The Impact of Artificial Intelligence on Inflation: Empirical Measurement and General-Equilibrium Dynamics in a Multi-Channel Framework

Salem Abo-Zaid*
University of Maryland Baltimore County

May 25, 2026

Abstract

This paper constructs a quarterly U.S. Artificial Intelligence (AI)-intensity index and uses it to quantify the contribution of AI to inflation. The broad AI index combines a set of measures that capture AI-related patents, capital services, investment and data-center activity. Across specifications, the results indicate that AI has made a positive, economically meaningful contribution to U.S. inflation, on the order of 0.2 annualized percentage points, between 2007 and 2025. The impact of AI on inflation has been rising in recent years. It is also shown that sectors with higher exposure to AI tend to have higher inflation rates. To interpret these findings, the paper develops a general equilibrium model in which AI influences inflation through multiple simultaneously active mechanisms. AI raises productivity and lowers marginal costs, exerting downward pressure on inflation. Conversely, AI may increase effective markups in product and labor markets, raise reference costs in AI-complementary activities, and stimulate investment, generating inflationary forces. The impact of AI on inflation, thus, is not pinned down by productivity alone, but also by the relative strength of market-power, cost channels and the extent to which AI propagates through demand and input-price spillovers. AI also affects the *long-run* inflation rate, which could rise by 1 percentage point, and possibly more, over the course of the next 10 years.

JEL classification: E31; E32; E47; E17; E52
Keywords: Inflation, Artificial Intelligence, Market Power, Investment, Data Centers, Patents

*University of Maryland-Baltimore County, Department of Economics, 1000 Hilltop Circle, Baltimore, MD 21250. Email: salem.abozaid@umbc.edu, Tel. +1 410-455-2498.

1 Introduction

The impact of artificial intelligence (AI) on inflation has become a central question for macroeconomic analysis and for monetary policy. A natural starting point is the familiar view of AI as a productivity-enhancing general-purpose technology: by improving efficiency, compressing unit costs, and expanding effective capacity, AI can act as a structural disinflationary force. At the same time, the macroeconomic footprint of AI is not confined to productivity. Scaling and deploying AI requires costly complementary inputs and infrastructure, can reshape product-market pricing and market power, and can alter labor-market rents and wage-setting wedges. Moreover, the diffusion of AI has increasingly taken the form of an investment and infrastructure wave—most visibly through rapid expansion of compute capacity and data-center build-out—that operates as a demand impulse as well as a supply-side transformation. The sign, magnitude, and persistence of AI's effect on inflation therefore cannot be inferred from a single channel or a single dataset; they are joint empirical and quantitative objects that depend on how these mechanisms interact in general equilibrium.

The contribution of this paper is threefold. First, it constructs a quarterly U.S. proxy for AI intensity that is designed for macroeconomic use and for robustness across measurement choices. Because AI is not directly observed as a single national-accounts series, the approach triangulates AI activity using multiple observables that speak to distinct margins of AI development and diffusion. The measurement system emphasizes AI-oriented capital formation and intangible investment, the accumulated services of AI-relevant capital, indicators of enabling infrastructure associated with large-scale deployment, and a proxy for innovative activity related to AI. These components are aligned to a common quarterly calendar and combined into composite summaries that allow the analysis to distinguish sustained accumulation from episodic innovation and late-sample infrastructure acceleration. The purpose is not to claim that any one indicator is a sufficient statistic for AI, but to provide a transparent and replicable mapping from observable time series to a latent "AI intensity" state that can be used across empirical and structural exercises, while also permitting targeted exclusions that clarify which empirical results are sensitive to investment, infrastructure, or slower-moving capacity measures.

Second, the paper constructs descriptive time series for the contribution of AI to inflation. For each AI proxy, quarterly movements in the AI measure are related to quarterly inflation outcomes under a set of alternative specification environments that address key features of inflation data, including persistence and common confounders. The fitted AI component from these projections is then translated into a quarterly "AI contribution" series in the same units as the macro outcome, allowing a direct comparison between realized inflation dynamics and the portion that co-moves with observed AI intensity under the maintained specification. This exercise is implemented for multiple inflation measures to reflect the fact that different price indices embed different exposures to imports, commodities, and basket composition.

The main empirical message that emerges is that the implied contribution of AI to inflation is positive on average and becomes more pronounced in the later part of the sample, with the late-sample increase aligning with the period in which enabling infrastructure and deployment-oriented indicators accelerate sharply. This pattern is especially salient when the AI proxy is constructed to capture the scaling of compute-intensive capacity, consistent with a macro environment in which AI diffusion is increasingly tied to infrastructure expansion and associated demand pressures rather than only to gradual cost-reducing efficiency gains. For the sample period 2007Q1-2025Q3, the average contribution of AI to inflation in the benchmark analysis is nearly 0.2 annualized percentage points when the broad composite of AI is considered, and rises to close to 1 annualized percentage point when a measure of AI capital services is used. Investment in AI contributes greatly to inflation as well, particularly when investment in data centers is accounted for. Alternative estimates point to even stronger impact of AI measures on inflation, which further cements the conclusions of the paper. Furthermore, the estimates suggest that across measures of AI, the contribution tends to be positive and significantly different from zero. The only exception is when AI patents, which in theory could be productivity-enhancing, are considered: the impact of AI-related patents on inflation is not statistically different from zero.

The paper also provides sector-level evidence on the inflationary effects of AI adoption. Using producer-price inflation at a detailed industry level, it asks whether sectors that were more exposed ex ante to AI-relevant task content experienced systematically different inflation dynamics after the generative-AI acceleration associated with late 2022. The analysis exploits cross-sector variation in occupational task exposure, distinguishing industries whose production structure is more intensive in analytical, information-processing, and reasoning tasks from less exposed sectors. Across complementary empirical designs, the results point to a clear pattern: sectors with higher AI exposure recorded significantly higher producer-price inflation in the post-2022 period. The estimated differential is economically meaningful, on the order of roughly three-quarters to one percentage point of annual producer-price inflation, and remains robust across alternative treatment definitions, exposure constructions, and sample designs. The sectoral evidence therefore reinforces the aggregate results of the paper. Rather than suggesting that AI has so far operated mainly as a pure cost-reducing force, it indicates that the adoption and deployment phase of AI has been associated with transitional cost pressures at the industry level. In this sense, the sector-level analysis provides direct evidence that the inflationary footprint of AI is visible not only in aggregate time series, but also in the cross-section of industries most exposed to AI-intensive production and adoption.

The third contribution of the paper is theoretical. The paper develops a unified quantitative dynamic framework that can rationalize an inflationary impact of AI. The model is a Dynamic Stochastic General Equilibrium (DSGE) framework with nominal rigidities and imperfect competition, augmented to allow AI to affect inflation and activity through several channels. On the supply side, AI raises productivity and lowers marginal cost, but it also requires costly complementary inputs associated with deployment, inference, software, data infrastructure, energy use, and organizational adjustment. On the pricing side, AI can alter firms' effective pricing wedges by improving demand learning and price-setting capacity. On the labor-market side, AI can affect wage-setting power and the sensitivity of wages to labor-market tightness. Finally, on the demand side, AI can stimulate investment in complementary capital and infrastructure. The empirical evidence points to the relevance of these mechanisms, while the model provides a disciplined way to study how they interact. Because the channels can be activated individually or jointly, the framework makes it possible to compare how each mechanism alters the joint dynamics of the output gap and inflation under a standard interest-rate rule.

The interaction of these channels clarifies why the net inflation effect of AI is not mechanically disinflationary even when productivity gains are present. A rise in productivity is a mechanism that exerts a systematic downward force on inflation by reducing unit costs, but it competes with forces that can raise marginal costs or desired markups during diffusion. AI-specific input costs arise because the deployment of AI at scale requires scarce complementary resources—compute, energy, cooling, networking, and specialized capital—whose shadow prices can rise when capacity expands rapidly or adjustment is slow. Endogenous markups can increase if AI improves the ability of firms to segment demand, optimize pricing, or pass shocks through more effectively, raising the sensitivity of inflation to cost and demand conditions. Labor-market wedges can rise if AI alters bargaining positions, monitoring technologies, or outside options, generating wage markdown dynamics that change the mapping from activity to marginal costs and inflation. Crucially, the demand channel is not an ancillary feature in the recent data environment. The scaling of AI has been accompanied by a surge in investment and infrastructure expenditure, with data-center expansion serving as a prominent manifestation of the deployment wave. In the model, this channel raises the output gap and exerts upward pressure on inflation, and it can reinforce inflationary supply-side distortions when complementary inputs are scarce. The empirical contribution results, which show a stronger positive inflation association in the later sample when deployment-oriented indicators accelerate, motivate placing greater emphasis on the demand and infrastructure-scaling margin than would be suggested by narratives that focus on productivity alone.

