Tasks to TFP: The Micro Origins of Aggregate Productivity

Joseba Martinez* Johan Moen-Vorum†

March 2026

Abstract

We introduce a new measure of innovation, task relevance, which links the technology covered by patents to the production tasks performed in the U.S. economy over the past century. We first create a representation of the task space based on textual descriptions of work, and second, measure the semantic relevance of U.S. patents to tasks using natural language processing tools. Unlike existing metrics, this relevance measure can be computed for any patent upon publication, while still correlating strongly with established measures of patent quality, such as citations, importance, and value. To demonstrate the explanatory power of task relevance as a measure of technological change, we revisit two classic questions. Cross-sectionally, we find that relevance explains a substantial share of labor market polarization in task wages and hours, in line with Autor and Dorn (2013). At the aggregate level, we demonstrate that: i) relevance shocks transmit in a manner consistent with classic aggregate technology shocks, raising output lowering prices, and have persistent positive effects on TFP, explaining a sizable fraction of its variance; and ii) in out-of-sample forecasts, task relevance significantly improves predictions of medium- to long-run TFP growth relative to benchmarks.

*London Business School & CEPR (jmartinez@london.edu)

†BI Norwegian Business School (johan.moen-vorum@bi.no)

Joseba Martinez thanks the LBS RAMD fund and the Wheeler Institute for Business and Development for financial support. Neha Kekre provided excellent research assistance.

1 Introduction

When Abramovitz (1956) called productivity growth a 'measure of our ignorance', he could hardly have imagined how prophetic those words would prove. For over half a century, the Solow residual has been the gold standard for measuring technical progress. Yet this approach remains a "black box", unable to disentangle the underlying causes of productivity growth. To advance our understanding of these causes, this paper employs text analysis to directly measure a fundamental aspect of technical change: the relevance of technologies for performing economically meaningful tasks.

We propose a new metric, *task relevance*, that formalizes this concept. The metric is agnostic as to whether technology replaces or complements humans or, indeed, improves capital's ability to perform tasks. Our goal, instead, is to construct an intuitive and simple metric that can be computed using publicly available sources and methods. Belying the simplicity of task relevance, we find that it adds information relative to established patent quality metrics and has significant explanatory power in both cross-sectional and aggregate applications.

Our study leverages recent advances in text analysis to quantify the relevance of patents to tasks. Drawing on two databases that describe the nature of work in the U.S. over the past century—the Dictionary of Occupational Titles (DOT) and the Occupational Information Network (O*NET)—we create a representation of the task space spanning our sample period. Using natural language processing techniques, we measure the textual relevance of all U.S. patents granted since 1900 to these tasks, thereby constructing a panel of task relevance measures spanning more than a century.

Our measure has advantages over existing approaches. First, relevance can be computed as soon as the text of a patent is published, unlike citation counts or the importance metric from Kelly et al. (2021), which can only be calculated retrospectively. Second, relevance can be computed for all patents, not just those linked to publicly traded firms—unlike the patent value measure of Kogan et al. (2017)[1]. It is also less confounded by market power, economies of scale, or other firm-level considerations. Third, because tasks can be classified along economically meaningful dimensions (manual vs. cognitive, routine vs. nonroutine, collaborative vs. independent), our relevance metric enables a rich study of the bias of technological progress.

Our focus is not the study of specific technologies, such as automation (Dechezleprêtre et al., 2020; Mann and Püttmann, 2023) or artificial intelligence (Webb, 2020; Hampole et al., 2025), but the methodology could be made specific by computing our relevance metric in subsets of patents. And, unlike Kogan et al. (2023), we are agnostic as to whether task relevance augments or substitutes workers.

We use our metric to conduct three exercises. At the patent level, we show that relevance is strongly correlated with citations, importance, and value, and we uncover systematic biases: patents associated with manual or routine tasks are less cited, important, and valued, while those linked to collaborative tasks are more so.

Aggregating our measure to the task level, we revisit the classic Autor and Dorn (2013) finding of labor market polarization. Recasting polarization in terms of tasks rather than occupations, we find the familiar U-shaped pattern of wage and hours growth across the task distribution. Projecting these outcomes onto task relevance reveals that relevance accounts for a substantial share of polarization in the middle-to-right

[1]See Kline et al. (2019) for an approach to extending patent value estimates beyond publicly traded firms by linking patent-holder wages to patent grants.

tail, but less so in the left tail—consistent with the Autor and Dorn (2013) result that technology primarily drives the upper-tail dynamics.

At the aggregate level, we connect task relevance to the macroeconomic literature on technology shocks. Within a time series framework, we demonstrate that shocks to aggregate relevance resemble classic technology shocks, causing output and prices to move in opposite directions. Consistent with this interpretation, we show that relevance shocks have persistent positive effects on TFP, explaining a sizable fraction of its variance.

Related literature: Patents became a standard metric for measuring technological progress following the work of Hall et al. (2000) and Hall et al. (2001), with citation-weighted patent counts emerging as the dominant approach following Hall et al. (2005). There is a growing interest in developing alternative measures of technological progress beyond patent citations. Kogan et al. (2017) use the stock market response to patent publications to determine the private value of patents. They find that the private value, as measured by excess stock returns, is correlated with other measures of the scientific value of the patent and with medium-run fluctuations in TFP and economic growth. Kelly et al. (2021) use textual analysis on patent data to measure the development of breakthrough innovations, which they identify as patents that are distinct from previous inventions but similar to subsequent work. Both measures are now established approaches in the literature.

A related literature examines the labor market effects of new technologies. Webb (2020) employs a keyword-based approach to identify patents related to artificial intelligence. Dechezleprêtre et al. (2020) propose a method for identifying automation technologies based on patent classification categories. Mann and Püttmann (2023) train a machine learning classifier to categorize 'automation' and 'non-automation' patents, which they use to create an aggregate metric for automation technologies. Kogan et al. (2023) use textual analysis to determine the similarity between occupational tasks and breakthrough innovations. Autor et al. (2024) classify patents as 'augmentation' and 'automation' patents, which they link with historical data on new job titles and U.S. Census data to study the creation of new job categories.

Structure of the paper: Section 2 describes the data, while 3 details the methodology for representing the task space, measuring the relevance of patents to tasks, and determining the skill requirements of tasks. Section 6 explores the resulting task relevance measure and how it relates to aggregate growth. Section 7 concludes.

2 Data

This analysis combines four main data sources. Patent data from Google Patents Public Datasets are employed to construct task-specific measures of technological progress. Task descriptions are sourced from the Dictionary of Occupational Titles (DOT) and the Occupational Information Network (O*NET) databases. Labor data are sourced from the Current Population Survey (CPS). Finally, macroeconomic data are supplied by Bergeaud et al. (2016) and Òscar Jordà et al. (2016).

Google Patents Public Datasets are a collection of compatible database tables from government, research, and private entities for patent data analysis. To construct our measure of technological progress, we use all patents granted in the U.S. between 1900 and 2018[2].

[2]Although patent data are available from the late 1800s, measures constructed using patents granted before 1900 appear

Task data are provided by the Dictionary of Occupational Titles (DOT) and the Occupational Information Network (O*NET) databases. The DOT, produced by the U.S. Department of Labor, catalogs over 13,000 distinct occupations from 1938 to the late 1990s, with each occupation described by a set of associated tasks. The DOT was replaced by the O*NET database, which is now the primary source of occupational information in the US. O*NET includes detailed information on nearly 1,000 occupations, each characterized by requisite knowledge, abilities, skills, and typical work activities. Work activities are organized in a taxonomy, with Detailed Work Activities (DWAs) representing the most granular level of detail. We use the combined set of DOT tasks and O*NET DWAs as our task database, effectively covering all major types of work performed in the U.S. since 1938.

To validate our method and study the effects of technological progress, we employ both micro-level labor data and macroeconomic data. At the micro level, we use labor data from IPUMS CPS. The CPS is a monthly survey conducted by the U.S. Bureau of the Census, covering over 65,000 households and documenting wages and hours worked per respondent. From these data, we aggregate responses into occupation-level estimates, which we convert into task-level estimates. We use CPS data covering the period 1977 to 2019. For the macroeconomic analysis, we draw on aggregate data from Bergeaud et al. (2016) on growth and productivity, as well as CPI data from Òscar Jordà et al. (2016), both starting in 1924.

3 The Task Relevance of Technology

We aim to construct an economically meaningful set of tasks and measure how relevant each new technology is to these tasks. Conceptually, our approach extends the canonical task framework of Autor et al. (2003) using text data on both work activities and patents.

1. We start from descriptions of work activities from the DOT and O*NET, which are the standard data sources used in the task literature, and cluster them into 145 groups that we interpret as tasks. These tasks span the main types of work performed in the U.S. economy since the 1940s.

2. We describe each task along three standard dimensions: manual vs. cognitive, routine vs. nonroutine, collaborative vs. independent.

3. We then use natural language processing to measure the semantic similarity between patents and work tasks, and interpret this similarity as a measure of how relevant the patented technology is to performing that task.

3.1 Representing the Task Space

We first construct a representation of tasks performed in the U.S. economy by embedding and clustering detailed task descriptions into a low-dimensional representation of the task space. Our dataset encompasses all major types of work performed from the 1930s to the present, drawing from two sources. From the DOT, we extract all tasks listed under each occupation. From the O*NET, we use the Detailed Work Activities (DWAs).

to be noisy. This is likely due to the lower quality of the digitized patent records from that era, which affects the reliability of language model outputs using this data as input.

Since the raw data contains thousands of task descriptions with significant overlap, we use a clustering approach to reduce dimensionality and remove duplicate information. This approach consists of three steps. First, we construct word embeddings—numerical vector representations of text—for each task using a pretrained language model. Second, we perform a dimensionality reduction on these vectors to facilitate the clustering. Third, we apply a clustering algorithm to the reduced task vectors.

To construct word embeddings, we use the BERT model introduced by Devlin et al. (2018), which pioneered the transformer-based paradigm that underpins recent advances in large language models (LLMs). Despite recent advances in the field, BERT continues to perform well on shorter text excerpts (Zhang et al., 2025), making it suitable for this analysis. To perform the dimensionality reduction, we employ the UMAP method of McInnes et al. (2020). Clustering is done using the HDBSCAN algorithm of Campello et al. (2013). Appendix A describes the methodology and the modeling choices in detail. The resulting clusters are stable across a range of parameter choices (see Appendix A.1).

Applying the clustering algorithm to the 47,794 task descriptions, we identify 145 clusters. In total, 20,591 descriptions are assigned to clusters, leaving 27,203 task descriptions unassigned. The unassigned task descriptions largely consist of specialized activities that appear infrequently across occupations. Focusing on the clustered tasks captures the major types of work that drive aggregate economic relationships, while avoiding noise from idiosyncratic job descriptions. Appendix Figure 14 presents the distribution of task descriptions across clusters. We represent each task cluster by its centroid, the average of the embedding vectors of all task descriptions assigned to the cluster. While averaging individual embeddings removes some information, this dimensionality reduction is essential for making the method computationally feasible. These centroids, representing the prototypical activities within each task cluster, are what we refer to as the 'tasks'.

Quantifying task characteristics. Studying technological progress through the lens of tasks allows us to capture how technology affects work differently depending on the nature of the tasks involved. Following the extensive literature on job polarization and skill-biased technological change, we adopt the canonical framework of Autor et al. (2003) for organizing task characteristics. This framework categorizes tasks along two dimensions: routine/nonroutine and manual/cognitive. We extend this framework by introducing a third dimension, the degree to which work is collaborative versus independent, as emphasized by Deming and Kahn (2018).

To allow for a nuanced representation of task characteristics, we use numerical scores rather than binary assignments for each dimension. Our approach involves two steps: we first generate numerical scores for each of the six categories in the framework: manual, cognitive, routine, nonroutine, collaborative, and independent. We then combine these scores to create one composite score for each dimension, representing whether the work is predominantly manual or cognitive, routine or nonroutine, and collaborative or independent.

As with the method for measuring task exposures detailed in Section 3.2, we employ word embeddings to determine the semantic similarity between tasks and framework categories. Each category is represented by a dictionary definition characterizing the type of work it encompasses, reported in Appendix Table 2. We assign each of the 20,591 task descriptions a score for each category, computed as the cosine similarity between the task description embedding and the category embedding. To obtain task-level scores, we

average these similarity scores across all descriptions assigned to the same task cluster. The resulting scores are denoted by C_j^v, with $v \in \{$ manual, cognitive, routine, nonroutine, collaborative, independent $\}$.

To construct composite scores for each dimension, we combine the scores of opposing categories. In particular, for each task j, we divide a category's score by the sum of its score and the score of its opposing category. The weighted manual score, for instance, is calculated as: $C_j^{mw} = C_j^{man}/(C_j^{man} + C_j^{cog})$. This score quantifies the extent to which a task is predominantly manual as opposed to cognitive: a score of 1 indicates that the task consists of purely manual work, while a score of 0.50 indicates equal requirements for manual and cognitive characteristics. The weighted routine and collaborative scores are denoted as C_j^{rw} and C_j^{cw}, respectively, following the same weighting approach. Since the cosine similarity is unitless, this weighting scheme enhances the interpretability of the cosine similarity scores.

As a validity check on the composite scores, we examine the most prominent terms associated with tasks at the extremes of each dimension. This is done by applying a term frequency–inverse document frequency (TF-IDF) weighting to the task descriptions and identifying the most prominent terms for the highest- and lowest-scoring tasks on each dimension. Of the 145 tasks, the task with the highest manual score is characterized by terms such as 'linen', 'clean', 'bed', and 'change', consistent with physical care work. In contrast, the most cognitive task features terms like 'scheduling', 'answering', and 'appointment', consistent with administrative work. Appendix Figure 15 presents word clouds for all six categories, and Appendix Table 3 presents correlations across the three task scores.

We also validate the composite scores against an established benchmark, the task input scores of Autor and Dorn (2013). To do this, we construct occupation-level scores by first assigning each task description the score of its parent task cluster and then taking the simple average across all task descriptions assigned to each occupation. We find statistically significant, positive but modest rank correlations for the 'Manual' dimension ($\rho \approx 0.21$) and between our 'Collaborative' dimension and their 'Abstract' dimension ($\rho \approx 0.27$). In contrast, we find no statistically significant correlation between our 'Routine' dimension and theirs ($\rho \approx 0.03$). Taken together, these rank correlations suggest that the two approaches capture related underlying task characteristics, albeit with important differences across dimensions. The modest magnitude of these correlations reflects differences in construct definition between the two approaches. The measures of Autor and Dorn (2013) are built from a narrow set of DOT variables, whereas our measures are designed to capture a broader range of task content. We provide a detailed comparison of the underlying constructs and variables in Appendix B.1.

Constructing task-level labor outcomes. We use labor data from the U.S. Census and the CPS to construct task-level hours and wages. We first map occupations in these datasets to task descriptions in the corresponding DOT or O*NET vintage. We then use the assignment of task descriptions to the 145 task clusters to obtain a concordance between occupations and task clusters. Two key assumptions underlie this approach. First, we assume that workers split their time equally across the task descriptions associated with their occupation. Second, we assume that the wage for each task description equals the occupation wage. These assumptions are necessary because the earliest editions of the DOT does not include data specifying the relative importance or time allocation across tasks. Under these assumptions, hours worked per task cluster corresponds to the sum of hours across all underlying task descriptions, and the task cluster wage is the average wage of those descriptions.

