share, output, employment, capital stock, and intangible assets.

We use regressions analogous to subsection 5.2 but replace the independent variable of interest – IDE– with total similarity. As shown in table J3, first, the associations between total similarity and the focal firm's growth in output, employment, and market share (relatively weak for employment) are all significantly negative. This provides validation that our updated measure behaves like a proxy for horizontal product competition. In addition, we observe that total similarity is not significantly associated with the devaluation of intangible assets as seen with our IDE measure, and its association with the depreciation of capital stock is weak.[28]

[28]The sample size in the main IDE dataset regressions is approximately 2% larger than in the total similarity dataset. This difference arises primarily due to two factors. First, there are 320 firms that reported 10-Ks in at least one of the past five years but not in the current year, yet they remain in Compustat and are therefore included in our dataset. A key reason for this is that some firms moved abroad and were no longer required to file 10-Ks. Second, one firm in the total similarity dataset—SAVIENT PHARMACEUTICALS INC in 2005—is missing from IDE due to its technology summaries exceeding the length limit for the embedding model.

			Output			
(1)	(2)	(3)	(4)	(5)	(6)	(7)
-0.006***	-0.013***	-0.015***	-0.020***	-0.020***	-0.025***	-0.029***
(0.002)	(0.003)	(0.003)	(0.005)	(0.006)	(0.006)	(0.007)
			Capital stock			
(1)	(2)	(3)	(4)	(5)	(6)	(7)
-0.003**	-0.005*	-0.006	-0.009*	-0.010*	-0.014**	-0.029***
(0.001)	(0.003)	(0.004)	(0.005)	(0.006)	(0.006)	(0.007)
			Employment			
(1)	(2)	(3)	(4)	(5)	(6)	(7)
-0.003*	-0.006**	-0.007**	-0.008*	-0.007	-0.010	-0.012
(0.002)	(0.003)	(0.003)	(0.004)	(0.006)	(0.006)	(0.007)
			Intangible asset			
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0.007	0.008	0.007	0.006	-0.003	-0.003	-0.009
(0.005)	(0.008)	(0.009)	(0.009)	(0.010)	(0.013)	(0.011)
			Market share			
(1)	(2)	(3)	(4)	(5)	(6)	(7)
-0.002	-0.008**	-0.011**	-0.015**	-0.015**	-0.019***	-0.023***
(0.002)	(0.003)	(0.004)	(0.005)	(0.006)	(0.006)	(0.007)

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE J3: This table shows the association between total similarity and firm growth outcomes other than profit growth. In particular, we run the regression specified in equation 2 while replacing profit growth with growth in capital stock, output, employment, intangible assets ($\frac{\text{intangible asset}}{\text{book equity}}$), and market share. We control for the log asset growth and profitability among other variables – firm and industry-level innovation value, log of the current profit, employment, and capital stock. In addition, we use fixed effects of year, industry code, and year cross industry category. The data spans from 2005 to 2015. All independent variables are standardized within each industry category and year, and all standard errors are clustered at the firm-year level. Asset growth is computed as the ratio between the total assets in year t and $t-1$ and profitability is defined as total sales minus cost of goods sold divided by total assets.

O Data appendix

In this section, we provide a detailed discussion of our data sources and the steps we take to merge different datasets. We also provide a complete variable dictionary.

O.1 Dataset construction

We acquire firm characteristics from COMPUSTAT's global dataset up to the fiscal year 2022. Then, we obtain the consumer price index and equipment deflator data fromm the Federal Reserve of St. Louis. Lastly, we obtain the firm and industry-level innovation value data from the KPSS data repository. All of the datasets are merged based on year and the unique firm identifier PERMNO.

The next step in our data processing is to scale firm growth variables with a dollar value to a common baseline. This is because the value of a dollar is not comparable across different years. In our study, we consider 4 outcomes, profit growth, capital stock growth, output growth, and employment growth. We scale the first three growth variables by dividing each firm's profit and output in a year by the consumer price index of that year and capital stock by the equipment deflator. Note that some of our independent variables in our analysis such as R&D expenditures also have dollar values; however, since we standardize all independent variables by year and industry category (or at least by year in Appendix F), we do not need to further scale these variables.

Next, we use the 30 industry categories defined by the Fama-French data repository. The categories include Personal and Business Services; Business Equipment; Healthcare, Medical Equipment, and Pharmaceutical Products; Retail; Petroleum and Natural Gas; Wholesale; Communication; Everything Else; Construction and Construction Materials; Fabricated Products and Machinery; Transportation; Restaurants, Hotels, and Motels; Recreation; Consumer Goods; Food Products; Chemicals; Automobiles and Trucks; Steel Works; Precious Metals, Non-Metallic, and Industrial Metal Mining; Apparel; Electrical Equipment; Printing and Publishing; Aircraft, Ships, and Railroad Equipment; Textiles; Beer and Liquor; Tobacco Products; and Coal.

The resulting data contains 193,518 firm-year data points. In addition, following the procedure outlined in section 2, we construct 63,205 innovation embeddings and 40,923 technology stacks. Note that the innovation stacks ranges from the year 1980 to 2015, and technology stacks are from 2005 to 2015. Finally, we merge the firm characteristics, technology embeddings, and innovation embeddings and drop missing values in IDE, current profit, profitability, $\frac{\text{total sales}-\text{cost of goods sold}}{\text{total asset}}$, capital stock, employment, asset growth, $\frac{\text{total asset in the current year}}{\text{total asset in the previous year}}$, PERMNO, year, industry code, and industry category. The final dataset contains 28,075 firm-year data points.

O.2 Variable dictionary

- Profit: total sales minus cost of goods sold. Depreciated by the consumer price index.
- Profit growth (T years ahead): The log of the focal firm's profit T years in the future minus the log of profit in the current year.
- Capital stock: Property, Plant, and Equipment (Gross) divided by equipment deflator.
- Current employment: The number of employees.
- Profitability: profit divided by total assets.
- Asset growth: Total assets in the current year divided by the total assets in the previous year.
- Output: total sales minus total divident. Depreciated by the consumer price index.
- PERMNO: unique firm identifier used to merge datasets.
- Industry code: The 3-digit SIC industry code.
- Industry category: The 30 industry categories defined by Fama-French.
- Capital expenditures, R&D expenditures, and market capitalization are based on the definition in COMPUSTAT.
- Firm and industry-level innovation values are based on the KPSS data repository.