A central diagnostic implication of the framework concerns the joint behavior of inflation and activity in response to AI-related fluctuations. Because both productivity and AI-induced investment tend to raise desired production and demand, a weak or negative output-gap response is informative: it implies that inflationary wedges operating through input costs, markups, or wage-setting frictions are strong enough to offset the expansionary supply impulse from productivity. Conversely, even when activity expands, inflation can rise if the expansion is driven by the demand and investment impulse while marginal costs and markups are pushed up by complementary-input scarcity or pricing wedges. In this sense, observing rising activity is not sufficient to infer disinflationary AI forces; it may reflect a strong investment-driven diffusion episode whose inflationary consequences are amplified by bottlenecks in the inputs required to deploy and serve AI. This diagnostic perspective helps reconcile mixed empirical findings and provides a structured way to interpret the timing and sign of the inflation response during periods of rapid AI scaling.

The quantitative analysis is carried out through a sequence of environments that progressively enrich the economic mechanisms while preserving interpretability. The paper begins with a static framework that isolates competing cost and markup forces, then embeds the logic in a one-sector New Keynesian setting to characterize dynamic propagation under monetary policy, then extends the analysis to a two-sector environment with heterogeneous AI intensity to make reallocation and relative-price channels explicit, and finally introduces endogenous AI investment and accumulation so that AI capability becomes a stock that builds gradually and generates persistence. This sequence allows the paper to separate purely contemporaneous cost and markup effects from investment-driven dynamics, and to show how sectoral heterogeneity can mask important cross-sector reallocations even when aggregate outcomes appear modest. Across environments, the qualitative ranking of mechanisms remains stable: the productivity channel is disinflationary in isolation, the demand and investment channel tends to raise inflation and amplify expansions, and the combination of AI-input costs and pricing and labor wedges can overturn the disinflationary effect of productivity and deliver an inflationary response even when output rises.

A complementary long-run exercise in the paper studies whether the effect of AI on inflation extends beyond short-run fluctuations. Using long-horizon projections, the analysis shows that higher AI intensity is associated with persistently higher future inflation paths, not merely temporary movements around the current inflation rate. This pattern remains present even in specifications that abstract from the data-center channel, indicating that the inflationary effect of AI is broader than any single infrastructure component. Specifically, the model projects a 1 percentage point, and possibly larger, increase in the inflation rate over a period of 10 years.

To interpret these findings, the paper develops a parsimonious New Keynesian framework in which AI diffuses gradually into effective production and affects inflation through both demand expansion and a composite supply-side wedge. The model matches the central empirical pattern: inflation adjusts gradually toward a higher long-run level as AI adoption spreads through the economy. In this way, the empirical projections and the structural framework both suggest that the impact of AI on inflation is not only a short-term phenomenon, but can persist over the medium to long run as diffusion continues.

A rapidly expanding literature reinforces the need for such a multi-channel perspective. Conceptual and forward-looking work emphasizes that AI does not follow a predetermined macro trajectory. Brynjolfsson and Unger (2023) describe AI as a set of "forks in the road" in which productivity, inequality, and industrial-organization outcomes depend on adoption patterns, institutional constraints, and policy choices. Their scenarios highlight that AI can generate either broad-based productivity gains or narrow labor-displacing improvements; either inclusive wage compression or polarization; and either concentrated "superstar" dynamics or more competitive ecosystems depending on openness, regulation, and diffusion. This perspective directly motivates the modeling choice in this paper: the inflation consequences of AI cannot be understood through a single mechanism, because the technology itself activates multiple wedges whose relative strength depends on the economic environment.

Scenario analyses echo these themes while emphasizing that the transitional macro path depends critically on policy. Nygaard et al. (2025) estimate that AI could add 0.1–1.5 percentage points to annual TFP growth, but they stress that the distribution of gains and the inflationary path depend on labor-market adjustment, retraining, and the speed of diffusion. Their finding that AI behaves like a demand shock in high-investment phases aligns closely with the infrastructure-scaling margin emphasized in this paper. Similarly, Melina and Villa (2025) show in a structural macro model that ICT and AI investment can raise inflation and output simultaneously, with the natural rate of interest initially falling and then rising persistently as productive capacity expands. These dynamics mirror the investment-driven demand channel in our model and highlight why AI cannot be treated as a purely supply-side force. A related exercise by Faria-e Castro and Ozkan (2026) uses the St. Louis Fed DSGE model to study AI optimism as a TFP news shock. Their analysis shows that expectations of future AI-driven productivity gains can raise current output and inflation even before productivity materializes, because households and firms increase spending and investment in anticipation of future gains. The result is especially relevant for the present paper because it formalizes a mechanism through which AI can be inflationary during the anticipation and deployment phase, even if the technology is ultimately productivity-enhancing. Their counterfactuals also show that the monetary-policy response matters: if policy accommodates the news shock, the inflationary effect becomes larger and more persistent.

Micro and firm-level evidence provides robust support for a productivity channel. Randomized experiments such as Noy and Zhang (2023) and Brynjolfsson et al. (2025) show that access to large language models reduces completion time, raises output quality, and disproportionately benefits lower-performing workers. Firm-level studies reinforce this picture: Alderucci et al. (2024) show that firms producing AI-related innovations subsequently experience faster growth in scale and productivity, while Tambe et al. (2020) find that AI adoption and AI-related hiring are associated with higher sales growth, higher measured productivity, and higher market valuations, especially among firms with complementary intangible assets and data capabilities. These findings justify modeling AI as a productivity-enhancing force at the micro level.

However, these gains do not automatically translate into aggregate productivity growth.

Brynjolfsson et al. (2017) describe a “productivity paradox” in which rapid advances in digital technologies coexist with modest aggregate productivity growth. They attribute this to diffusion lags, measurement problems, and the need for complementary organizational capital. Acemoglu (2024) similarly argues that the aggregate impact of AI depends on the breadth of tasks affected, the costs of reorganization, and whether applications expand output or merely redistribute rents. These insights motivate the empirical strategy of this paper: we construct a quarterly AI intensity measure precisely because aggregate data do not yet exhibit a clear AI-driven acceleration, and because the inflation consequences of AI may arise even when aggregate productivity does not move strongly.

A second strand documents that AI can reduce effective marginal costs while simultaneously drawing heavily on scarce complementary inputs. Sectoral evidence from Borowski et al. (2025) shows that higher AI adoption is associated with lower producer-price inflation in European industries, consistent with cost compression in sufficiently AI-intensive activities. At the same time, syntheses by Luccioni et al. (2024) and Luccioni et al. (2024) highlight that AI workloads are substantially more compute- and energy-intensive than traditional digital services. They estimate that inference for state-of-the-art conversational systems can require significantly more energy than a standard web search, and that AI workloads already represent a meaningful share of data-center electricity use. Energy-system assessments project that data-center power demand could rise sharply as AI diffusion accelerates, implying a front-loaded wave of complementary investment and resource absorption. These findings motivate the AI-input-cost channel in our model: AI can raise marginal costs when complementary inputs are scarce or adjustment is slow.

A third strand examines how AI affects pricing behavior and market structure. Using job-posting data, Adams et al. (2025) show that firms adopting AI-powered pricing tools experience faster growth in sales, employment, assets, and markups, and exhibit stronger stock-price reactions to monetary policy surprises. Reinforcement-learning pricing algorithms can sustain supra-competitive prices in inflationary environments (Tinoco et al., 2025). These findings support a markup channel in which AI improves firms’ ability to segment demand, optimize pricing, or pass shocks through more effectively. Related work shows that AI interacts with market structure in ways that can amplify pricing power. Models treating generative AI as a “centripetal technology” show that AI can reduce variety and raise concentration in the long run (Turegeldinova et al., 2025). Evidence from automation more broadly indicates that technology adoption tends to raise sales concentration and generate “superstar firms” (Firooz et al., 2025). These dynamics connect naturally to the economics of intangible capital, where scalability and non-rivalry amplify differences in firm capabilities (Crouzet et al., 2022). Our model captures these forces through an endogenous markup wedge that can rise with AI intensity.