Economic properties of the task space. In total, there are 145 tasks in our task representation. Mapping these to Census and CPS data results in an unbalanced panel, with fewer active tasks in the early part of the sample than in the later part. This pattern reflects both the emergence of new occupations over time and the more granular nature of the O*NET data that underlies the later part of the sample. Table 1 reports summary statistics for task-level wages and hours over the full sample. The average hourly task wage is $1.77 with a standard deviation of $0.77, and the dataset covers 145 unique tasks and 1,062 task–year observations. More detailed year-by-year statistics are provided in Appendix Table 5.

Table 1: Summary statistics, task outcomes

	Mean	Standard deviation
Hours worked per task ($\times 10^9$)	1.679	2.179
Hourly task wage (real, 1983 $)	1.77	0.74
Number of unique tasks		145
Task–year observations		1,062

Note: This table reports average hours worked per task and average hourly task wages. Observation years correspond to major DOT/O*NET task-classification updates. The underlying labor data are drawn from the U.S. Census and the CPS, and cover the period 1940-2020.

The number of active tasks in the data increases sharply between 1980 and 1990. This reflects the transition from the DOT to the O*NET classification system, which provides a much more granular description of work activities. Importantly, the underlying distribution of economic activity is smooth across this transition: both total hours worked and average hourly wages evolve without noticeable breaks (see Appendix Table 5). The jump is in task granularity, not in the underlying labor input.

The level and evolution of total hours are broadly consistent with aggregate hours worked in the U.S. economy as reported by the BEA's National Income and Product Accounts. The correspondence is especially close from 1990 onward. In earlier decades, coverage in the historical microdata is more limited, leading to some divergence. Average task-level wages exhibit a smooth upward trend throughout the full sample, consistent with the long-term growth of real wages.

To assess changes in the composition of work, we use the task characteristic scores described above. Based on the weighted manual score, we classify all tasks in the top quartile as 'manual' and those in the bottom quartile as 'cognitive'. We define 'routine' and 'nonroutine' tasks, and 'independent' and 'collaborative' tasks using the same approach. For each year, we sum the total number of hours within each task type. Figure 1 reports the share of total hours worked that falls into each category. Panel A shows the allocation of hours between manual and cognitive work. In 1940, more than 80% of all hours worked are in manual tasks, while the share of cognitive work is nearly zero. The low initial share of cognitive work reflects the fact that most cognitive tasks in our data appear later in the sample. Over time, the share of cognitive work steadily increases, while the share of manual work declines. By 2020, almost 40% of hours worked are attributed to the top quartile of cognitive tasks, while manual tasks account for only 11%.

Panels B and C of Figure 1 show analogous decompositions for routine versus nonroutine and independent versus collaborative work. The patterns across panels are similar: hours shift away from routine,

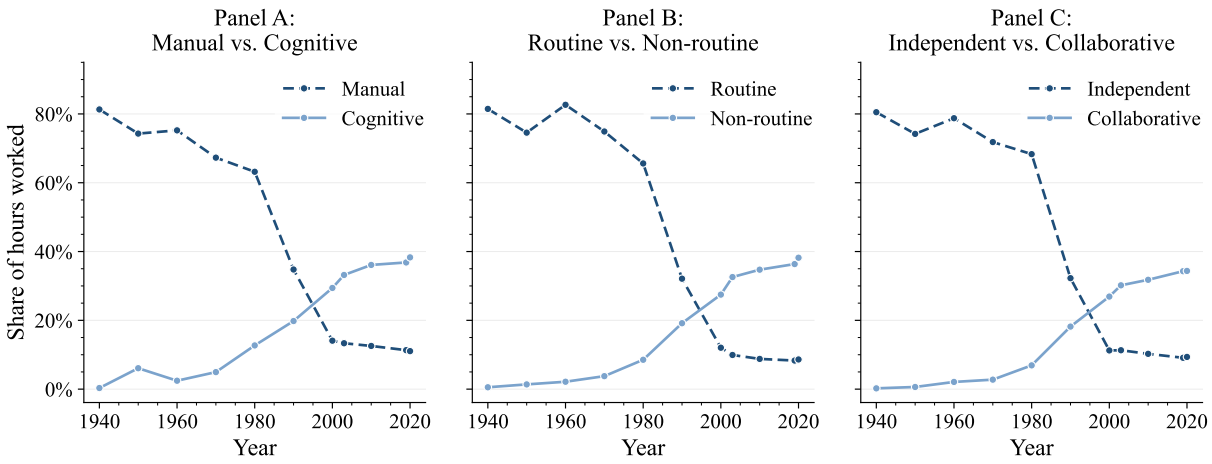


Figure 1: Share of hours worked by task characteristics

Note: This figure shows the share of hours worked in the most manual versus most cognitive tasks (Panel A), the most routine versus most nonroutine tasks (Panel B), and the most independent versus most collaborative tasks (Panel C). Tasks are classified using the top and bottom quartiles of the corresponding task characteristic scores C_j^{mw}, C_j^{rw}, and C_j^{cw}. For each year, total hours within task types are summed and normalized by total hours in the economy.

independent work and toward nonroutine, collaborative work over time. The strong correlation across the three dimensions reflects the structure of our task space: most manual tasks are also routine and performed independently, whereas most cognitive tasks are nonroutine and collaborative. Appendix Table 3 documents the correlations between task characteristic scores.

Figure 2 plots the median hourly task wage over time, together with the 10th and 90th percentiles of the task-wage distribution. While median wages rise steadily, the more striking pattern is the increase in wage dispersion. The interdecile range (P90–P10) nearly triples over the period, growing from \$0.45 in 1950 to \$1.28 in 2019. This widening reflects increasing job polarization documented in prior work (Autor et al., 2003; Autor and Dorn, 2013). We exclude 1940 and 2020 from this figure due to data quality concerns (see Appendix C for details).

To assess how the increasing wage dispersion relates to task characteristics, Figure 3 plots the wage premium for cognitive (relative to manual), nonroutine (relative to routine), and collaborative (relative to independent) tasks over time. For each year, we estimate the wage premium as the difference in mean log wages between each pair of categories and report the implied percentage difference. Confidence intervals are constructed using a stratified nonparametric bootstrap with 1,000 replications. Because cognitive, nonroutine, and collaborative tasks are relatively scarce in the early part of the sample, the corresponding premiums are less precisely estimated. Figure 3 therefore shows estimates from 1970 onwards; Appendix Figure 17 reports the full sample.

Figure 3 shows that the wage premium of cognitive relative to manual tasks was statistically indistinguishable from zero before the 1980s. Since then, the wage premium has increased sharply, reaching a peak of 35% in 2010. The wage premiums of nonroutine and collaborative work have followed similar patterns, albeit with a less pronounced increase. This pattern mirrors the rise in occupation-level wage premiums for higher education, which increased rapidly in the 1980s before plateauing between 2010 and 2015 (Valletta, 2018).

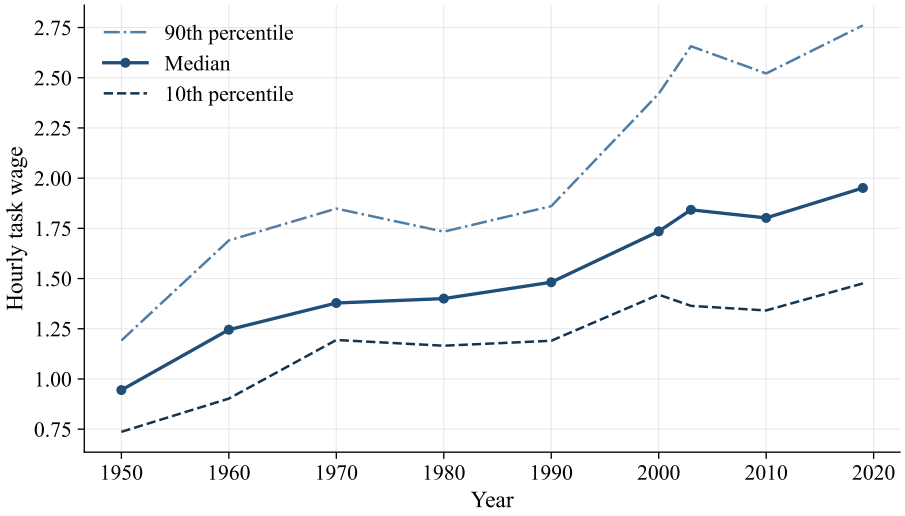


Figure 2: Distribution of task wages

Note: This figure shows the median hourly wage per task each sample year from 1950 to 2019, together with the 10th and 90th percentile of the task wage distribution. The underlying wage data are drawn from the U.S. Census and the CPS.

3.2 Measuring Task Relevance

Having constructed a representation of the task space and characterized each task along economically meaningful dimensions, we now turn to measuring the relevance of patents to these tasks. To do this, we utilize U.S. patent data provided by the Google Patents Public Datasets. Although patents are an imperfect measure of innovation, their comprehensive coverage dating back to the 1800s makes them a valuable resource for studying technical change. We use new patents as a measure of the flow of new ideas and apply word embeddings to relate these innovations to specific tasks. As described in Section 3.1, word embeddings offer a parsimonious way to represent the content of texts numerically. Consequently, the similarity between two texts can be established by comparing their vector representations. The underlying premise of this approach is simple: textual similarity between an idea and a task indicates that the idea is relevant for performing the task.

To formalize our measure, we represent the economy as a set of discrete production tasks $\mathcal{J} = \{1, \ldots, J\}$. Each task $j \in \mathcal{J}$ is characterized by a vector $v_j \in \mathcal{S}$, where $\mathcal{S} \subset \mathbb{R}^d$ is the semantic space. The vector v_j corresponds to the centroid of the task cluster defined in Section 3.1. Similarly, let $v_k \in \mathcal{S}$ represent the embedding vector for patent k. We define the relevance of patent k to task j as the cosine similarity between the task and patent embedding:

$$r_{j,k} = \frac{v_j \cdot v_k}{\|v_j\| \, \|v_k\|} \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean (L2) norm. The higher the cosine similarity $r_{j,k}$, the more semantically proximate the patent and task texts are, and, in our interpretation, the higher the relevance of the patent to the task.

For each patent, we generate word embeddings v_k by applying the BERT model to patent titles. We use

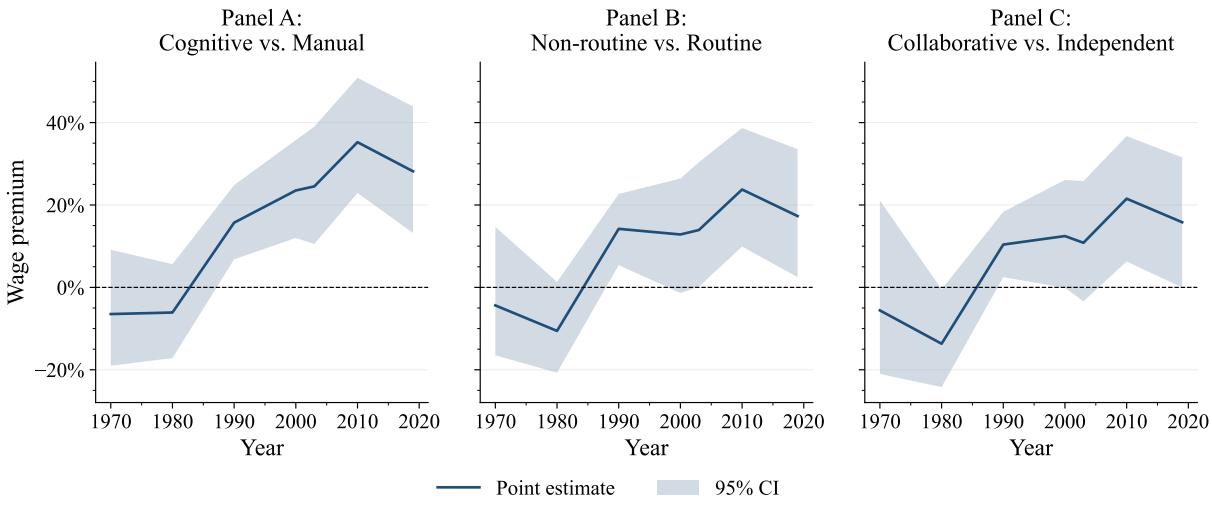


Figure 3: Wage premiums

Note: This figure shows the percentage wage premium for cognitive (relative to manual), nonroutine (relative to routine), and collaborative (relative to independent) tasks over time. Wage premiums are estimated as the difference in mean log wages between each pair of categories (e.g., cognitive vs. manual) in each year. We construct 95% confidence intervals using a stratified nonparametric bootstrap with 1,000 replications. The underlying wage data are drawn from the U.S. Census and the CPS and cover the period 1970 to 2019.

titles, rather than the abstract or the entire patent text, since these contain more textual information that is not directly related to the underlying idea, thus increasing noise. Similar to the method for constructing task embeddings v_j, we first use BERT to generate embeddings for each word in the patent title, and then average these to obtain a single embedding vector for the patent, excluding stopwords. Having generated embedding vectors v_k for all patents in the sample, we calculate the pairwise relevance $r_{j,k}$ of each patent to each task.

In the applications presented in Sections 4 and 6, we aggregate our task relevance metric in four ways. First, to create a patent-level measure of relevance, we compute the average relevance over tasks:

$$r_k = \frac{1}{J} \sum_{j=1}^{J} r_{j,k}. \tag{2}$$

Second, to compute the flow of ideas relevant to task j at time t, we compute the average task relevance for task j, for all patents filed at time t:

$$r_{j,t} = \frac{1}{|\mathcal{K}_t|} \sum_{k \in \mathcal{K}_t} r_{j,k}, \tag{3}$$

where $\mathcal{K}_t$ denotes the set of patents filed at time t, and $|\mathcal{K}_t|$ is the count of patents filed at time t. We measure relevance using the patent filing date, which provides the earliest observable record of the underlying ideas. By averaging relevance across all patents filed at time t, we abstract from fluctuations and growth in the volume of patenting. While the number of patents may also carry information, it is likely more affected by institutional changes in the patenting process and shifting incentives to file. In

contrast, our average measure is less sensitive to such distortions and thus provides a cleaner proxy for the quality of ideas.

To construct an aggregate measure of relevance, we average $r_{j,t}$ across all tasks:

$$r_t = \frac{1}{J}\sum_{j=1}^{J} r_{j,t} \tag{4}$$

Finally, to measure the rate of progress in different technology fields, we average $r_{j,t}$ across all tasks and all patents classified within the technology field g:

$$r_{g,t} = \frac{1}{J}\sum_{j=1}^{J} r_{j,g,t} \tag{5}$$

where $r_{j,g,t}$ represents the task relevance of task j in technology field $g \in \mathcal{G}$, calculated as $r_{j,g,t} = \frac{1}{|\mathcal{K}_{g,t}|}\sum_{k\in\mathcal{K}_{g,t}} r_{j,k}$, where $\mathcal{K}_{g,t}$ denotes the set of patents filed at time t in technology field g. We identify technology fields using the International Patent Classification (IPC) codes. From 650 four-digit IPC categories, we assign patents to 164 technology categories using a mapping proposed by Berkes et al. (2022), which is based on inventor information. These categories are further aggregated into 16 broader technology fields corresponding to the IPC classes to which each category belongs. These 16 fields make up our set of technology fields $\mathcal{G}$.

Aggregate task relevance and its drivers. The left panel of Figure 4 plots r_t in the sample from 1900 to 2024. Aggregate task relevance exhibits an upward trend, suggesting that technology has become increasingly relevant to production tasks over time. The rate of increase accelerates in the 1920s, 1940s, and late 1980s, and slows down in the late 1990s and post-GFC, consistent with historical episodes that witnessed faster and slower productivity growth. In Section 6, we use a time-series model to study the relationship between r_t and aggregate productivity formally.

To understand the drivers of aggregate task relevance, we employ the decomposition proposed by Olley and Pakes (1996) across technology fields. This approach decomposes the aggregate task relevance r_t into two components: the unweighted average task relevance of each field, and a covariance term capturing the covariance of patent shares and field-specific task relevance. The decomposition can be expressed as

$$r_t = \bar{r}_t + \sum_{g\in\mathcal{G}} (r_{g,t} - \bar{r}_t)(s_{g,t} - \bar{s}_t) \tag{6}$$

where $\bar{r}_t$ is the unweighted mean of task relevance across technology fields at time t, $s_{g,t}$ represents the share of patents in field g at time t, and $\bar{s}_t$ is the unweighted mean of patent shares across all fields. The first term captures broad-based shifts in technology. An increase in this term indicates that technology fields are, on average, becoming more relevant to production tasks. The second term, which we denote by Ω_t, measures the extent to which patenting is concentrated in high-relevance fields. A growing covariance indicates that patent shares and relevance are becoming more positively correlated.

Figure 4 illustrates this decomposition: aggregate task relevance, r_t (the solid line in the left panel), can be decomposed into the unweighted mean, $\bar{r}_t$ (the dashed line in the left panel), and the covariance

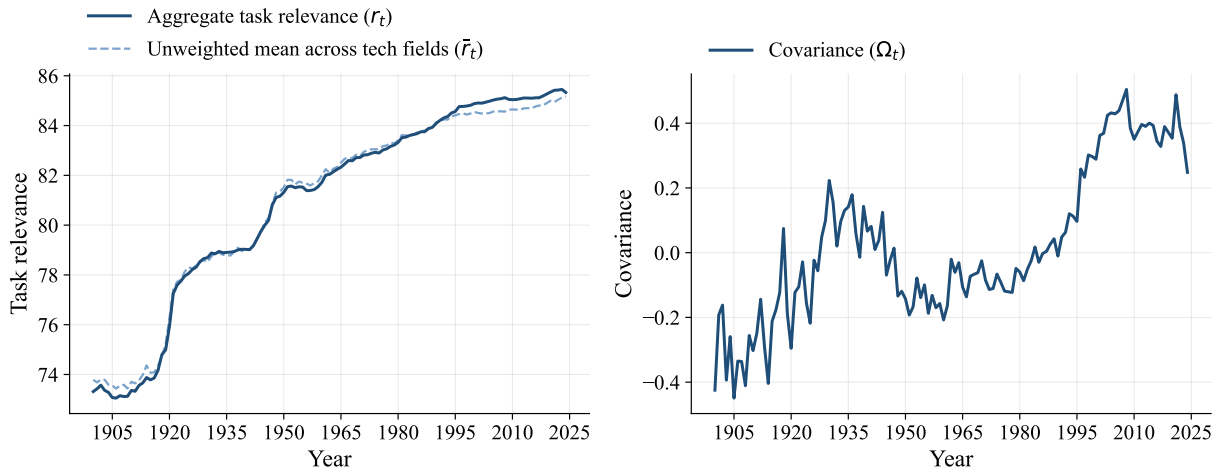


Figure 4: Task relevance of technology, 1900-2024

Note: The left panel of this figure plots the aggregate task relevance of technology, estimated using all patents granted in the U.S. from 1900 to 2024 (see Equation 4), as well as the unweighted mean of task relevance across technology fields. The right panel shows the covariance term from the Olley-Pakes decomposition of task relevance across technology fields over the same period (see Equation 6).

term, Ω_t (right panel). The right panel reveals two periods of technological reallocation over the past century, characterized by a rising covariance term. The first occurred between 1905 and 1935, and the second from roughly 1980 to 2010. These periods represent paradigm shifts, during which a small group of technologies had an outsized effect on aggregate task relevance.

To quantify the relative importance of broad-based technological progress and reallocation, Figure 5 decomposes the change in aggregate task relevance by decade. Two decades stand out as periods of particularly rapid technological progress: the 1920s and the 1940s. In both, growth stems primarily from increases in the unweighted mean, indicating that broad-based technological progress drove the rise in task relevance. Across most of the century, the covariance term plays a relatively minor role. The exception is the period from 1900 to 1930 and from 1980 to 2010, when the covariance term remains positive for an extended period.

The covariance of patent shares and task relevance increases either if fields with above-average task relevance grow their patenting share, or if fields with above-average patent shares improve their average task relevance. Formally, we can decompose the growth in covariance[3] into a reallocation effect and a quality effect:

$$\Delta\omega_g \approx \underbrace{(r_g - \bar{r})\Delta s_g}_{\text{Reallocation Effect}} + \underbrace{(s_g - \bar{s})\Delta r_g}_{\text{Quality Effect}} \tag{7}$$

where $\Delta\omega_g$ represents the contribution of field g to the overall change in covariance ($\Delta\Omega$) over the period. Variables without time subscripts (r_g, s_g, $\bar{r}$, $\bar{s}$) denote values at the start of the respective period, while Δ denotes the change from the start to the end of the period.

We determine each field's contribution by applying the reallocation and quality decomposition to the

[3]Equation 7 follows from a standard discrete first-order Taylor expansion. The exact decomposition of the total change in covariance, $\Delta\Omega = \sum_g \Delta\omega_g$, includes a second-order interaction term, $\sum_{g\in\mathcal{G}} \Delta s_g \Delta r_g$, which we omit for expositional clarity. In our empirical results, this residual term is quantitatively negligible.

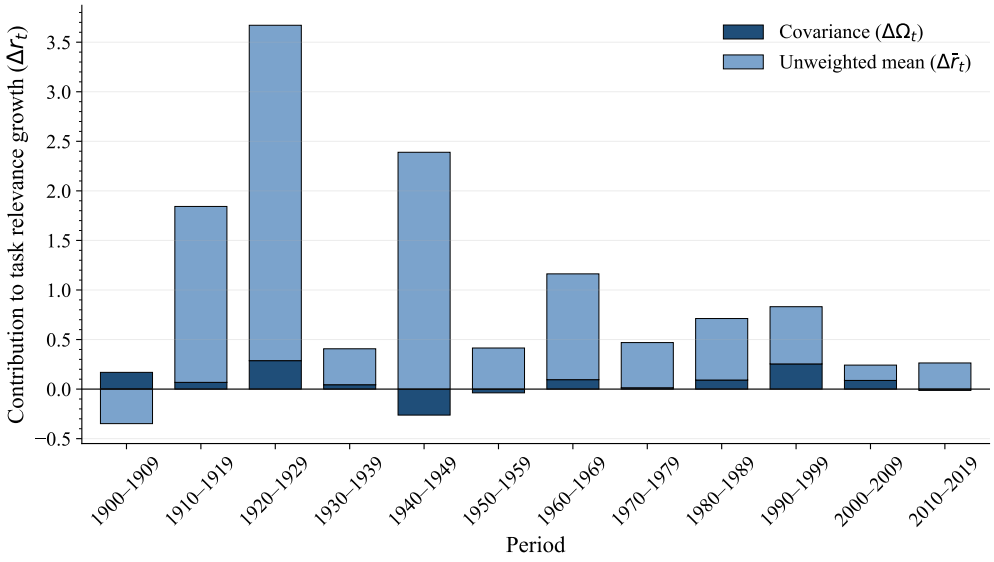


Figure 5: Contributions to task relevance growth per decade

Note: This figure shows the contributions of the unweighted mean and covariance to the change in aggregate task relevance per decade. Contributions are calculated by taking the difference between end-of-period and start-of-period values in Equation 6.

periods 1905–1930 and 1980–2010. The results are reported in Panels A and B of Figure 6, respectively, with a full overview reported in Appendix D. Between 1905 and 1930, 'Transporting' had the largest positive effect, followed by 'Shaping' and 'Chemistry'. This development aligns with historical accounts of the early twentieth century, where inventions from the Second Industrial Revolution, including the internal combustion engine and electricity, were maturing (Field, 2011; Gordon, 2016). Notably, we find that quality improvements within fields drove task relevance during this period, particularly for transportation patents.

In contrast, the quality effect was muted during the 1980–2010 period. Instead, aggregate task relevance was in large part driven by a reallocation towards 'Information technology'. This pattern is consistent with the arrival of a new general purpose technology, where growth arises from the economy shifting resources towards the expanding technology sector (Jovanovic and Rousseau, 2005).

4 Task Relevance as a Measure of Patent Quality

Next, we validate our measure by relating it to three widely-used measures of patent quality: citations, importance (Kelly et al., 2021), and value (Kogan et al., 2017). First, we compare each metric to the average task relevance for each patent (Equation 2). We then create three additional measures that capture the relevance of patents across different types of tasks. Specifically, we rank all tasks by their manual, routine, and collaborative score, and compute the ratio of a patent's relevance to above-median versus below-median tasks along each dimension. These ratios enable us to assess the relationship between existing importance metrics and the technological bias of patents, with higher scores indicating greater manual versus cognitive, routine versus nonroutine, and collaborative versus independent bias in each case. Figure 7 plots patent-level bin scatters of our new measure against the three existing ones; the

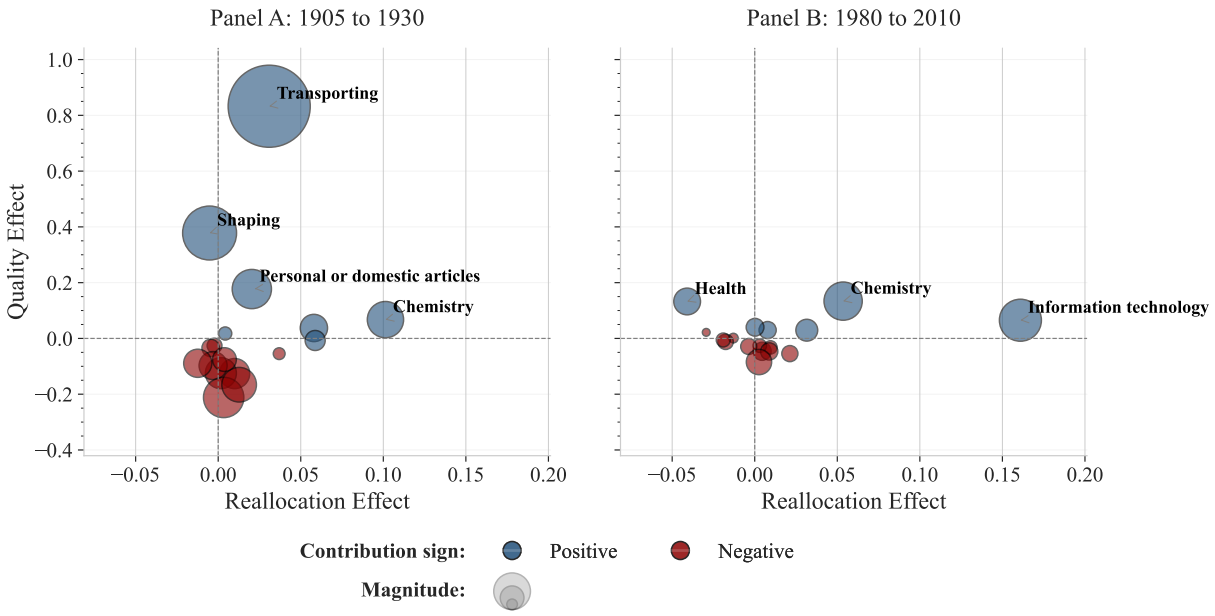


Figure 6: Decomposition of the covariance term: Reallocation vs. Quality

Note: This figure decomposes the change in the covariance term of Equation 6 into a reallocation effect and a quality effect. Panel A shows the decomposition for the period 1905-1930, and Panel B for the period 1980-2010. The size of each bubble corresponds to the absolute value of each field's contribution (|Reallocation effect + Quality effect|).

corresponding regression tables are in Appendix F.

Citations. First, we investigate the association between our task relevance measure and patent citations. The left column of Figure 7 shows binned scatter plots of patent citations against mean (top), manual (second), cognitive (third), and collaborative (bottom row) relevance, after controlling for year-by-technology fixed effects. All patent citations up until the last year of our sample (2022) are included in our measure. We can therefore interpret the top row as indicating a strong predictive relationship between mean relevance and patent citations, since our measure can be computed contemporaneously with patent publication. In contrast, citation counts are accumulated with long lags. Collaborative relevance is also positively associated with citations, whereas higher manual and routine relevance are predictive of lower future citations.

Importance. The middle column of Figure 7 shows the association between the four dimensions of task relevance and the Kelly et al. (2021) five-year importance metric. The correlations have the same sign as those for citations. Because the importance measure is forward-looking, we can, again, regard our results as predictive: more relevant patents become more influential. In Appendix F Table 9 we present the results of a linear probability model using the Kelly et al. (2021) breakthrough measure as the left-hand side variable. A one-standard-deviation increase in relevance is associated with a 0.5% higher chance of a breakthrough patent, whereas a similar increase in manual relevance has a coefficient of -3%.

Value. The right column of Figure 7 shows the association between the four dimensions of task relevance and the Kogan et al. (2017) patent value metric. The correlation between mean relevance and value (top right panel) is positive but noisy. However, there is a strong negative association between manual and routine relevance and patent values, and a particularly strong positive association between collaborative relevance and value. The effects are large: within year-technology field, a one standard deviation increase in collaborative relevance is associated with a 13% increase in value. For comparability with the other bin scatters, we do not include firm fixed effects in this exercise, but we do so in Appendix Table 10. Interestingly, within firm, the sign of the coefficients flips, so that higher manual and cognitive (collaborative) relevance are positively (negatively) associated with value. The within-firm magnitudes remain large, with a 1 standard deviation increase in collaborative relevance associated with a nearly 1% decline in patent value.

In summary, there is strong evidence that our new measure is associated with three widely used measures of patent quality. Moreover, because our measure is constructed using text available when a patent is published, it provides information relative to citations and patent importance, both of which can only be measured with significant delay (5 years for the measure of importance we have used here, and much longer for patent citations). Our analysis also suggests that task relevance, as measured on the patent publication date, is predictive of patent value, which is fully realized at the patent grant date.

5 Task Relevance and Labor Market Polarization

To investigate how task relevance influences wages and hours across different types of work, we revisit the classic finding by Autor and Dorn (2013) on labor market polarization and examine to what extent task relevance can account for it.