A fourth strand highlights how AI affects wage-setting and labor-market dynamics. Task-based analyses show that AI tools raise productivity especially for lower-performing workers, while firm-level adoption reshapes occupational structure and increases STEM-intensive roles (Babina et al., 2024). Automation and AI can weaken workers’ bargaining power, flattening the Phillips curve (Basso and Rachedi, 2025). These findings motivate the wage-markdown channel in our model, in which AI alters the sensitivity of wages to labor-market conditions and changes the mapping from activity to marginal cost.

The empirical literature sketched above paints a nuanced picture that aligns well with the structure of the model. At the micro level, there is evidence that AI raises productivity and improves measured performance where it is adopted. There is also compelling descriptive evidence that scaling AI is capital- and energy-intensive, requiring large investments in specialized inputs and infrastructure. Studies of pricing and competition suggest that AI can raise markups and strengthen market power in some settings, and related work on automation indicates that new technologies can weaken workers' bargaining power and flatten the Phillips curve. Finally, AI adoption appears to be associated with substantial complementary investments in intangible capital and organizational change that can generate a significant demand impulse.

At the same time, aggregate output and productivity data have not yet exhibited a clear, AI-driven takeoff. This is broadly consistent with an environment in which local productivity and investment channels are strong but are mediated by diffusion lags, measurement frictions, and general-equilibrium forces—including cost pressures from scarce complementary inputs and changes in market structure—that can delay or offset a visible aggregate acceleration. In our framework, the output gap plays a diagnostic role. If micro evidence suggests that AI should raise activity through productivity and complementary investment, but the measured output gap does not rise markedly, a natural interpretation is that other channels—higher AI-related input costs, rising desired markups, and altered wage-setting dynamics—may be restraining measured aggregate activity even as they influence inflation.

The empirical evidence, therefore, supports a modeling strategy in which AI enters both the supply side and the demand side of the economy, and in which the net effect on inflation is the outcome of offsetting forces. Our quantitative analysis below uses this structure to map different observed combinations of output-gap and inflation dynamics into statements about the relative strength of the AI channels suggested by the literature.

The remainder of the paper is organized as follows. Section 2 discusses the measurement of AI and presents the construction of the AI indicators used throughout the analysis. Section 3 quantifies the contribution of AI to inflation using these measures. Section 4 then turns to sector-level evidence and presents Difference-in-Differences results. Section 5 introduces a simple static framework that highlights the opposing supply-side forces through which AI can affect inflation, while Section 6 embeds the same logic in a baseline New Keynesian model. Section 7 extends the analysis to a richer non-linear environment with endogenous AI adoption, substitution between labor and automation, and multiple channels through which AI influences inflation; this section also presents two-sector analysis. Section 8 examines the impact of AI on long-run inflation rate, and Section 9 concludes.

2 Measuring AI: Construction of a Quarterly U.S. AI Activity Index

This section describes the construction of a set of quarterly U.S. indicators intended to proxy for the intensity of AI-related activity. The guiding principle is to triangulate AI's macroeconomic footprint using observables that capture distinct margins of development and diffusion: investment in AI-relevant tangible and intangible capital, the resulting installed base of capital services, complementary infrastructure, and the flow of codified AI innovations. Because no single series provides a sufficient statistic for AI, we work with a small dashboard of component indicators and combine them into composite indices that are suitable for time-series and structural analysis.

2.1 Quarterly panel and normalization

The data sample covers the period 2007Q1-2025Q3, and the sample that covers data centers starts in 2014Q1. All series are aligned to a common quarterly panel over the sample period.

When the underlying data are observed at higher frequency, they are mapped to quarter-end dates and aggregated to the quarterly level using a fixed rule, typically a within-quarter mean for flow variables and a within-quarter mean for high-frequency indicators such as survey responses or benchmark statistics. For comparability across heterogeneous units, each level series is converted into an index normalized to 100 at its first available observation. Missing values are retained as missing rather than imputed; when we later construct composite indicators, they are based on the average of the standardized components that are actually observed in a given quarter.

2.2 Component indicators

The core measurement system rests on four main components that correspond closely to the channels emphasized in macroeconomic models of technology adoption and diffusion: AI-oriented investment, AI-related capital services, data-center infrastructure, and AI innovation as proxied by patents.[1]

AI-oriented investment. AI investment is proxied by real private fixed investment in categories that are tightly linked to the deployment and production of AI technologies. The primary AI-investment index aggregates real investment in information processing equipment, software, and research and development. Each of these components is taken from the national accounts in constant-price form, aligned to quarter-end dates, and summed to form a single real investment flow that is then converted into an index normalized to 100 at its first observation. This construction is meant to capture both the tangible capital that supports compute-intensive AI applications and the intangible knowledge capital that underpins AI-related innovation and deployment.

AI capital services. Investment flows are complemented with a stock-and-services measure designed to capture the effective installed base of AI-relevant capital. Starting from real investment in software, computers, communications equipment, research and development, and non-residential structures, we construct capital stocks using a standard perpetual-inventory method with asset-specific depreciation rates that are converted to quarterly equivalents. For each asset, the capital stock evolves as the previous-period stock net of depreciation plus current-period investment. We then aggregate these asset-level stocks into an economy-wide capital-services index using a Tornqvist aggregation scheme, where growth in the aggregate is the share-weighted average of asset-level log stock growth rates, with weights proportional to the nominal investment shares of each asset. The resulting capital-services series is normalized to 100 at its first observation. Conceptually, this measure is closer to the productive capacity associated with AI-related capital than raw investment, because it incorporates both accumulation and depreciation dynamics and attaches higher weight to assets that account for a larger share of total spending.

Data-center infrastructure. To capture the complementary infrastructure required to deploy AI at scale, we construct an indicator of data-center construction activity in the United

[1]Ideally, one would also want to account for the capability frontier of AI systems and for the diffusion or adoption of AI across firms and production processes. We examined indicators intended to capture these dimensions, including a model-performance-based capability measure and a survey-based adoption measure. However, these series begin only recently and are therefore not available for a long enough period to be included in the construction of the AI measures used in the time-series analysis below.

States. The underlying series measures the volume of new data-center projects over time. Observations are mapped to quarter-end dates and aggregated to a quarterly series, which is then normalized to 100 at its first observation. This component is intended to proxy for the expansion of compute and storage capacity that often accompanies the diffusion of AI systems into production. Because data-center construction begins later in the sample and displays pronounced high-frequency variation, this indicator is especially informative about the recent period of rapid scale-up.

AI innovation: patents. Innovation is proxied by a count-based measure of AI-related patents. We identify patents that are tagged as AI-related in an underlying micro dataset and aggregate them to the quarterly level. For each quarter, we count the number of AI patents granted (or filed, depending on the timing convention), assign the observation to the quarter-end date, and convert the resulting count series into an index normalized to 100 at its first observation. This indicator is designed to capture time variation in the flow of codified AI-relevant inventive activity, recognizing that patent counts are volatile and episodic but can still convey medium-run trends in innovative intensity.

2.3 Standardization and composite indices

Because the component indicators differ in units, scale, and volatility, aggregation is performed in standardized units. Let $A_{j,t}$ denote the level index for component j at quarter t (for example, the investment index, the capital-services index, the data-center construction index, or the AI-patent index). For each component, we compute a standardized series

$$Z_{j,t} \;=\; \frac{A_{j,t} - \bar{A}_j}{s(A_j)}\,, \tag{1}$$

where $\bar{A}_j$ and $s(A_j)$ denote the sample mean and standard deviation of component j computed over the available quarters. Standardization is applied separately to each component.

The broad AI activity index averages, in standardized units, four conceptually distinct components: the AI-oriented investment index based on information processing equipment, software, and research and development; the AI-related capital-services index; the data-center construction index; and the AI-related patent index. This composite is intended to summarize broad AI intensity while remaining transparent about the underlying building blocks and their interpretation.