To do so, we construct wage and hours measures for each task in our data. We combine employment data from IPUMS CPS (Flood et al. (2024)) with work tasks from the DOT and O*NET databases. Our sample covers the years when the Bureau of Labor Statistics updated its occupational coding system: 1977, 1990, 2000, 2003, 2010, and 2019. For each year, we extract individual wage and hours data from the CPS and aggregate to occupation-level variables using survey weights.

The linking procedure proceeds in two steps. First, we map CPS occupation codes to the corresponding DOT (through 1990) or O*NET codes (2000 onward). Second, we apply the occupation-task mapping from Section 3.1 to assign each occupation to its associated tasks. The task wage is the average wage across all occupations assigned to that task. For task hours, we assume that workers distribute hours equally across their assigned tasks.

We first estimate the relationship between task relevance, wages, and hours worked using panel regressions with year and, in some specifications, task fixed effects. We also explore heterogeneity by task type, following the approach in Section 3.1. The full results are reported in Appendix G. Within task-year, hours and wages are positively associated with relevance, particularly for collaborative tasks.

To analyze the evolution of tasks across the wage distribution, we adapt the approach by Autor and Dorn (2013) from occupations to tasks. We use task–year data from 1990 and 2019 to construct changes in log hours, log wages, and relevance at the task level. A key finding from the literature on technology and worker outcomes is that different types of technology may have differential effects on workers. To explore

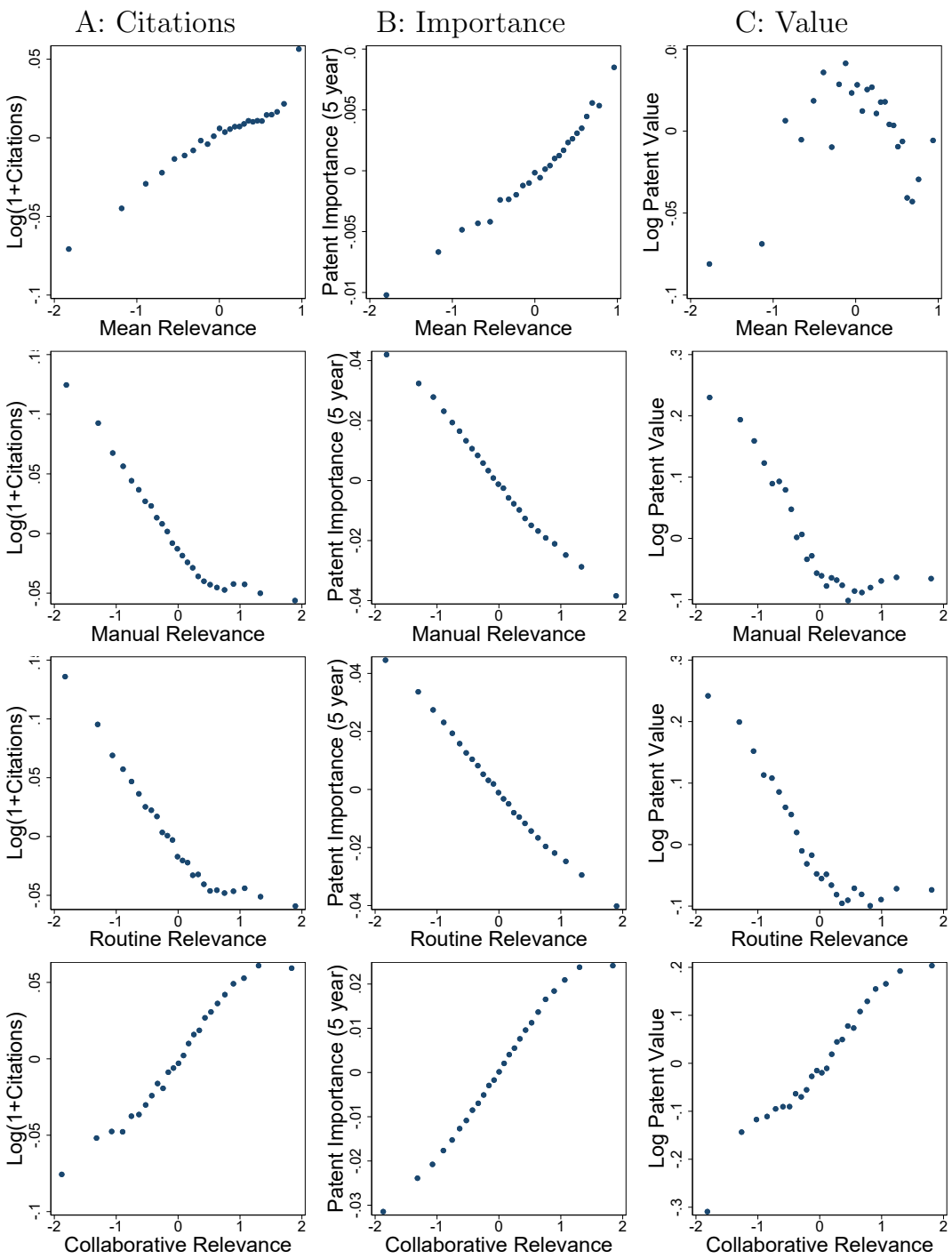


Figure 7: Patent Relevance Validation: bin scatter plots against patent citations, importance, and value.

This figure plots the relation between task relevance and citations (left column), importance (middle column), and value (right column). Moving top to bottom across rows, the regressors are task relevance, manual relevance, routine relevance, and collaborative relevance, where the latter three are constructed as ratios of average relevance in tasks above the median versus those below the median, ranked by each of the three characteristics. Before plotting, we trim the sample by excluding patents with Task Relevance below the 5th or above the 95th percentile of its distribution and standardize the relevance measures. For each panel, both the outcome and the indicated task measure are residualized with respect to filing-year × technology-field fixed effects (technology fields defined as 3-digit IPC codes). The binned scatter plots display the mean of the residualized outcome against the mean of the residualized regressor within 25 equal-frequency bins of the regressor.

this heterogeneity, we construct measures of task relevance within different technology classes, defined as one-digit IPC categories, by calculating the average task relevance of patents within each technology class in a given year to the tasks in our panel[4].

We then compute smoothed changes in hours and wages across the initial task-wage distribution, ranking tasks by their initial (1990) wage, using locally weighted regressions. To assess the contribution of task relevance, we regress changes in hours and wages on changes in task relevance over the same period and compute the (smoothed) fitted values from this regression. Figure 8 plots the (smoothed) data and the part of the variation explained by task relevance (regression tables are in Appendix G).

The solid blue line in the top and bottom panels of Figure 8 confirms that polarization is also evident when looking at the labor market through the lens of tasks rather than occupations. Both hours (top panel) and wages (bottom panel) display the characteristic U-shape across the initial (1990) task-wage ranking. Tasks in the lower and upper percentiles experience large gains in log hours, whereas middle-ranked tasks show declines—i.e., a hollowing-out of the middle. For wages, changes are high at the bottom, trough around the middle of the distribution, and rise again toward the top. Together, these patterns mirror the classic polarization profile—strong growth at the tails and relative weakness in the middle when tasks rather than occupations are ordered by their initial wage.

The dashed series ("Explained by Task Relevance") summarizes fitted values from cross-section regressions of long differences in hours or wages on long differences in Relevance. In both panels, the explained component rises with the initial rank. It tracks the upper tail fairly closely, indicating that increases in task relevance account for a substantial share of the growth among top-ranked tasks (especially for hours). The fitted model also performs reasonably well in explaining the changes in both hours and wages from the trough to the right tail, but less well in explaining the increases in the lower tail. This is not surprising through the lens of the Autor and Dorn (2013) framework, in which demand for low-skill services plays a critical role in driving the increase in low-skill employment and wages in response to technological changes. Our simple empirical model is silent on the composition of demand, but it performs well in explaining the aspects of polarization that are driven by technology.

6 Task Relevance and Aggregate Growth

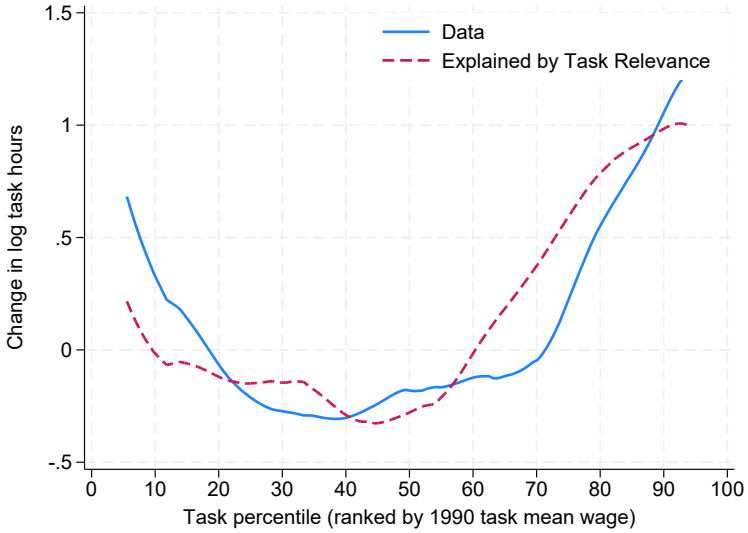
6.1 Impulse responses and variance decomposition

We now turn to the relationship between task relevance and aggregate output, prices, and productivity. We show two results. First, we demonstrate that shocks estimated using the time series of aggregate relevance (Equation 4) exhibit the hallmark of technology shocks, as they move output and prices in different directions. Second, we demonstrate that relevance shocks have a significant positive impact on total factor productivity and account for a substantial portion of the variance in TFP.

Our empirical approach utilizes lag-augmented local projections, as proposed by Montiel Olea and Plagborg-Møller (2021). This framework addresses a key concern with technology indicators, namely that their high persistence can render standard inference methods invalid. The local projections can be

[4]See , where instead of averaging all patents in year t, we average within one-digit IPC codes.

(a) Smoothed changes in task hours by task wage percentile, 1990–2019.



(b) Smoothed changes in task wages by task wage percentile, 1990–2019.

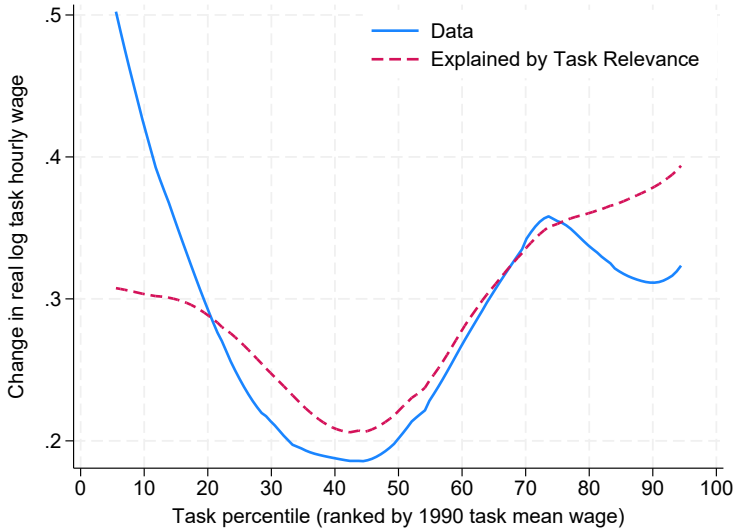


Figure 8: Polarization in task hours and wages (data and explained by task relevance), 1990–2019.

For each task cluster i, we compute long differences between 1990 and 2019: $\Delta \log(\text{hours})_i = \log(\text{hours}_{i,2019}) - \log(\text{hours}_{i,1990})$ and $\Delta \log(\text{wage})_i = \log(\text{wage}_{i,2019}) - \log(\text{wage}_{i,1990})$. The x–axis ranks tasks by their 1990 mean wage and maps ranks to a 0–100 percentile scale. The solid line in each panel plots a LOWESS (locally weighted regression) smooth of the raw change against the percentile rank using bandwidth 0.4. The dashed line plots a LOWESS smoothing of fitted values from cross-section OLS regressions of the form $\Delta y_i = \alpha + \sum_j \beta_j \, \Delta\text{Relevance}_{i,j} + \varepsilon_i$, estimated separately for $\Delta y_i \in \{\Delta \log(\text{hours})_i, \ \Delta \log(\text{wage})_i\}$. $\text{Relevance}_{i,j,t}$ is the mean relevance of task i across patents in technology class j in year t; $\Delta\text{Relevance}_{i,j} = \text{Relevance}_{i,j,2019} - \text{Relevance}_{i,j,1990}$. Technology classes are one-digit IPC codes ($j \in \{A...H\}$).

represented by the following empirical specification:

$$y_{t+h} = \alpha_h + \beta x_t + \sum_{l=1}^{p} \beta_l x_{t-l} + \sum_{l=0}^{p} \delta_l y_{t-l} + u_{t+h} \tag{8}$$

where y_t represents real GDP per capita, the Consumer Price Index (CPI), or TFP; x_t represents the technology indicator; h is the impulse response horizon; and p denotes the number of lags included in the estimation. The parameter α_h represents the intercept, with the subscript h indicating that we estimate separate regressions for each horizon, and u_{t+h} is the error term.

We first estimate Equation 8 using GDP data from Bergeaud et al. (2016) and CPI data from Òscar Jordà et al. (2016) from 1900-2018. We include 10 lags in the estimation to sufficiently control for persistence in the data, yielding an effective sample size of 108 observations at horizon 1. Following Montiel Olea and Plagborg-Møller (2021), we construct bootstrap confidence intervals using their recommended wild autoregressive bootstrap with the percentile-t method. We standardize the aggregate relevance time series by dividing the logged series by the standard deviation of its first difference. Thus, the impulse responses can be interpreted as the percentage change in real GDP per capita and CPI to a one-standard-deviation innovation in the typical period-to-period change in relevance.

The results for GDP and CPI are plotted in Figure 9, showing the hallmarks of a technology shock. GDP (left panel) increases gradually, reaching a peak mean estimate of 3% after approximately 8 years. Prices, measured by CPI, fall over the entire response horizon.

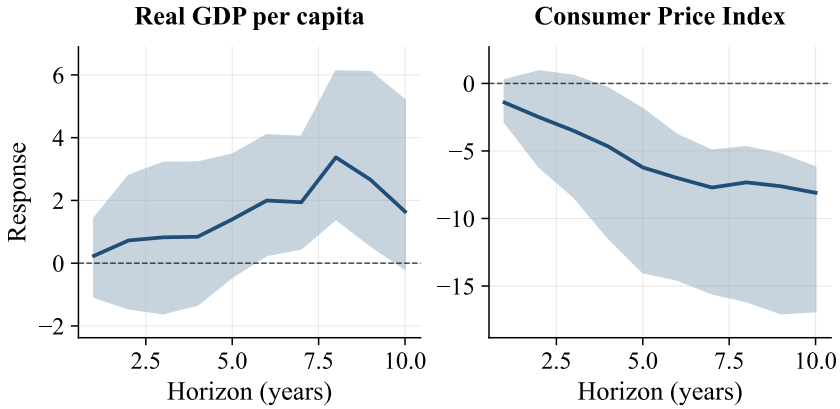


Figure 9: Impulse responses of Real GDP and CPI to Relevance shock

Note: The left (right) panel shows the impulse response of real GDP per capita (CPI) to a relevance shock. Relevance is standardized by the standard deviation of its first difference. The impulse responses are obtained from Equation 8, which is estimated using U.S. data over the sample period 1900-2018. Shaded areas represent 90% confidence intervals.

To further investigate the usefulness of relevance as a measure of technology, we also estimate the response of TFP. We also compare the performance of relevance to two prominent patent-based measures of aggregate technological progress, Value/GDP (Kogan et al. (2017)) and Breakthroughs/Capita (Kelly et al. (2021)). The model and setup are as described above; however, the sample size, including the two additional measures, spans 1920-2014. Once again, we standardize each technology indicator by dividing the logged series by the standard deviation of its first difference. Thus, the impulse responses can be

interpreted as the percentage change in TFP to a one-standard-deviation innovation in the typical period-to-period change in each indicator.

Panel A of Figure 10 shows the impulse response of TFP resulting from Equation 8. Each panel corresponds to one estimation, using relevance, value, and breakthrough, respectively, as the aggregate technology indicator. All three indicators generate increases in TFP that materialize 5 to 10 years following the shock. Relevance generates the quickest response, with significant effects appearing five years after the shock. These results indicate that relevance performs comparably to established measures, such as value and breakthrough, in capturing the latent process of technological progress.

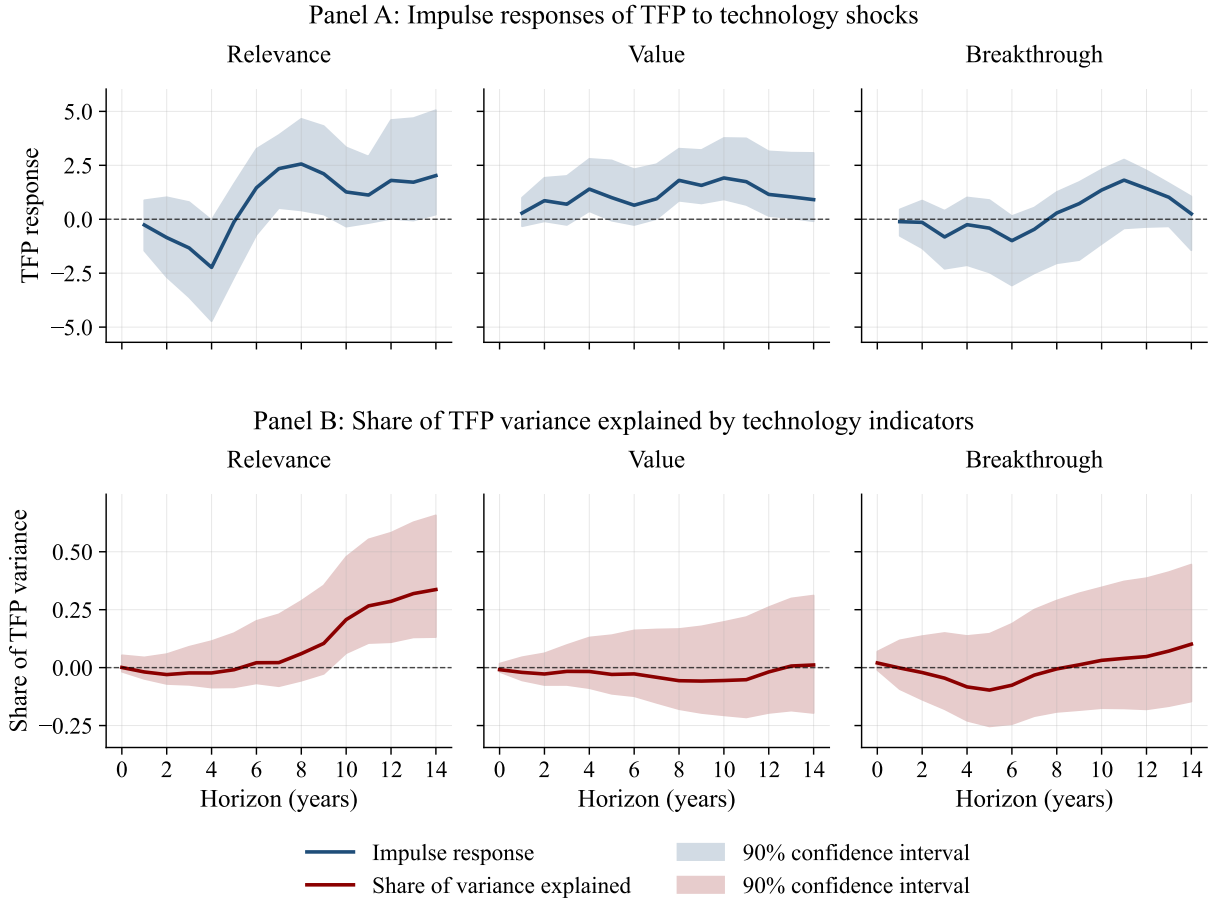


Figure 10: Impulse responses and forecast error variance decomposition, TFP

Note: Panel A of this figure shows the impulse responses of total factor productivity (TFP) to a shock to three technology indicators: relevance, value, and breakthrough. Each indicator is standardized by the standard deviation of its first difference. The impulse responses are obtained from Equation 8, which is estimated using U.S. data over the sample period 1920-2014. Shaded areas represent 90% confidence intervals. Panel B shows the share of forecast error variance of TFP explained by the same technology indicators. The forecast error variance decomposition is calculated using the framework of Gorodnichenko and Lee (2020), using data in first differences of the log-levels, with their proposed finite-sample bias correction. Shaded areas represent 90% bootstrap confidence intervals with bias correction.

To complement the impulse responses, we quantify how much of the variance in TFP each indicator explains. We do so using the R^2-based forecast error variance decomposition (FEVD) of Gorodnichenko and Lee (2020), estimated within the local projections framework.

The FEVD estimation differs from the impulse response analysis in two ways. First, we use log levels of each indicator rather than the per-period normalization applied above. Second, because the Gorodnichenko and Lee (2020) framework requires the explanatory variable to be a white-noise shock, we transform Equation 8 to first differences.[5] We then use bootstrapping to correct for finite-sample bias, following their recommended procedure. Appendix E provides the technical details and, for completeness, reports the FEVDs obtained without bias correction.

Panel B of Figure 10 shows the share of the forecast error variance of TFP explained by the technology indicator. Each panel corresponds to one estimation, using relevance, value, and breakthrough, respectively, as the explanatory variables. Among these, relevance consistently explains the largest share of variance in TFP, with positive estimates at the 90% confidence level from year 9 onward and reaching more than 20% after 10 years. In contrast, the variance share explained by Breakthrough and Value remains non-significant at all forecast horizons[6]. Importantly, since all three technology indicators likely contain measurement error, these estimates should be considered lower bounds for their true contribution to TFP, given that the estimator is downward biased in such cases (Gorodnichenko and Lee, 2020). These results suggest that task relevance captures a dimension of technological progress that is particularly informative for productivity dynamics, and may be useful for efforts to incorporate the effects of innovation into economic and budgetary projections (Gullo et al., 2025).

6.2 Forecasting productivity

The variance decomposition demonstrates that task relevance explains a significant share of TFP variation from 10 years onwards. A natural question is whether this relationship carries genuine predictive content out of sample. We address this question through a forecasting exercise. We compare forecasts of cumulative TFP growth from three models: a random-walk benchmark, a standard macroeconomic baseline, and a model augmented with task relevance.

The most parsimonious benchmark is a random walk with drift. Under this model, the optimal forecast of future TFP growth is simply the mean growth observed over the estimation window. Against this benchmark, we compare a baseline macroeconomic model comprising TFP, capital input, labor quality, and R&D investment, and an augmented specification that includes aggregate task relevance (Equation 4). By including R&D investment in the baseline, we ensure it captures the aggregate innovation effort, so that any forecasting gain from the augmented model reflects the *content* of innovation rather than its volume.

All three specifications are estimated using a rolling-window Bayesian VAR with a conjugate Minnesota-style prior. The Minnesota prior shrinks the coefficient matrices toward a random walk, with the degree of shrinkage governed by the overall tightness parameter λ and a lag decay parameter d that penalizes coefficients on more distant lags. To guard against overfitting in the larger model, we follow Bańbura et al. (2010) and scale the tightness parameter inversely with the number of variables: $\lambda_m = \bar{\lambda}/n_m^2$, where $\bar{\lambda} = 0.35$ and n_m is the number of endogenous variables in specification m. The lag decay parameter is set to $d = 1$ across all specifications, and the lag length is set to $p = 6$.

[5] This transformation alters the interpretation of the FEVD. While the impulse responses assess the long-run impact of technology shocks on the level of TFP, the FEVD quantifies the variation in TFP growth attributable to technology shocks.

[6] The negative point estimates for value and breakthrough appear because the non-bias-corrected estimates are close to zero, and after subtracting the finite-sample bias from the estimates, they turn negative.

All variables enter in first differences of $100 \times$ log levels. TFP, capital input, and labor quality are drawn from Fernald (2014); R&D investment is measured as real GDP in research and development from the Bureau of Economic Analysis.[7] Forecasts are generated by iterating the posterior mean VAR coefficients forward, and the forecast target is cumulative TFP growth from the forecast origin to horizon h. The estimation uses a rolling window of $W = 60$ quarters, and we evaluate forecasts at horizons ranging from $h = 4$ to $h = 32$ quarters (1 to 8 years). The sample spans 1947:Q1 to 2020:Q4.

To ensure the exercise reflects the information set of a real-time forecaster, task relevance is lagged by six quarters (18 months) relative to the forecast origin. This reflects the statutory delay between patent filing and publication, and means that all forecasting gains reported below are achievable using only information that would genuinely have been available at the time of the forecast.[8]

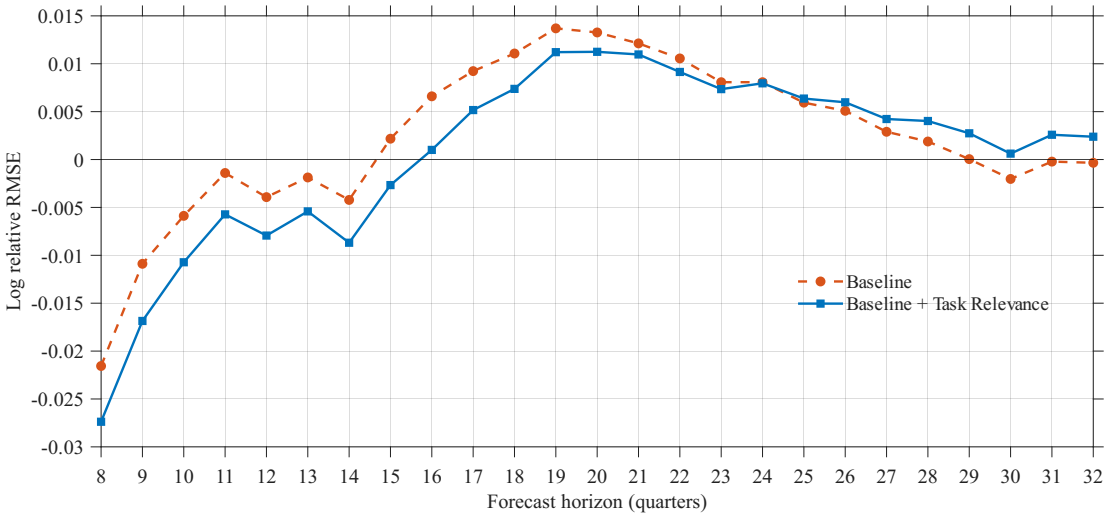


Figure 11: Out-of-sample TFP forecasting: log relative RMSE by horizon

Note: This figure plots $\log(\text{RMSE}_{\text{RW}}/\text{RMSE}_m)$ for each model m at each forecast horizon, so that positive values indicate the model outperforms the random walk benchmark. The baseline model includes TFP growth, capital input growth, R&D investment growth, and labor quality growth. The augmented model adds task relevance, lagged by 6 quarters to reflect realistic publication delays. Estimated using rolling-window Bayesian VARs with a conjugate Minnesota prior ($W = 60$ quarters, $p = 6$ lags). All variables enter in first differences of $100 \times$ log levels.

Figure 11 summarizes the out-of-sample forecasting performance; the underlying RMSE values at every fourth quarter are reported in Table 12 in Appendix H. Both the baseline and augmented models beat the random walk at horizons beyond four years, confirming that standard macroeconomic variables contain useful information for TFP forecasting. At medium-term horizons of 4 to 5 years, the baseline and augmented models perform comparably. However, the two models diverge at longer horizons: from approximately 6 years onward, the task-relevance-augmented model establishes a clear advantage over the baseline. At the 8-year horizon, the augmented model reduces cumulative forecast error by approximately 2% relative to the random walk, while the baseline model underperforms the random walk at this horizon.

This pattern is consistent with the in-sample evidence. The impulse response analysis showed that

[7] FRED series Y694RX1Q020SBEA

[8] Because our measure is computed from granted patents, and the grant date typically exceeds the 18-month publication lag, this is a conservative assumption. In practice, the measure could potentially be constructed from published applications, which become available earlier.

task relevance shocks take approximately 5 years to produce statistically significant effects on TFP, and the forecast error variance decomposition reached significance at comparable horizons. The out-of-sample results confirm that this slow-building but persistent relationship translates into forecasting gains at the horizons where it would be expected to—and where long-run economic projections are most exposed to uncertainty about the pace of technological change.

Several caveats are worth noting. First, at the longest horizons (h = 28–32 quarters), the number of non-overlapping forecast observations is limited, so the long-horizon results should be interpreted in conjunction with the in-sample evidence. However, the internal consistency between the two analyses—both pointing to effects emerging at approximately 5 years—strengthens the overall case. Second, the Minnesota prior with tightness scaled by model size provides a degree of protection against overfitting in the augmented specification, but the results could be explored further with alternative shrinkage methods or prior structures. Third, the exercise uses granted patents; as noted above, a practitioner with access to published applications could potentially construct the measure with shorter lags, which may improve forecasting performance at shorter horizons.

7 Conclusion

This paper develops a novel methodology for measuring technological progress —capturing the relevance of technology to work tasks. The main assumption underlying this method is that the semantic similarity between a patent and a task description contains information about the relevance of the patented technology in performing that task. We validate our metric by showing that it correlates strongly with established indicators of patent quality (citations, value, and importance). We demonstrate that our metric has explanatory power in both cross-sectional and aggregate applications. Furthermore, in an out-of-sample forecasting exercise, we show that adding task relevance to standard macroeconomic models improves TFP forecasts at longer horizons, suggesting the measure contains information about the trajectory of productivity that existing indicators do not capture.

References

Abramovitz, M. (1956): "Resource and Output Trends in the United States since 1870," in *Resource and Output Trends in the United States since 1870*, National Bureau of Economic Research, Inc, NBER Chapters, 1–23.

Allaoui, M., M. L. Kherfi, and A. Cheriet (2020): "Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study," in *Image and Signal Processing: 9th International Conference, ICISP 2020*, Marrakesh, Morocco, vol. 12119, 317–325.

Asyaky, M. S. and R. Mandala (2021): "Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP," in *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–6.