To examine mechanisms and assess robustness, we also construct several alternative indices. A composite that excludes data centers averages the standardized AI investment index, capital-services index, and AI-patent index. This series is especially useful because it extends back further in time and is less influenced by the sharp late-sample swings in data-center construction. A composite that excludes capital services combines the standardized AI investment index, data-center index, and AI-patent index, thereby placing relatively greater weight on contemporaneous investment flows and infrastructure deployment. Finally, a parsimonious supply-side AI capacity index averages the standardized capital-services and AI-patent indices, focusing on the accumulated productive capacity and documented innovative output associated with AI.

2.4 Interpretation and link to the GDP deflator

The resulting measurement system delivers a set of interpretable component indicators and a family of composite AI activity indices in standardized units. The component series preserve economic meaning: the investment index reflects flows of AI-related capital formation, the capital-services index captures the installed base of productive capacity, the data-center index tracks complementary infrastructure, and the patent index summarizes the flow of codified inventions. The composite indices then translate these heterogeneous movements into a single, standardized measure that can be used as a forcing variable in reduced-form regressions and in structural models. Having several ways to summarize AI activity is a feature rather than a complication. The presence of multiple composite definitions allows us to check that our qualitative and quantitative conclusions are not an artifact of a single measurement choice.

In the empirical work that follows, we primarily relate these AI activity indices to the behavior of the GDP price deflator. The GDP deflator is the most natural macroeconomic price index for our purposes for three reasons. First, it is conceptually closest to the marginal-cost term that appears in New Keynesian Phillips curves, since it measures the price of domestically produced value added rather than the cost of a consumption basket. Second, it aggregates price movements across the broad set of sectors where AI-related capital, infrastructure, and innovations are likely to affect production costs and markups, rather than focusing on a more restricted set of consumer expenditures.

Throughout, we use the AI activity indices as standardized measures of AI intensity that can be interpreted as external shifters of technology, input costs, and effective demand in the macroeconomic environment, and we use the GDP deflator as the central price index for quantifying how these AI-related forces map into observed inflation dynamics.

2.5 AI measurement components and composite indices

Figure 1 reports the individual indicators that enter our measurement system. The investment-related components—*Software*, *Information Processing Equipment*, and *Research & Development*—display a clear upward trajectory over the full sample, with a pronounced acceleration in the post-2016 period and especially in the most recent years. While these three series differ in smoothness and short-run variability, their common pattern is a sustained rise that is consistent with a broad-based increase in AI-related capital formation and intangible accumulation. The *Data Centers* series is distinct in two respects: it begins later (reflecting data availability) and exhibits the sharpest late-sample surge, capturing the rapid scale-up of compute-intensive infrastructure. This steep increase toward the end of the sample is informative precisely because it is not a mild extension of prior trends; rather, it signals a discrete intensification of infrastructure deployment that is tightly linked to the recent wave of large-scale AI adoption.

The remaining indicators are designed to capture complementary dimensions beyond investment flows. The *Capital (AI Capital Services)* index rises smoothly and persistently, reflecting the cumulative nature of capital services constructed from investment dynamics and depreciation; its monotone-like profile underscores that the capital-services channel is a slow-moving but powerful contributor to the overall AI signal. The *Patents* series is more volatile and episodic, featuring bursts and pullbacks that are typical of innovation counts based on discrete events; nevertheless, the medium-run movement is upward, indicating a sustained increase in AI-relevant inventive activity.

Figure 2 aggregates these ingredients into composite indices that emphasize different mechanisms. The *Investment: Excluding Data Centers* panel provides a broad summary of AI-related

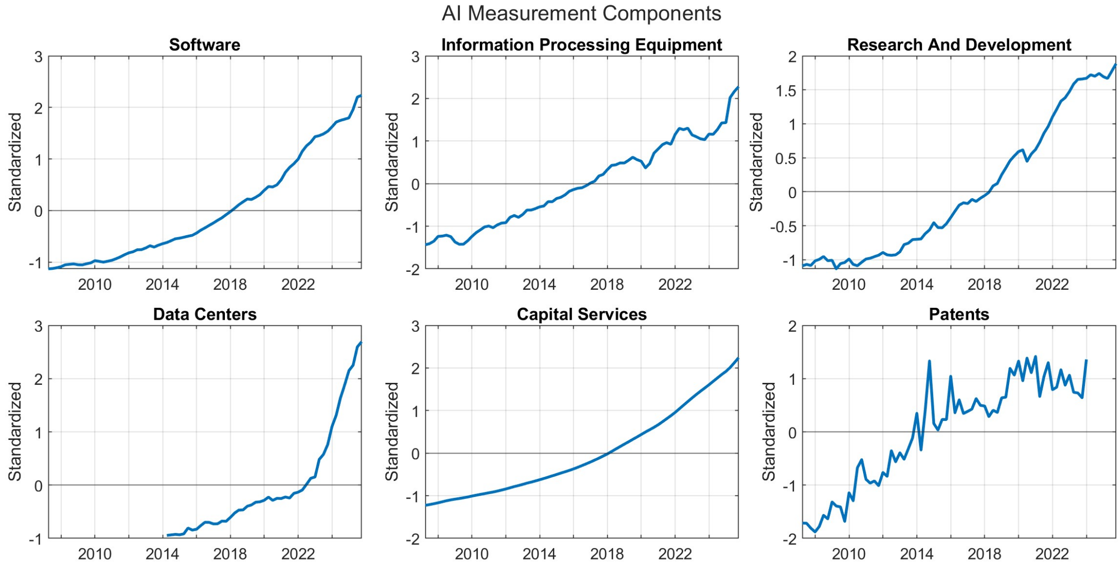


Figure 1: Individual AI-related indicators (standardized). The figure reports the time paths of the component series used to summarize AI-related investment, capital deepening, innovation, and infrastructure. All series are shown in standardized units (mean zero, unit variance over their available sample), so differences in levels reflect relative movements rather than original measurement units.

investment activity excluding data centers and inherits the gradual upward movement visible in the underlying investment components. The *Investment: Including Data Centers* panel adds the data-center series to this investment index; as a result, it starts later and exhibits a more pronounced increase near the end of the sample, reflecting the surge in compute infrastructure. Comparing these two panels clarifies an important measurement point: including data centers materially changes the late-sample dynamics, not by altering earlier trends, but by amplifying the most recent acceleration.

The *Capital Services & Patents* panel isolates two conceptually distinct forces: accumulated productive capacity, as captured by capital services, and innovative output, as captured by patents. Relative to the pure investment measures, this series reflects a combination of persistence from capital accumulation and additional variation associated with patent activity. Turning to the broader composites, the *Composite: Excluding Data Centers & Capital Services* panel removes both data centers and capital services and therefore places relatively greater weight on the remaining investment component together with patents. The *Composite: Excluding Data Centers* index combines the main AI investment signal excluding data centers with capital services and patents, yielding a long and internally consistent series that summarizes investment, accumulation, and innovation in a single measure. Finally, the *Composite: Including Data Centers* variant begins when the data-center series becomes available and then rises more rapidly toward the end of the sample, reinforcing that data centers primarily affect the most recent part of the period under study. In the analysis that follows, we use this range of AI measures to assess the impact of artificial intelligence on inflation.

3 Quantifying the contributions of AI to inflation

This section constructs descriptive "AI contribution" time series for inflation. The goal is *not* a structural decomposition, but a transparent accounting device: for each AI proxy we estimate its average same-quarter association with inflation, and we translate the fitted AI component into a

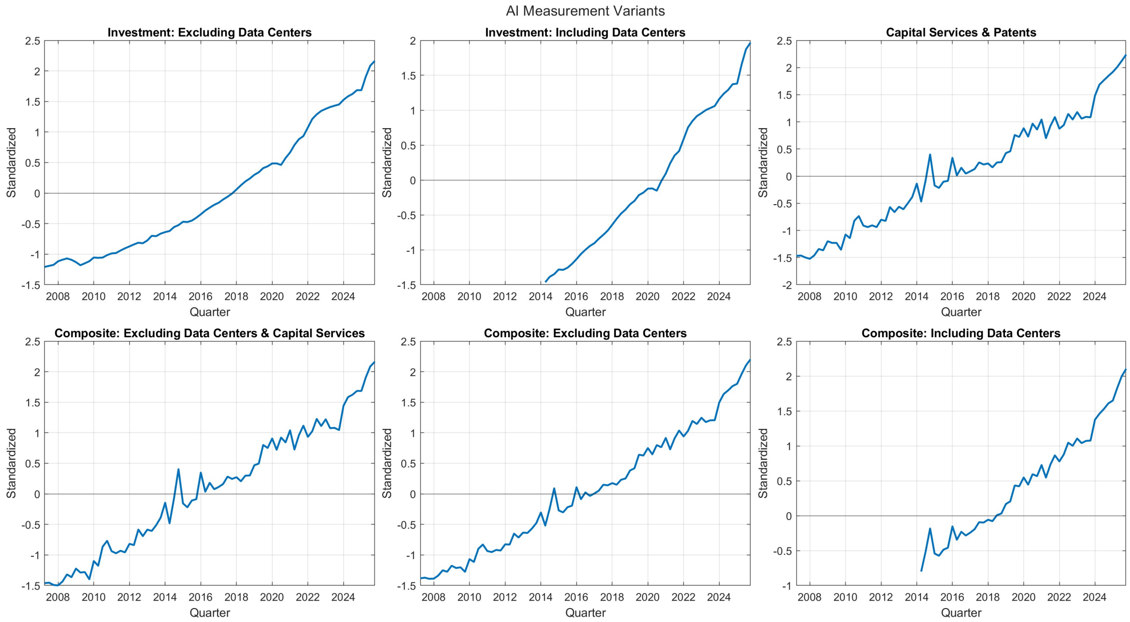


Figure 2: Composite indices (standardized) constructed from the component indicators. Top row: an AI-investment index (constructed from information processing equipment, software, and research & development), the corresponding investment index including data centers, and a composite of capital services and patents. Bottom row: broader composites that combine investment, capital services, and patents, with variants that either include data centers or exclude capital services. Series that incorporate data centers start later due to data availability for the data-center measure.

quarterly series that can be compared across AI measures and across specification environments. Because the exercise is reduced-form, the resulting contributions should be read as *conditional associations* under the maintained specification, not as causal effects.

The contribution methodology has three building blocks. First, we define the inflation outcome at the quarterly frequency and place it on a common unit basis. Second, we transform each AI proxy into a quarterly regressor that captures quarter-to-quarter movements in standardized AI intensity. Third, we estimate a contemporaneous projection and map the fitted AI component into a contribution series. We then repeat the exercise under alternative specification environments—a simple benchmark projection and a projection augmented with controls—to assess the stability of the sign and the time pattern of the implied contributions. Further details on the construction of the contribution series, the control variables, and the alternative empirical designs are provided in Appendix B.

3.1 Macroeconomic outcome (GDP-deflator inflation) and sample

Let t index quarters. Our headline inflation outcome for the contribution analysis is the *Gross Domestic Product: Implicit Price Deflator* (GDP deflator). The GDP deflator is the most natural inflation counterpart in this exercise because it is a national-accounts concept tied to *domestic production*. The GDP deflator measures the price of domestically produced final output and excludes imports by construction. This alignment matters for interpretation: if AI is hypothesized to affect inflation primarily through production-side channels (e.g., productivity, marginal costs, markups, and sectoral reallocation within domestic value added), then a domestic-output price index is the closest empirical object to the model's inflation concept.

Formally, we define GDP-deflator inflation as

$$\pi_t^{GDP} \equiv 400\big(\log P_t^{GDP} - \log P_{t-1}^{GDP}\big), \tag{2}$$

where P_t^{GDP} denotes the GDP deflator. The rate π_t^{GDP} is expressed in annualized q/q percentage points. The main sample period is 2007Q1–2025Q3, with shorter effective windows when required by data availability for specific AI variants.

3.2 Transforming AI measures into quarterly movements

Let Z_t denote a standardized AI measure at quarterly frequency. Because the underlying AI indicators differ in units, scale, and volatility, the contribution exercise is carried out using changes in standardized AI intensity. For each AI proxy, the regressor used in the inflation projection is therefore

$$x_t \equiv \Delta Z_t \equiv Z_t - Z_{t-1}. \tag{3}$$

This transformation has two advantages. First, it places the AI measures on a common scale, so that investment-based measures, capital-services measures, data-center measures, patent measures, and composite indices can be compared within the same empirical framework. Second, it focuses the contribution exercise on quarter-to-quarter movements in AI intensity rather than on differences in the original units of the underlying series.

For composite indices, Z_t is the standardized composite itself. For individual AI components, the raw level series is first normalized and standardized, and the contribution regressor is then the first difference of that standardized series. Thus, the estimated AI contribution reflects the association between GDP-deflator inflation and quarterly changes in standardized AI intensity. The AI regressor and the GDP-deflator inflation outcome are aligned to the same quarterly calendar prior to estimation.

3.3 Baseline contribution model and interpretation

The starting point is a contemporaneous linear projection of GDP-deflator inflation on quarterly AI movements,

$$\pi_t^{GDP} = \alpha + \beta x_t + u_t, \tag{4}$$

where $x_t = \Delta Z_t$ is the quarterly change in the standardized AI measure and u_t is a residual. The coefficient β summarizes the average same-quarter association between changes in standardized AI intensity and GDP-deflator inflation over the estimation window.

In practice, the regression in Equation (4) is estimated by ordinary least squares on the set of quarters for which both π_t^{GDP} and x_t are observed, possibly including a simple indicator for the 2020–2022 pandemic and the surge in inflation episodes as an additional regressor. This dummy absorbs the level effect of that exceptional period; for the purposes of defining the AI contribution, it is treated as part of the conditioning set and does not alter the definition of the AI component below.

Given the estimated slope $\widehat{\beta}$ from Equation (4), we define the quarterly AI-attributed fitted component, or "AI contribution," as

$$\widehat{c}_t^{AI} \equiv \widehat{\beta}\, x_t. \tag{5}$$

This definition follows directly from the fitted value of the linear projection. In Equation (4), the fitted inflation rate is

$$\widehat{\pi}_t^{GDP} = \widehat{\alpha} + \widehat{\beta} x_t.$$

The term $\widehat{\beta}x_t$ is therefore the part of fitted inflation that is mechanically associated with the observed AI movement in quarter t. Equivalently, it is the difference between the fitted value using the observed AI movement and the fitted value that would obtain, within the same projection, if the AI movement were set to zero while the intercept and any other conditioning variables were held fixed. This is why $\widehat{\beta}x_t$ is interpreted as the AI-associated fitted component of inflation.

By construction, $\widehat{c}_t^{AI}$ has the same units as the dependent variable, π_t^{GDP}, namely annualized percentage points. The AI regressor itself is a change in standardized AI intensity, so $\widehat{\beta}$ measures the change in annualized GDP-deflator inflation associated with a one-unit quarterly increase in the standardized AI measure. The contribution series is defined on the full quarterly grid for which x_t is available; outside the overlapping estimation sample, $\widehat{c}_t^{AI}$ is left missing. Within the estimation window, the regression pins down a single slope $\widehat{\beta}$, and the time variation in the contribution series comes entirely from the observed movements in x_t.

This fitted component is best read as the portion of GDP-deflator inflation in quarter t that is linearly associated with the observed AI movement in that quarter, under the benchmark specification. It is not a structural decomposition: omitted variables, simultaneity, and measurement error in both π_t^{GDP} and x_t can affect $\widehat{\beta}$ and therefore the attribution in Equation (5). These concerns motivate a second specification that conditions explicitly on major non-AI inflation drivers.

3.4 Controls-augmented contribution model

To assess whether the benchmark contribution patterns are robust to plausible confounders, we estimate a controls-augmented projection for GDP-deflator inflation,

$$\pi_t^{GDP} = \alpha + \beta x_t + \theta' \mathbf{X}_t + \varepsilon_t, \tag{6}$$

where $\mathbf{X}_t$ is a vector of quarterly controls capturing major non-AI influences on domestic inflation, as well as measurement-related persistence. In the baseline implementation, $\mathbf{X}_t$ includes a global supply-bottlenecks index, real transfer payments to households deflated by the GDP deflator, an energy-price indicator, a measure of the size of the Federal Reserve's balance sheet, a proxy for tariff developments, and the same dummy variable used in the benchmark specification.