Autor, D., C. Chin, A. Salomons, and B. Seegmiller (2024): "New Frontiers: The Origins and Content of New Work, 1940–2018*," *The Quarterly Journal of Economics*, qjae008.

Autor, D. H. and D. Dorn (2013): "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market," *American Economic Review*, 103, 1553–97.

Autor, D. H., F. Levy, and R. J. Murnane (2003): "The Skill Content of Recent Technological Change: An Empirical Exploration," *Quarterly Journal of Economics*, 118, 1279–1333.

Bańbura, M., D. Giannone, and L. Reichlin (2010): "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25, 71–92.

Bergeaud, A., G. Cette, and R. Lecat (2016): "Productivity Trends in Advanced Countries between 1890 and 2012," *Review of Income and Wealth*, 62, 420–444.

Berkes, E., K. Manysheva, and M. Mestieri (2022): "Global Innovation Spillovers and Productivity: Evidence from 100 Years of World Patent Data," *FRB of Chicago Working Paper No. 2022-15*.

Campello, R. J., D. Moulavi, and J. Sander (2013): "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, Springer, 160–172.

Dechezleprêtre, A., D. Hemous, M. Olsen, and C. Zanella (2020): "Automating labor: evidence from firm-level patent data," CEP Discussion Papers dp1679.pdf, Centre for Economic Performance, LSE.

Deming, D. and L. B. Kahn (2018): "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals," *Journal of Labor Economics*, 36, S337–S369.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018): "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint (https://arxiv.org/abs/1810.04805)*.

Fernald, J. (2014): "A quarterly, utilization-adjusted series on total factor productivity," Federal Reserve Bank of San Francisco.

Field, A. J. (2011): *A Great Leap Forward: 1930s Depression and U.S. Economic Growth*, Yale University Press.

Flood, S., M. King, R. Rodgers, S. Ruggles, J. R. Warren, D. Backman, A. Chen, G. Cooper, S. Richards, M. Schouweiler, and M. Westberry (2024): "IPUMS CPS: Version 12.0 [dataset]," .

Gordon, R. J. (2016): *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War*, Princeton University Press, rev - revised ed.

Gorodnichenko, Y. and B. Lee (2020): "Forecast Error Variance Decompositions with Local Projections," *Journal of Business & Economic Statistics*, 38, 921–933.

Gullo, T., B. Page, D. Weiner, and H. Williams (2025): "Estimating the Economic and Budgetary Effects of Research Investments," *Tax Policy and the Economy*, 39, 53–81.

Hall, B., A. Jaffe, and M. Trajtenberg (2001): "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools," NBER Working Papers 8498, National Bureau of Economic Research, Inc.

Hall, B. H., A. Jaffe, and M. Trajtenberg (2005): "Market Value and Patent Citations," *The RAND Journal of Economics*, 36, 16–38.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2000): "Market Value and Patent Citations: A First Look," Working Paper 7741, National Bureau of Economic Research.

Hampole, M., D. Papanikolaou, L. D. Schmidt, and B. Seegmiller (2025): "Artificial Intelligence and the Labor Market," Working Paper 33509, National Bureau of Economic Research.

Jovanovic, B. and P. Rousseau (2005): "General Purpose Technologies," in *Handbook of Economic Growth*, ed. by P. Aghion and S. Durlauf, Elsevier, vol. 1, Part B, chap. 18, 1181–1224, 1 ed.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021): "Measuring Technological Innovation over the Long Run," *American Economic Review: Insights*, 3, 303–20.

Kline, P., N. Petkova, H. Williams, and O. Zidar (2019): "Who Profits from Patents? Rent-Sharing at Innovative Firms*," *The Quarterly Journal of Economics*, 134, 1343–1404.

Kogan, L., D. Papanikolaou, L. D. Schmidt, and B. Seegmiller (2023): "Technology and Labor Displacement: Evidence from Linking Patents with Worker-Level Data," Working Paper 31846, National Bureau of Economic Research.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017): "Technological Innovation, Resource Allocation, and Growth*," *The Quarterly Journal of Economics*, 132, 665–712.

Mann, K. and L. Püttmann (2023): "Benign Effects of Automation: New Evidence from Patent Texts," *The Review of Economics and Statistics*, 105, 562–579.

McInnes, L., J. Healy, and J. Melville (2020): "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," .

Montiel Olea, J. L. and M. Plagborg-Møller (2021): "Local Projection Inference Is Simpler and More Robust Than You Think," *Econometrica*, 89, 1789–1823.

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297.

Petukhova, A., J. P. Matos-Carvalho, and N. Fachada (2025): "Text clustering with large language model embeddings," *International Journal of Cognitive Computing in Engineering*, 6, 100–108.

Seegmiller, B., D. Papanikolaou, and L. D. Schmidt (2023): "Measuring document similarity with weighted averages of word embeddings," *Explorations in Economic History*, 87, 101494, methodological Advances in the Extraction and Analysis of Historical Data.

Valletta, R. G. (2018): "Recent Flattening in the Higher Education Wage Premium: Polarization, Skill Downgrading, or Both?" .

Webb, M. (2020): "The Impact of Artificial Intelligence on the Labor Market," Working Paper.

Zhang, J., Y. Huang, S. Liu, Y. Gao, and X. Hu (2025): "Do BERT-Like Bidirectional Models Still Perform Better on Text Classification in the Era of LLMs?" .

Òscar Jordà, M. Schularick, and A. M. Taylor (2016): *Macrofinancial History and the New Business Cycle Facts*, University of Chicago Press, 213–263.

A Clustering task descriptions

This appendix describes the methodology for constructing the 145 task clusters used in the main analysis. The procedure involves three steps: embedding task descriptions, reducing dimensionality, and clustering.

Word embeddings build on the core principle that semantically similar texts have similar vector representations. Since Devlin et al. (2018) introduced the Bidirectional Encoder Representations from Transformers (BERT), there has been significant interest in transformer-based models, underpinning many recent advances in large language models (LLMs). Despite recent advances, the BERT model continues to perform well on shorter text excerpts (Zhang et al., 2025), which makes it suitable for this analysis. Using BERT ("bert-base-uncased"), we construct a representation for each task description by first creating an embedding vector for each word in the description and averaging word vectors to represent the task.[9] We exclude 'stopwords'—words with little semantic content ('the', 'is', 'at', and so on)—from the final embeddings. This process yields a high-dimensional vector for each task description in our data.

Next, we perform a dimensionality reduction on the embedding vectors that reduces their dimension from 768 to 10. Dimensionality reduction is increasingly used in clustering applications, as it improves performance compared to using the raw data (Allaoui et al., 2020). We use the Uniform Manifold Approximation and Projection (UMAP) by McInnes et al. (2020) for dimensionality reduction, as this method performs well on word embeddings (Asyaky and Mandala, 2021).

Finally, we run the clustering algorithm on the UMAP-reduced vectors. For this task, we use Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) by Campello et al. (2013), which we select for two main advantages: it automatically selects the optimal number of clusters, and it labels observations as noise if they do not fit into any identified cluster.

While the clustering approach involves several parameter choices, our objective is not to identify the 'one true' categorization of work tasks, but rather to obtain a representative and compact characterization of the task space. We select parameters to balance two competing objectives: capturing sufficient detail to distinguish different types of tasks, while maintaining a manageable number of clusters for empirical implementation. As noted, for UMAP, we reduce the embedding vectors to 10 dimensions to preserve most of the semantic structure while maintaining stable clustering results. For HDBSCAN, we set the minimum cluster size to 45 to ensure each cluster represents economically meaningful work activities that span multiple occupations.

The clustering exercise yields a silhouette score[10] of 0.42, indicating reasonably distinct clusters (values between 0.3–0.7 are typically acceptable for large language model embeddings; see Petukhova et al., 2025). Rather than indicating poor clustering performance, this score reflects the inherent overlap between work activities in the economy, where activities often share common elements.

[9] For instance, given the task 'Direct financial operations', the model first constructs an embedding for each word based on the word itself and the context in which it appears, then computes the final embedding by averaging these individual embeddings. Our task embedding method is an unweighted version of one proposed by Seegmiller et al. (2023).

[10] The silhouette score measures how well separated the clusters are. The silhouette ranges from -1 to 1, with high values indicating that the clusters identify more distinct kinds of work.

A.1 Parameter selection and cluster stability

We use the Adjusted Rand Index (ARI) to assess the stability of our clustering results. The ARI measures the similarity between two data groupings, adjusted for chance agreement. The ARI between two groupings is defined as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}} \tag{9}$$

where n_{ij} is the number of objects in both cluster i of the first grouping and cluster j of the second grouping, a_i is the number of objects in cluster i of the first grouping, b_j is the number of objects in cluster j of the second grouping, and n is the total number of objects. ARI scores range from -1 to 1, where 1 indicates identical clusterings and 0 indicates agreement no better than random chance.

To assess clustering stability, we vary the two parameters with the greatest impact on results: UMAP dimensions and HDBSCAN minimum cluster size. We hold other algorithm parameters at standard values: number of neighbors used for UMAP graph construction (UMAP neighbors = 12) and the minimum number of neighbors required for a point to be considered a core point (HDBSCAN minimum neighborhood size = 5). Specifically, we test UMAP dimensions of $[4, 6, 8, 10, 12]$ components while holding minimum cluster size at 45, then test HDBSCAN minimum cluster sizes of $[30, 35, 40, 45, 50]$ while holding UMAP dimensions fixed at 10. Figure 12 reports the pairwise ARI scores obtained by varying UMAP dimensions, while Figure 13 reports those obtained when varying minimum cluster size.

Figure 12 shows that using 6, 8, or 10 UMAP components yields stable clustering results, with pairwise ARI scores of 0.70, 0.71, and 0.75. In contrast, 4 and 12 components produce substantially different clusters, with pairwise ARI scores of between -0.01 and 0.00 when compared to clusters with 6, 8, or 10 components. Given stability across 6-10 components, we select 10 components to retain more information from the original word embeddings when constructing clusters.

Figure 13 shows that minimum HDBSCAN cluster sizes between 30 and 45 yield stable clustering results, with pairwise ARI scores ranging from 0.84 to 0.95, while the clusters resulting from a minimum cluster size of 50 are completely unrelated (pairwise ARI scores of 0.00). We select a minimum cluster size of 45 to capture only the most common work tasks.

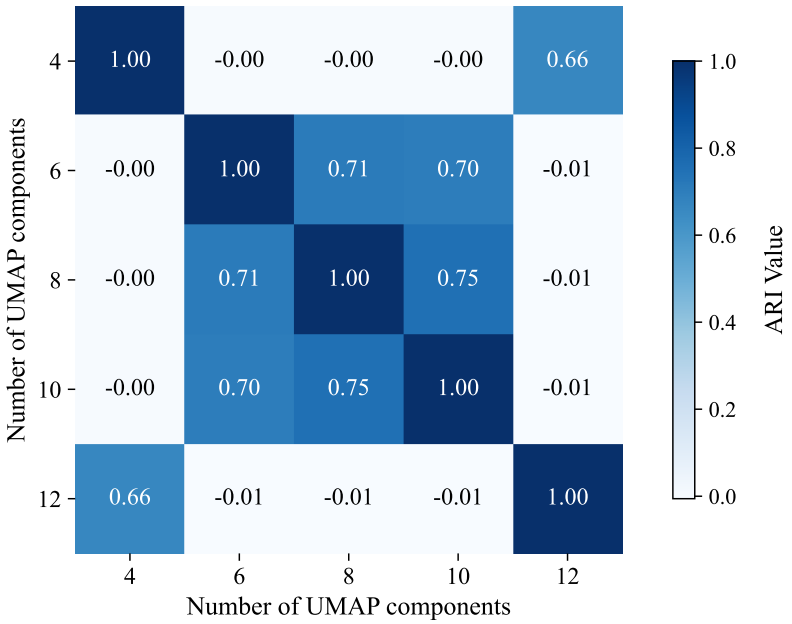


Figure 12: Pairwise Adjusted Rand Index, varying UMAP dimensions

Note: This figure shows the pairwise Adjusted Rand Index (ARI) scores obtained when varying the UMAP dimensions $[4, 6, 8, 10, 12]$. The other algorithm parameters are kept at UMAP neighbors = 12, HDBSCAN minimum neighborhood size = 5, and HDBSCAN minimum cluster size = 45.

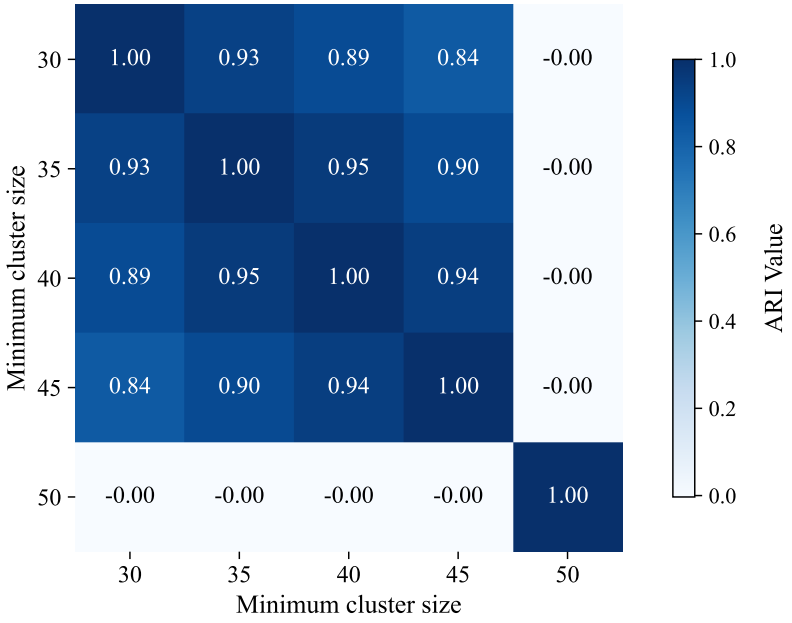


Figure 13: Pairwise Adjusted Rand Index, varying HDBSCAN minimum cluster size

Note: This figure shows the pairwise Adjusted Rand Index (ARI) scores obtained when varying the HDBSCAN minimum cluster size of $[30, 35, 40, 45, 50]$. The other algorithm parameters are kept at UMAP dimensions = 10, UMAP neighbors = 12, and HDBSCAN minimum neighborhood size = 5.