Equation (6) is again estimated by ordinary least squares on the maximal overlapping sample. The AI contribution series in the controls case is defined analogously to Equation (5):

$$\widehat{c}_t^{AI,\,\mathrm{ctrl}} \equiv \widehat{\beta}^{\,\mathrm{ctrl}} x_t, \tag{7}$$

where $\widehat{\beta}^{\,\mathrm{ctrl}}$ is the slope coefficient on x_t in Equation (6). The difference relative to the benchmark is interpretive: $\widehat{c}_t^{AI,\,\mathrm{ctrl}}$ isolates the part of GDP-deflator inflation that is linearly associated with AI movements *conditional on* the chosen set of non-AI drivers. Comparing the benchmark and controls-based contribution series therefore provides a simple robustness check on the sign, magnitude, and time profile of the inferred AI component.

3.5 Contributions of AI to Inflation: Results

Table 1 reports the windowed contributions of AI to inflation across alternative constructions of the AI measure and across two identification environments: a benchmark specification and an otherwise similar specification that conditions additionally on a standard set of macro controls. The headline pattern is one of robustness. Whenever AI is measured in a broad, composite way, the estimated contribution to GDP-deflator inflation is positive, economically meaningful, and very similar across the two specifications. For the composite that includes all components, including data centers, the mean contribution lies in the 0.16–0.20 range in annualized percentage points, the median is around 0.20–0.25, and roughly three quarters of the quarters in the relevant sample window record a positive contribution. The 90% confidence intervals for these composites lie comfortably above zero in both specifications. This alignment across benchmark and control-augmented environments suggests that the inference is not being driven by a particular conditioning set, but rather reflects a common underlying mechanism that survives reasonable perturbations to the empirical environment.

The table also makes clear that a positive inflation contribution does not require an investment-only channel. For the GDP deflator, the composite measures that exclude data centers & capital services or exclude data centers continue to deliver small but clearly positive mean contributions, on the order of a few hundredths of a percentage point per year, with about two thirds of the quarters showing positive contributions in both specifications. These patterns are suggestive of channels through which AI raises inflation beyond measured investment alone: adoption and diffusion can increase effective demand for complementary inputs, reallocate activity toward AI-intensive production margins, and generate bottlenecks or capacity constraints that elevate marginal costs. In that sense, the non-investment composites show that the overall inflation contribution does not collapse to zero when the investment component is dialed back; instead, they indicate that broader diffusion and complementary-input channels are themselves inflationary, albeit at a more moderate quantitative scale than the pure investment proxies.

Case	Variant	Mean	CI	Median	Cum.	Pos. share	N
Benchmark	Investment: Excluding Data Centers	0.633	(0.515, 0.751)	0.567	12.424	89.2	74
	Investment: Including Data Centers	1.105	(0.905, 1.304)	0.962	13.544	95.7	46
	Capital Services	1.137	(1.017, 1.257)	0.991	23.414	100.0	74
	Patents	0.003	(-0.002, 0.007)	0.001	0.044	50.7	67
	Data Centers	0.441	(0.296, 0.585)	0.282	5.199	82.6	46
	Capital Services & Patents	0.022	(0.007, 0.037)	0.023	0.404	63.5	74
	Composite: Excluding Data Centers & Capital Services	0.030	(0.009, 0.051)	0.024	0.550	63.5	74
	Composite: Excluding Data Centers	0.058	(0.030, 0.086)	0.051	1.079	66.2	74
	Composite: Including Data Centers	0.196	(0.117, 0.275)	0.241	2.276	76.1	46
Controls	Investment: Excluding Data Centers	0.305	(0.248, 0.362)	0.273	5.801	89.2	74
	Investment: Including Data Centers	0.297	(0.243, 0.351)	0.259	3.473	95.7	46
	Capital Services	1.233	(1.103, 1.363)	1.074	25.626	100.0	74
	Patents	0.001	(-0.001, 0.004)	0.000	0.023	50.7	67
	Data Centers	0.284	(0.191, 0.377)	0.181	3.317	82.6	46
	Capital Services & Patents	0.019	(0.006, 0.032)	0.020	0.347	63.5	74
	Composite: Excluding Data Centers & Capital Services	0.014	(0.004, 0.024)	0.011	0.258	63.5	74
	Composite: Excluding Data Centers	0.033	(0.017, 0.049)	0.029	0.611	66.2	74
	Composite: Including Data Centers	0.163	(0.097, 0.229)	0.201	1.896	76.1	46

Table 1: Summary of AI contribution series across cases and AI variants (windowed mean and 90% CI; median reported). Mean/median are in annualized percentage points. Cumulative is log-compounded: $100\,(\exp(\sum_t c_t/400) - 1)$. Pos. (%): the share of quarters with positive contributions to inflation; $\widehat{c}^{AI} > 0$ within the window.

At the same time, the investment and capital-deepening channels are quantitatively important in the GDP deflator. The investment-only series deliver relatively large contributions, with mean annualized effects in the neighborhood of 0.30–1.10 percentage points depending on the variant and specification, and with positive contributions in roughly 89% to 96% of quarters. The capital-services measure is even stronger: the mean contribution is around 1.1–1.2 percentage points in both specifications and the share of positive quarters is effectively 100%. This configuration is consistent with the mechanics emphasized in the paper: measured AI capital services and investment are direct ways through which AI diffusion is accompanied by higher utilization of AI-related inputs and complementary resources, which propagate inflationary pressure through marginal costs and pricing decisions. The broader composite measures reconcile these forces with the remaining AI components, producing contribution estimates that are smaller than for investment or capital services alone but more representative of an "all-components" AI measure.

The contrast between the capital-services series and the pure patent-based measure is particularly informative. When the AI proxy is restricted to patents alone, the implied contribution to GDP-deflator inflation is essentially neutral: the mean contribution is close to zero in both specifications, the 90% confidence intervals straddle zero, and the share of positive quarters is near one half. This stands out as the only variant in which the contribution is both economically and statistically negligible. By contrast, the measure that combines capital services and patents yields a small but clearly positive contribution, with mean effects around 0.02 percentage points per year and confidence intervals that lie above zero. One interpretation is that patents largely capture the innovation and idea-creation margin of AI, which is more closely tied to gradual productivity improvements and less directly to contemporaneous utilization and capacity pressure. When patents are considered in isolation, the inflationary forces associated with deployment and capital deepening are largely absent, and the net contribution is close to zero. Once the patent series is combined with the capital-services measure, however, the shorter-run inflationary signal coming from realized AI capital deepening is strong enough to pull the composite into a positive, albeit modest, range.

It is useful to summarize the evidence for the GDP deflator by averaging across the two specifications for each AI variant. This aggregation highlights an additional form of robustness: except for the patent-only proxy, all AI measures yield positive average contributions to GDP-deflator inflation. The composite that includes data centers delivers an average mean contribution of about 0.18 percentage points per year across the two specifications, while the composites that exclude data centers & capital services or exclude data centers cluster around about 0.02 and 0.05 percentage points, respectively. The investment-only and capital-services variants are larger, with average contributions in the vicinity of roughly 0.47–0.70 percentage points and about 1.19 percentage points, respectively. In each of these cases, the share of quarters with positive contributions is high and the confidence intervals lie above zero. Taken together, these patterns indicate that whenever the AI measure loads meaningfully on realized deployment, capital deepening, and associated infrastructure, the contribution of AI to GDP-deflator inflation is both economically non-trivial and robust to reasonable changes in specification, whereas the purely patent-based measure stands out as the unique case with an essentially neutral impact.

3.6 Contribution of AI to Inflation: An Evolving and Rising Influence Over Time

This subsection summarizes what the time-series contribution measures imply about the evolution of AI-driven inflationary pressure over time for the GDP deflator. Figure 3 reports, for each AI indicator or composite, the estimated contemporaneous contribution of that measure to quarterly GDP-deflator inflation. Bars display the raw quarterly contributions in the benchmark and in the specification that conditions on a standard set of macro controls, while the solid lines plot the corresponding four-quarter trailing moving averages.