A.2 Overview of task clusters

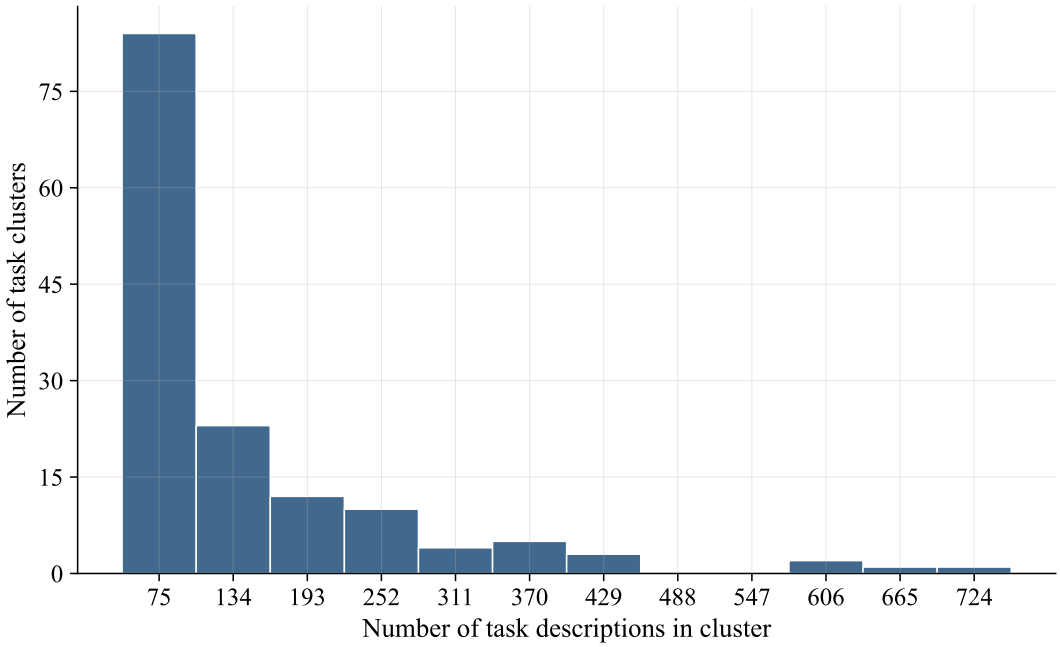


Figure 14: Histogram of task descriptions per cluster

Note: This figure shows a histogram of the number of task descriptions assigned to each cluster using the clustering approach described in Section 3.1. The clustering is done using the UMAP algorithm of dimensionality reduction and HDBSCAN for clustering. In total, there are 145 task clusters.

B Task characteristics

Table 2: Task category definitions

Category	Definition
Cognitive	'Cognitive work encompasses tasks that demand mental processing, including problem-solving, decision-making, and creative thinking, often requiring the application of acquired knowledge, critical analysis, and innovative approaches'
Manual	'Manual work encompasses the utilization of physical labor, dexterity, and skill to accomplish tasks, often involving tangible materials and hands-on interaction'
Routine	'Routine work refers to the performance of repetitive tasks that follow established procedures, requiring minimal adaptation'
Nonroutine	'Nonroutine work involves tasks that demand adaptability, critical thinking, and creative problem-solving, as they often present complex or unpredictable challenges'
Collaborative	'Collaborative work involves groups pooling their diverse skills, perspectives, and resources to cooperatively achieve a shared goal through teamwork and cooperation'
Independent	'Independent work refers to the process of an individual autonomously performing tasks and making decisions, relying on their skills, knowledge, and initiative to accomplish goals'

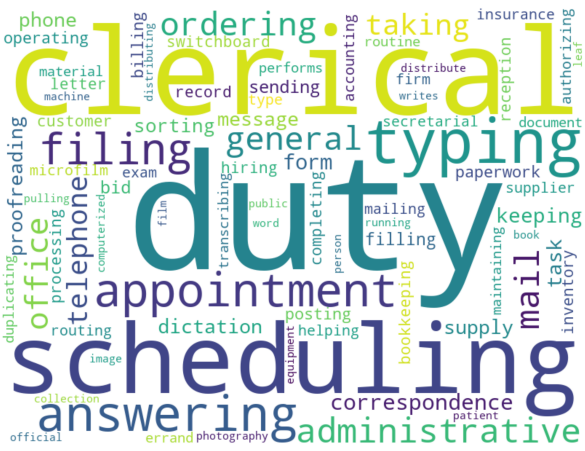
Note: This table reports the definitions for each category of our task framework (Section 3.1), which builds on the categorizations of Autor et al. (2003) and Deming and Kahn (2018). We quantify the extent to which each task is associated with these categories by computing the cosine similarity between individual task embeddings and the embedding vectors for each category definition. See Section 3.1 for details.

Table 3: Correlation of task characteristics

	Manual (C^{mw})	Routine (C^{rw})	Collaborative (C^{cw})
Manual (C^{mw})	1.00		
Routine (C^{rw})	0.41	1.00	
Collaborative (C^{cw})	−0.24	−0.76	1.00

Note: This table shows the correlations among the weighted task characteristics scores of the framework proposed in Section 3.1. The framework categorizes work as either cognitive or manual, routine or nonroutine, and collaborative or independent. These dimensions are quantified by the task scores $C_j^{mw} = C_j^{man}/(C_j^{man} + C_j^{cog})$, $C_j^{rw} = C_j^{rout}/(C_j^{rout} + C_j^{nonr})$, and $C_j^{cw} = C_j^{col}/(C_j^{col} + C_j^{indep})$, where C_j^{man}, C_j^{cog}, C_j^{rout}, C_j^{nonr}, C_j^{col}, and C_j^{indep} represent the task scores for the manual, cognitive, routine, nonroutine, collaborative, and independent categories of task j, respectively.

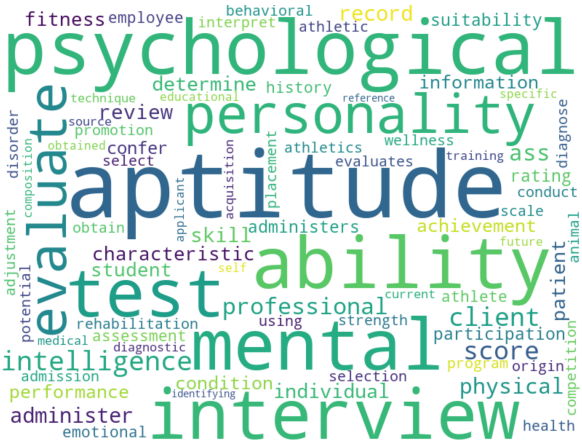
Table 3 presents the correlations among the weighted task characteristic scores across all tasks. Several patterns are apparent. Manual work tends to be more routine, with the two dimensions showing a correlation of 0.41. Manual tasks are also more likely to be independent rather than collaborative, with a negative correlation of -0.24. Lastly, collaborative tasks are predominantly nonroutine, showing a correlation of -0.75.



(a) Most cognitive



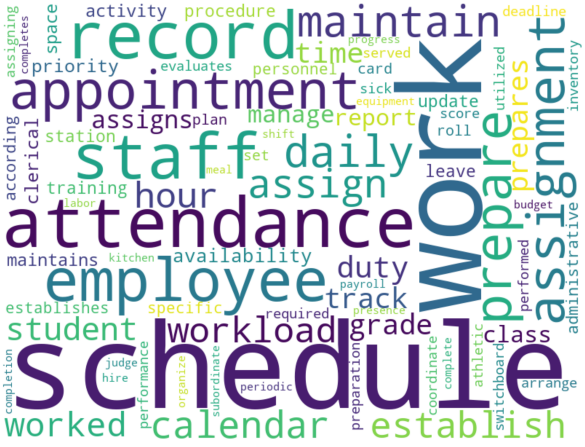
(b) Most manual



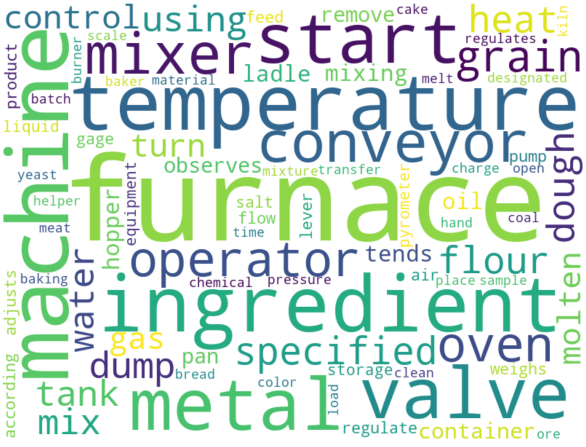
(c) Most nonroutine



(d) Most routine



(e) Most collaborative



(f) Most independent

Figure 15: Word clouds for tasks at the extremes of each dimension in the task framework

Note: This figure shows word clouds for the tasks with the highest and lowest scores on each of the three dimensions (manual–cognitive, routine–nonroutine, collaborative–independent). Word clouds are constructed by applying TF-IDF weighting to each task, where each task is represented by the concatenation of all underlying task descriptions.

B.1 Comparison with existing classifications

Table 4 reports the rank correlations between our task-based measures and the Autor and Dorn (2013) task input scores. To do so, we convert our task-level scores to occupation-level scores by assigning each task description the score of its parent task cluster and averaging across all task descriptions in each occupation. Let C_o^{dw} denote the occupation-level score for dimension $d \in \{\text{Manual, Routine, Collaborative}\}$, obtained by averaging task-level scores C_j^{dw} across all tasks j in occupation o. The primary pre-specified comparisons are along the diagonal (Manual–Manual, Routine-–Routine, Collaborative—Abstract); off-diagonal entries are reported for completeness.

The modest magnitude of these correlations is attributable to differences in construct definition. Autor and Dorn (2013) construct measures from a narrow set of DOT variables, whereas our measures capture a broader range of task content. In particular, their 'Routine' measure relies on 'finger dexterity' and 'setting limits, tolerances, and standards,' effectively targeting manufacturing and clerical tasks susceptible to early computerization. In contrast, our 'Routine' scores capture repetitiveness across all domains, including interpersonal and gross-motor tasks. Our 'Manual' scores encompass all kinds of physical labor, including hands-on interaction, while the benchmark measure reflects demand for 'eye–hand coordination.' Our 'Collaborative' scores encompass cooperation and coordination, which are distinct from the hierarchical 'Direction, Control, and Planning' variable used in the benchmark. Thus, our measure complements the canonical index. It captures work characteristics at the task- rather than the occupation level, and it identifies variation that falls outside the strict definitions of technological substitutability.

Table 4: Occupation-level correlation of task characteristics

	Autor–Dorn (2013) task input measures		
Task-based measures	Manual	Routine	Abstract
Panel A: Occupation-level scores			
Manual (C_o^{mw})	0.21	0.03	−0.40
	(0.000)	(0.541)	(0.000)
Routine (C_o^{rw})	0.17	0.03	−0.35
	(0.000)	(0.575)	(0.000)
Collaborative (C_o^{cw})	−0.15	0.08	0.27
	(0.001)	(0.073)	(0.000)
Panel B: Quartile-based scores			
Manual (C_o^{mw})	0.23	0.06	−0.39
	(0.000)	(0.202)	(0.000)
Routine (C_o^{rw})	0.17	0.05	−0.35
	(0.000)	(0.243)	(0.000)
Collaborative (C_o^{cw})	−0.15	0.05	0.28
	(0.001)	(0.272)	(0.000)

Note: This table reports Spearman rank correlation coefficients between occupation-level versions of our task-based measures (rows) and the task input measures of Autor and Dorn (2013) (columns). Panel A uses occupation-level scores obtained by assigning each task description the score of its parent task cluster and averaging across all task descriptions in each occupation. Panel B assigns occupations to quartiles based on each measure and computes rank correlations using quartile ranks (1–4). Two-sided p-values are shown in parentheses.

C Task labor outcomes

Table 5: Summary statistics per year, task outcomes

Year	Total hours ($\times 10^{10}$)	Avg. hourly wage (real, YYYY $)	Number of tasks
1940	2.928	0.169	38
1950	2.498	0.989	62
1960	4.686	1.315	48
1970	10.453	1.458	57
1980	13.708	1.420	54
1990	18.689	1.504	144
2000	23.112	1.871	128
2003	22.896	1.943	132
2010	23.899	1.878	133
2019	28.169	2.096	133
2020	27.160	2.642	133

Note: This table reports total hours worked, average hourly wages, and the number of active tasks for each observation year in the sample. Total hours are the sum of hours worked across all tasks in each year, expressed in units of 10^{10} hours. The average hourly wage is task-level hourly wages. Observation years correspond to major DOT/O*NET task-classification updates. The underlying labor data are drawn from the U.S. Census and the CPS, and cover the period 1940-2020.

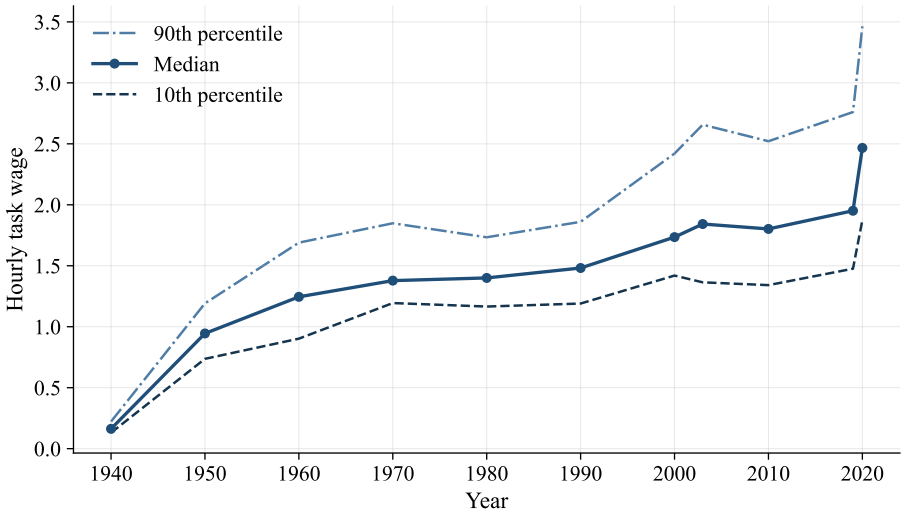


Figure 16: Distribution of task wages

Note: This figure shows the median hourly wage per task each sample year from 1940 to 2020, together with the 10th and 90th percentile of the task wage distribution. The underlying wage data are drawn from the U.S. Census and the CPS.

The wage data for 1940 and 2020 appear to be anomalies. In 1940, the level and dispersion of task wages appear significantly lower than in subsequent years in ways that are not consistent with aggregate wage series, likely reflecting limitations of early microdata. In 2020, wages exhibit a sharp jump relative to 2019, coinciding with well-documented measurement issues and compositional changes in labor force

participation during the COVID-19 pandemic.