The first two panels focus on investment-based measures. When AI is proxied by investment excluding data centers, the contribution to GDP-deflator inflation is already positive and nontrivial through most of the sample. The moving-average lines drift upward over the 2010s, with contributions becoming more persistently above zero well *before* the pandemic, and then step up further during the more recent years. Once data centers are explicitly included in the investment proxy, the pattern becomes even more pronounced. The contributions of AI to inflation rise steadily from the mid-2010s onward, and display a sharp increase in the early-2020s. Crucially, the elevated contributions in the last part of the sample do not simply mirror the temporary inflation spike: even after aggregate inflation begins to moderate, the investment-based AI contribution to the GDP deflator remains clearly above its pre-2015 baseline. This suggests that AI-related investment, particularly when it incorporates data-center build-out, has become an increasingly important and persistent source of inflationary pressure.

The third panel turns to capital services, a narrower but conceptually direct measure of AI-related productive capacity. Here the GDP-deflator contributions show a pronounced upward trend: starting from modest values in the late 2000s, the capital-services contribution rises gradually through the 2010s and then accelerates in the early-2020s, reaching its highest levels at the end of the sample. The moving-average series rarely dips below zero after the mid-2010s, and by the early-2020s it implies contributions on the order of a percentage point or more at annualized rates. This pattern indicates that once AI is embodied in installed capital and services, the associated utilization and capacity pressures translate into substantial and increasingly persistent inflationary contributions.

By contrast, the fourth and fifth panels show that when the AI proxy is restricted to patents or to the data-centers measure, the implied contributions differ sharply. The patent-only series is centered very close to zero throughout the sample, with frequent sign changes and only occasional short-lived spikes; the moving average hovers near the horizontal axis. The data-centers series, by contrast, is positive over most of its available sample and rises materially in the later years, consistent with the view that compute infrastructure has become an increasingly important source of inflationary pressure.

The sixth panel, which combines capital services and patents, yields contributions that are positive but modest in magnitude. These patterns are consistent with the idea that patenting primarily captures an innovation and idea-creation margin that is more closely tied to gradual productivity improvements than to contemporaneous demand or capacity constraints. When patents are combined with capital services, the resulting series retains a positive inflationary signal, but one that remains much smaller than the contribution from capital services alone.

The bottom-row composite measures provide the clearest view of the overall evolution of AI-related inflation pressure. The composite that excludes data centers & capital services yields relatively small contributions early in the sample, with both positive and negative quarters, but becomes more persistently positive over time. By the late 2010s the moving-average contribution is consistently above zero, and it rises further during and after the early-2020s inflation episode. The composite that excludes data centers behaves similarly, though at slightly higher levels: its contributions fluctuate around zero in the late 2000s, turn more systematically positive in the mid-2010s, and remain elevated thereafter. These conservative composites—which deliberately downweight either capital deepening or data-center infrastructure—thus already indicate that broad AI activity has shifted from being a negligible contributor to GDP-deflator inflation

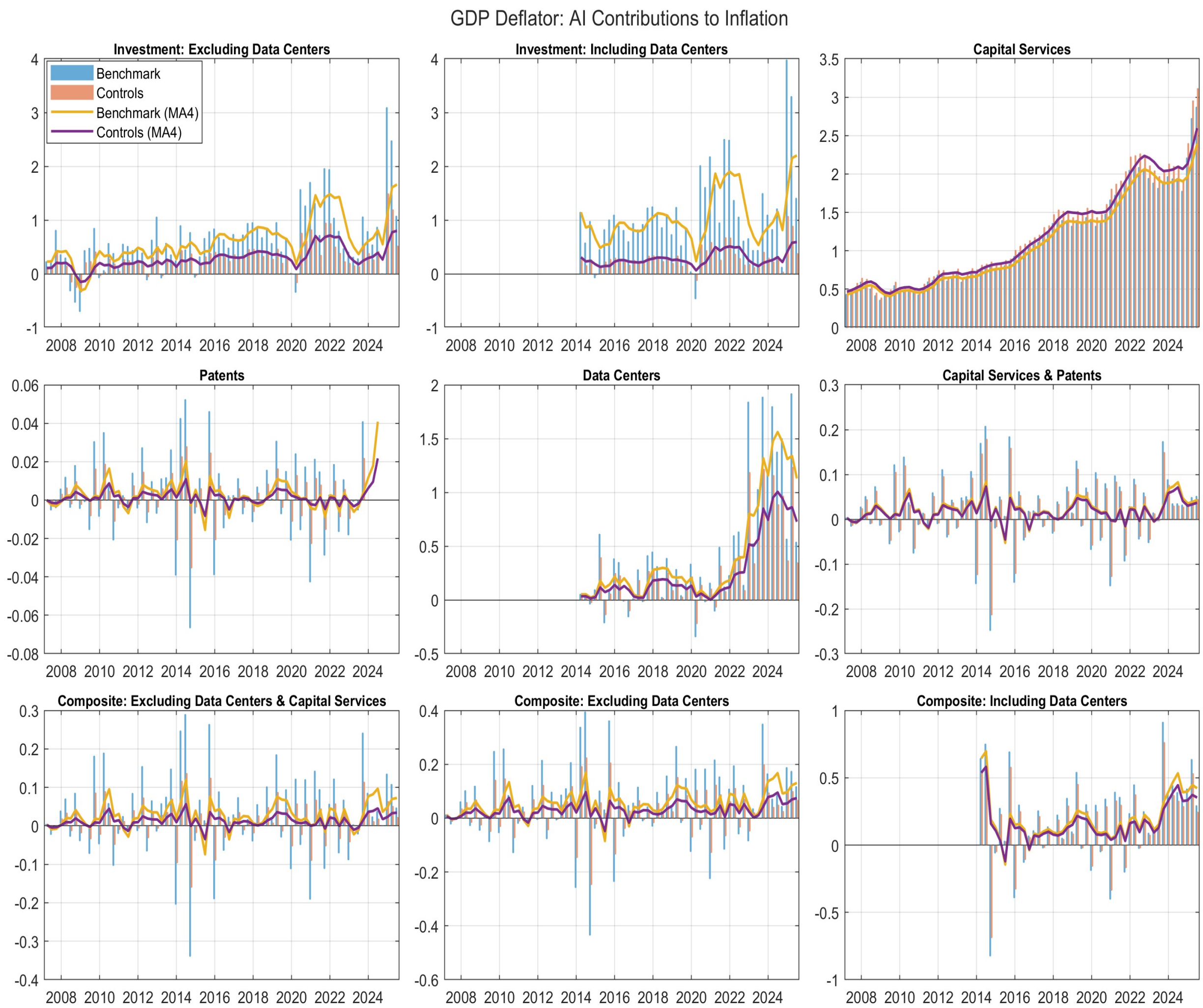


Figure 3: GDP deflator: AI contributions to inflation across alternative AI-related capital and composite variants. Bars display the raw quarterly contributions (Benchmark and Controls cases), while solid lines plot the corresponding four-quarter trailing moving averages. Panels report, clockwise from the top-left, investment excluding data centers, investment including data centers, capital services, patents, data centers, the capital-services-and-patents composite, the composite excluding data centers & capital services, the composite excluding data centers, and the composite including data centers.

toward being a modest but persistent positive driver.

The final panel, which reports the composite including data centers, shows the strongest and most policy-relevant evidence. From the late-2010s onward, the moving-average contribution becomes clearly positive and rises above the levels implied by the excluding-data-centers composite, reflecting the growing importance of AI infrastructure. The contribution steps up further around 2021–2022 and reaches its highest values in 2023–2024, remaining elevated relative to earlier years. The timing aligns with the rapid emergence and diffusion of generative AI technologies (such as *ChatGPT*) and the associated acceleration in data-center and compute-intensive investment. In this sense, the late-sample surge in the composite contribution appears to reflect both a continuation of a pre-existing upward trend in AI intensity and an additional push associated with the generative-AI acceleration.

Taken together, the time-series evidence in Figure 3 points to a clear and economically meaningful shift. The contribution of AI to inflation increases over time across a wide range of measures, with the rise beginning well before the COVID era and becoming especially pronounced once AI-related capital services, investment, and data-center infrastructure scale up in the late-2010s and early-2020s. Narrow patent-based proxies remain essentially neutral, but broader measures that load on realized deployment and infrastructure consistently indicate that AI has evolved into a significant and rising source of inflationary pressure.