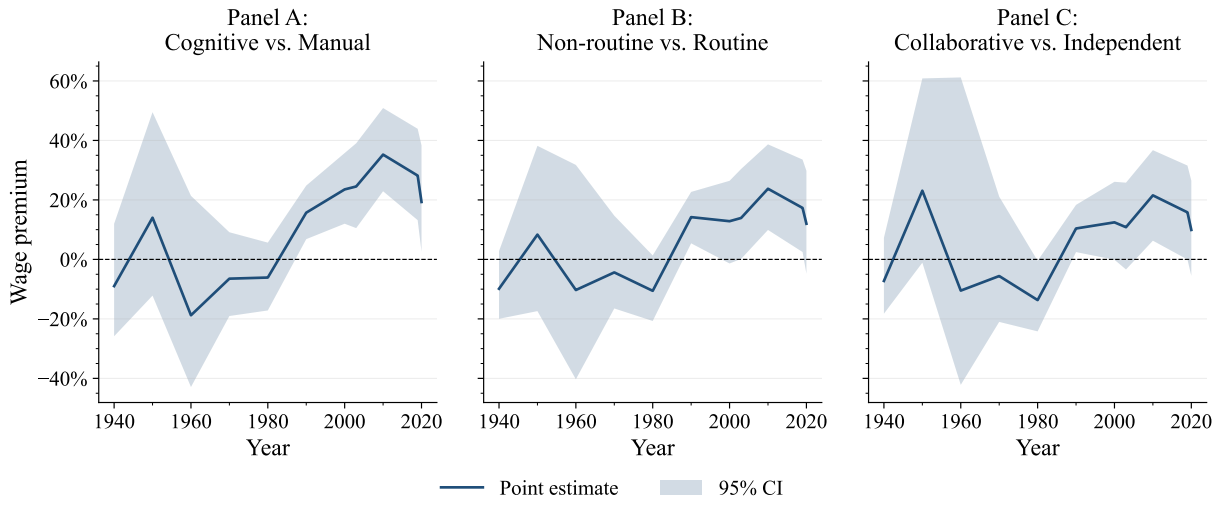


Figure 17: Wage premiums

Note: This figure shows the percentage wage premium for cognitive (relative to manual), nonroutine (relative to routine), and collaborative (relative to independent) tasks over time. Wage premiums are estimated as the difference in mean log wages between each pair of categories (e.g., cognitive vs. manual) in each year. We construct 95% confidence intervals using a stratified nonparametric bootstrap with 1,000 replications. The underlying wage data are drawn from the U.S. Census and the CPS and cover the period 1940 to 2020.

D Decomposition of covariance change by technology field

Table 6: Decomposition of covariance change by technology field

	1905–1930			1980–2010		
Technology Field	**Reallocation**	**Quality**	**Total**	**Reallocation**	**Quality**	**Total**
Building	0.058	0.037	0.095	0.004	-0.046	-0.042
Chemistry	0.101	0.067	0.169	0.054	0.134	0.187
Electricity	0.059	-0.007	0.051	0.032	0.030	0.061
Engineering in general	0.037	-0.055	-0.018	0.010	-0.032	-0.023
Engines or pumps	-0.002	-0.026	-0.028	0.003	-0.026	-0.023
Food	0.010	-0.126	-0.116	0.021	-0.054	-0.033
Health	0.002	-0.125	-0.123	-0.041	0.133	0.092
Information technology	-0.005	-0.034	-0.039	0.161	0.065	0.226
Instruments	-0.003	-0.097	-0.101	0.008	0.030	0.038
Lighting; heating	0.004	0.018	0.022	-0.013	0.001	-0.012
Mining	0.003	-0.212	-0.208	-0.004	-0.029	-0.033
Personal or domestic articles	0.020	0.177	0.197	0.009	-0.047	-0.038
Separating; mixing	-0.012	-0.089	-0.102	-0.018	-0.012	-0.030
Shaping	-0.005	0.378	0.373	-0.029	0.022	-0.008
Textiles or flexible materials	0.013	-0.166	-0.153	-0.019	-0.006	-0.025
Transporting	0.031	0.833	0.864	0.000	0.040	0.040
Weapons; blasting	0.004	-0.076	-0.072	0.003	-0.085	-0.082

Note: This table presents the field-specific contributions to the change in the covariance term of the Olley-Pakes decomposition (see Equation 7). The 'Reallocation' column corresponds to $(r_g - \bar{r})\Delta s_g$, and the 'Quality' column corresponds to $(s_g - \bar{s})\Delta r_g$. The 'Total' column is the sum of these two terms (approximating the total contribution $\Delta\omega_g$).

E Forecast error variance decomposition with local projections

Using Equation 8 as our starting point, the differenced specification used in the Gorodnichenko and Lee (2020), henceforth GL, framework can be expressed as:

$$y_{t+h} - y_{t-1} = \alpha_h + \sum_{l=1}^{p} \beta_l^h z_{t-l} + \sum_{l=1}^{p} \gamma_l^h \Delta y_{t-l} + f_{t+h|t-1} \tag{10}$$

where $f_{t+h|t-1}$ is the forecast error for the h-period ahead value of the endogenous variable. Here, $z_t = \Delta x_t$ represents the identified structural shock series.

The framework defines the forecast error variance decomposition within the local projections setting as:

$$s_h \equiv \frac{var(\sum_{k=0}^{h} \psi_{h-k} z_{t+k})}{var(f_{t+h|t-1})} \tag{11}$$

where $f_{t+h|t-1} = \sum_{k=0}^{h} \psi_{h-k} z_{t+k} + v_{t+h|t-1}$ decompose the forecast errors due to innovations in z and other sources of variation $v_{t+h|t-1}$. This quantity can be understood as an R^2 of the population projection of $f_{t+h|t-1}$ on the relevant shock series.

GL's preferred estimator for the FEVD, denoted as $\hat{s}_h$, is obtained by following two main steps:

1. Estimate the forecast errors $\hat{f}_{t+h|t-1}$ as residuals from a local projection regression in Equation 10.
2. Regressed the estimated forecast errors on the identified shock series: $\hat{f}_{t+h|t-1} = \alpha_{z,0} z_{t+h} + \cdots + \alpha_{z,h} z_t + \tilde{v}_{t+h|t-1}$. The estimator $\hat{s}_h$ is the R^2 from this second regression.

While the R^2 estimator is asymptotically unbiased, it can exhibit substantial finite-sample biases. To address this and improve inference, GL propose a bias-correction step implemented using a VAR-based bootstrap. This procedure involves simulating artificial time series from an estimated VAR model, re-estimating the FEVDs on these simulated series, and using the distribution of these bootstrap estimates to correct the original point estimate.

For the construction of confidence intervals, GL recommend using bootstrap methods over analytical expressions for asymptotic variance, due to their better finite-sample performance and ability to produce smoother confidence bands. Confidence intervals are constructed using the quantiles of the bias-corrected bootstrap distribution, which aims to center the interval around the bias-corrected FEVD estimate while accounting for the shape of the distribution.

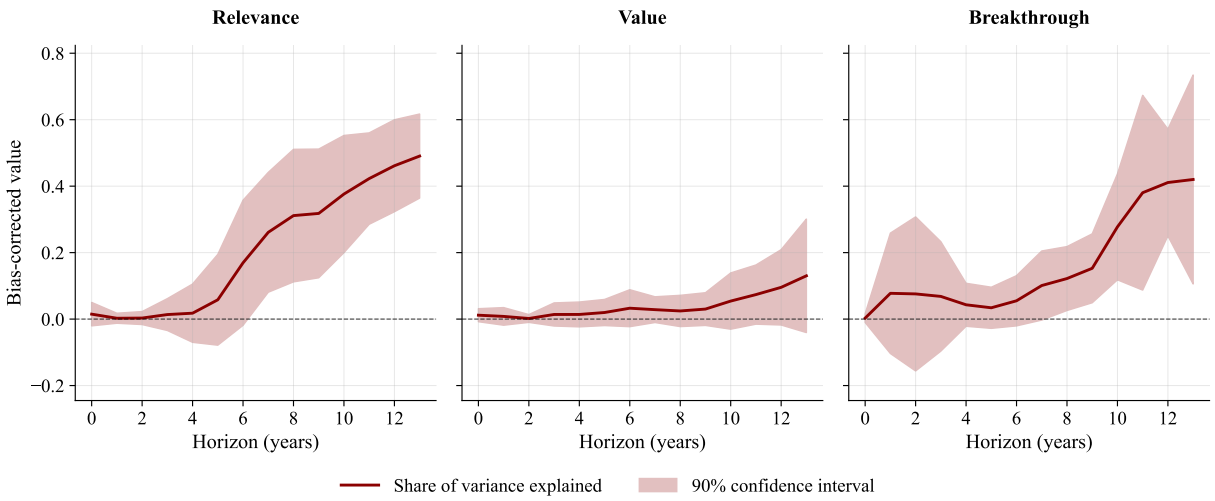


Figure 18: Non-bias-corrected share of TFP variance explained by technology indicators

Note: This figure shows the share of forecast error variance of TFP explained by three technology indicators: Relevance, Value, and Breakthrough. The forecast error variance decomposition is calculated using the framework of Gorodnichenko and Lee (2020), without their proposed finite-sample bias correction. The estimates use U.S. data over the sample period 1924-2014. Shaded areas represent 90% confidence intervals based on asymptotic standard errors with pre-whitening of the non-bias-corrected estimates.

F Validation: regression tables

Table 7: Validation – Log Forward Citations

	(1)	(2)	(3)	(4)
Relevance	0.0198***			
	(0.000387)			
Manual Relevance		-0.0526***		
		(0.000450)		
Routine Relevance			-0.0549***	
			(0.000448)	
Collaborative Relevance				0.0409***
				(0.000418)
Observations	7199571	7199571	7199571	7199571
R-squared	0.342	0.344	0.344	0.343
Year × Technology FE	Yes	Yes	Yes	Yes

RHS variables standardized by filing year. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Validation – Importance

	(1) lqsim05	(2) lqsim05	(3) lqsim05	(4) lqsim05
Relevance	0.00344***			
	(0.0000512)			
Manual Relevance		-0.0234***		
		(0.0000582)		
Routine Relevance			-0.0238***	
			(0.0000580)	
Collaborative Relevance				0.0165***
				(0.0000544)
Observations	6398712	6398712	6398712	6398712
R-squared	0.658	0.667	0.667	0.663
Year × Technology FE	Yes	Yes	Yes	Yes

RHS variables standardized by filing year. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Validation – Breakthrough LPM

	(1)	(2)	(3)	(4)
Task Relevance	0.00512***			
	(0.000149)			
Manual Relevance		-0.0303***		
		(0.000116)		
Routine Relevance			-0.0313***	
			(0.000117)	
Collaborative Relevance				0.0229***
				(0.000114)
Observations	7114302	7114302	7114302	7114302
R-squared	0.227	0.234	0.235	0.231
Year × Technology FE	Yes	Yes	Yes	Yes

RHS variables standardized by filing year. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Validation – Log Value

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Relevance	0.0148***				0.000202			
	(0.00138)				(0.000608)			
Manual Relevance		-0.0982***				0.00893***		
		(0.00155)				(0.000713)		
Routine Relevance			-0.102***				0.00825***	
			(0.00154)				(0.000708)	
Collaborative Relevance				0.133***				-0.00942***
				(0.00148)				(0.000668)
Observations	2344236	2344236	2344236	2344236	2342872	2342872	2342872	2342872
R-squared	0.127	0.128	0.128	0.130	0.836	0.836	0.836	0.836
Year × Technology FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	No	No	No	Yes	Yes	Yes	Yes

RHS variables standardized by filing year. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

G Task relevance, wages, and hours in the cross-section

We estimate the relationship between task relevance, wages, and hours worked using panel regressions with year and, in some specifications, task-cluster fixed effects. Relevance is defined as the mean relevance of a given task across all patents in a given year, standardized for comparability. To capture heterogeneity by task type, we construct three indicators, following the methodology described in Section 3.1: the Manual dummy equals one if a task's manual-intensity score is above the sample median; the Routine and Collaborative dummies are defined analogously. We interact these dummies with Relevance to test whether the link between task relevance and outcomes varies systematically across task types. Results are shown in Table 11.

Wages. Panel A of Table 11 shows the results of wage regressions. Across tasks (column 1), there is no significant unconditional association between relevance and wages. Once interactions are introduced, however, strong heterogeneity emerges. Manual (column 2) and routine (column 3) tasks exhibit negative interaction effects, indicating that increases in task relevance are associated with lower wage returns relative to other tasks. By contrast, collaborative tasks (column 4) display a positive and significant interaction, suggesting that wages in collaboration-intensive tasks respond more strongly to task relevance. In the within-task specification (column 5), controlling for task-cluster and year effects, the coefficient on Relevance is large and highly significant, indicating that a one-standard-deviation increase in relevance corresponds to a substantial wage gain. The within-task R^2 is much higher than in the across-task regressions, underscoring that variation in wages is best explained by comparing relevance within tasks.

Hours. Panel B of 11 shows the results of hours regressions. Relevance is positively and strongly correlated with hours, both across and within tasks, in all specifications except when manual task controls are included. Hours and relevance are both negatively correlated with manual scores, which explains a part of the positive cross-task association between hours and relevance.

Table 11: Regression Results

Panel A. Dependent variable: log(wage)					
	(1)	(2)	(3)	(4)	(5)
Relevance	-0.007	0.037**	0.042***	-0.039***	0.399***
	[0.010]	[0.015]	[0.015]	[0.012]	[0.047]
Manual (0/1)		-0.117***			
		[0.017]			
Manual (0/1) × Relevance		-0.080***			
		[0.018]			
Routine (0/1)			-0.128***		
			[0.017]		
Routine (0/1) × Relevance			-0.091***		
			[0.018]		
Collaborative (0/1)				0.101***	
				[0.017]	
Collaborative (0/1) × Relevance				0.071***	
				[0.018]	
R^2	0.208	0.279	0.295	0.262	0.810
Within R^2	0.001	0.091	0.110	0.069	0.114
Observations	724	724	724	724	722
Fixed effects	Year	Year	Year	Year	Task & Year
Panel B. Dependent variable: log(hours)					
	(1)	(2)	(3)	(4)	(5)
Relevance	0.242***	0.137	0.173*	0.203***	1.329***
	[0.060]	[0.097]	[0.097]	[0.072]	[0.380]
Manual (0/1)		-0.437***			
		[0.106]			
Manual (0/1) × Relevance		0.104			
		[0.113]			
Routine (0/1)			-0.262**		
			[0.107]		
Routine (0/1) × Relevance			0.068		
			[0.114]		
Collaborative (0/1)				0.383***	
				[0.107]	
Collaborative (0/1) × Relevance				0.037	
				[0.113]	
R^2	0.028	0.051	0.036	0.046	0.573
Within R^2	0.022	0.046	0.031	0.040	0.021
Observations	724	724	724	724	722
Fixed effects	Year	Year	Year	Year	Task & Year

Notes: Robust standard errors in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

H Forecasting TFP

This appendix reports the root mean squared error (RMSE) from the out-of-sample forecasting exercise described in Section 6.2. Table 12 presents the RMSE for each model at forecast horizons corresponding to every fourth quarter.

Table 12: Root Mean Squared Error by forecast horizon

Horizon (quarters)	Random walk	Baseline	Augmented
4	1.619	1.683	1.695
8	2.406	2.459	2.473
12	3.029	3.041	3.054
16	3.684	3.660	3.681
20	4.456	4.398	4.406
24	5.288	5.246	5.246
28	6.072	6.061	6.048
32	6.855	6.858	6.839

Notes: Root mean squared error (RMSE) from rolling-window out-of-sample forecasts of cumulative TFP growth. The random walk forecasts cumulative growth using the mean observed over the estimation window. The baseline model includes TFP growth, capital input growth, R&D investment growth, and labor quality growth. The augmented model adds aggregate task relevance, lagged by 6 quarters to reflect patent publication delays. Both the baseline and augmented models are estimated using Bayesian VARs with a conjugate Minnesota prior ($W = 60$ quarters, $p = 6$ lags, $\bar{\lambda} = 0.35$ scaled by model size). All variables enter in first differences of $100 \times$ log levels. Sample: 1947:Q1–2020:Q4.