Case	Variant	Mean	CI	Median	Cum.	Pos. share	N
Benchmark	Investment: Excluding Data Centers	1.049	(0.855, 1.244)	0.938	12.827	95.7	46
	Investment: Including Data Centers	1.105	(0.905, 1.304)	0.962	13.544	95.7	46
	Capital Services	2.467	(2.256, 2.679)	2.314	32.810	100.0	46
	Patents	0.005	(-0.008, 0.017)	0.004	0.045	53.8	39
	Data Centers	0.441	(0.296, 0.585)	0.282	5.199	82.6	46
	Capital Services & Patents	0.032	(0.005, 0.059)	0.042	0.371	69.6	46
	Composite: Excluding Data Centers & Capital Services	0.044	(0.006, 0.081)	0.042	0.504	69.6	46
	Composite: Excluding Data Centers	0.086	(0.038, 0.134)	0.095	0.999	71.7	46
	Composite: Including Data Centers	0.196	(0.117, 0.275)	0.241	2.276	76.1	46
Controls	Investment: Excluding Data Centers	0.240	(0.196, 0.284)	0.214	2.798	95.7	46
	Investment: Including Data Centers	0.297	(0.243, 0.351)	0.259	3.473	95.7	46
	Capital Services	1.761	(1.610, 1.911)	1.651	22.442	100.0	46
	Patents	0.012	(-0.020, 0.044)	0.011	0.119	53.8	39
	Data Centers	0.284	(0.191, 0.377)	0.181	3.317	82.6	46
	Capital Services & Patents	0.057	(0.009, 0.105)	0.075	0.663	69.6	46
	Composite: Excluding Data Centers & Capital Services	0.053	(0.008, 0.098)	0.050	0.607	69.6	46
	Composite: Excluding Data Centers	0.091	(0.041, 0.142)	0.101	1.054	71.7	46
	Composite: Including Data Centers	0.163	(0.097, 0.229)	0.201	1.896	76.1	46

Table 2: Summary of AI contribution series across cases and AI variants (windowed mean and 90% CI; median reported). Mean/median are in annualized percentage points. Cumulative is log-compounded: $100\,(\exp(\sum_t c_t/400) - 1)$. Pos. (%): the share of quarters with positive contributions to inflation; $\widehat{c}^{AI} > 0$ within the window.

Comparing the longer 2007Q1–2025Q3 sample with the shorter 2014Q1–2025Q3 window shows that AI's estimated contribution to inflation is generally stronger in the later period even when data centers are excluded. For example, the Benchmark Composite excluding Data Centers rises from 0.058 to 0.086 annualized percentage points, while the corresponding Controls measure rises from 0.033 to 0.091. Similarly, Benchmark Investment excluding Data Centers increases from 0.633 to 1.049. This suggests that the stronger contribution of AI to inflation in recent years is not driven solely by data centers. Rather, the broader set of AI-related investment and capital-deepening channels also appears to have become more inflation-relevant in the later sample.

3.7 Robustness Across Alternative Empirical Designs and Inflation Measures

The baseline contribution estimates for the GDP deflator are constructed in a relatively parsimonious environment. To assess whether the positive contribution of AI to inflation is specific to that setup, we consider three complementary empirical designs that alter the mapping from AI movements to inflation while keeping the underlying accounting logic unchanged. Table 3 reports the resulting contribution for the GDP deflator.

The first alternative design is a Factor-Augmented Vector Autoregression (FAVAR)-lite exercise. In a full FAVAR, a large macroeconomic information set is summarized by a small number of latent factors, and the variables of interest evolve jointly with those factors in a vector-autoregressive system. The version implemented here is called "lite" because it uses the factor-summary logic of a FAVAR without estimating a full structural FAVAR system. Specifically, the relevant AI information block is summarized by its first principal component, and the quarterly movement in that factor is used as the AI regressor in the contribution projection. This approach is useful because it reduces idiosyncratic noise in individual AI indicators and emphasizes the common AI-related movement across the underlying components. Appendix B provides the precise implementation.

Case	Variant	Mean	CI	Median	Cum.	Pos. share	N
FAVAR-lite	Investment: Excluding Data Centers	0.633	(0.515, 0.751)	0.567	12.424	89.2	74
	Investment: Including Data Centers	1.049	(0.855, 1.244)	0.938	12.827	95.7	46
	Capital Services	1.137	(1.017, 1.257)	0.991	23.414	100.0	74
	Patents	0.003	(-0.002, 0.007)	0.001	0.044	50.7	67
	Data Centers	0.441	(0.296, 0.585)	0.282	5.199	82.6	46
	Capital Services & Patents	0.008	(0.000, 0.016)	0.006	0.137	58.2	67
	Composite: Excluding Data Centers & Capital Services	0.229	(0.121, 0.336)	0.236	2.255	76.9	39
	Composite: Excluding Data Centers	0.027	(0.010, 0.045)	0.022	0.456	59.7	67
	Composite: Including Data Centers	0.426	(0.297, 0.555)	0.423	4.240	84.6	39
TVP Rolling	Investment: Excluding Data Centers	0.822	(0.573, 1.070)	0.247	11.959	92.7	55
	Investment: Including Data Centers	1.950	(1.475, 2.425)	1.646	14.070	96.3	27
	Capital Services	2.171	(1.712, 2.630)	1.750	34.788	90.9	55
	Patents	0.001	(-0.025, 0.027)	0.011	0.015	56.2	48
	Data Centers	1.061	(0.708, 1.414)	0.607	7.422	88.9	27
	Capital Services & Patents	0.013	(-0.010, 0.037)	0.017	0.183	60.0	55
	Composite: Excluding Data Centers & Capital Services	0.018	(-0.007, 0.043)	0.021	0.245	67.3	55
	Composite: Excluding Data Centers	0.040	(0.013, 0.066)	0.049	0.549	69.1	55
	Composite: Including Data Centers	0.212	(0.143, 0.281)	0.204	1.443	77.8	27
Local Projections	Investment: Excluding Data Centers	0.489	(0.398, 0.581)	0.438	9.471	89.2	74
	Investment: Including Data Centers	0.982	(0.805, 1.160)	0.856	11.959	95.7	46
	Capital Services	3.523	(3.151, 3.895)	3.069	91.884	100.0	74
	Patents	-0.022	(-0.058, 0.014)	-0.006	-0.366	49.3	67
	Data Centers	0.517	(0.347, 0.686)	0.330	6.122	82.6	46
	Capital Services & Patents	0.017	(0.005, 0.029)	0.018	0.313	63.5	74
	Composite: Excluding Data Centers & Capital Services	0.038	(0.011, 0.066)	0.031	0.714	63.5	74
	Composite: Excluding Data Centers	0.088	(0.046, 0.131)	0.078	1.649	66.2	74
	Composite: Including Data Centers	0.174	(0.104, 0.244)	0.214	2.017	76.1	46

Table 3: Robustness across alternative empirical designs for the GDP deflator. The table reports summary statistics for the implied AI contribution series across AI variants and alternative empirical mappings from AI movements to inflation dynamics. Mean and median are in annualized percentage points (q/q annualized). Cumulative is log-compounded: $100\,(\exp(\sum_t c_t/400) - 1)$. Pos. share: the share (in %) of quarters with positive contributions within the window. The 90% CI is for the windowed mean contribution and is computed as $\bar{c} \pm t_{1-0.100/2,\,N-1}\, s/\sqrt{N}$.

For the GDP deflator, the FAVAR-lite design delivers contribution estimates that are similar in magnitude and ordering to the baseline. For example, for the comprehensive AI composite that includes data centers, the alternative design implies an average contribution of about 0.43 annualized percentage points with a high positive-share statistic, compared with about 0.20 percentage points in the benchmark case. Investment-based measures also remain sizable, and narrow patent-based measures remain small in magnitude. The pattern across variants is therefore stable: investment-heavy measures yield large positive contributions, capital-services measures are larger, and narrow innovation-based measures remain close to zero.

The second design is a rolling or time-varying implementation that re-estimates the mapping from AI movements to inflation over moving windows. This design allows the response of inflation to AI to drift over time as the economy evolves, as AI diffuses across sectors, or as the composition of AI-related spending changes. For the GDP deflator, the resulting contributions remain positive for the main investment and composite variants. The comprehensive composite including data centers yields an average contribution of roughly 0.21 annualized percentage points, with a positive share close to 80%. Investment measures again show substantial positive contributions, and the capital-services-only variant implies even larger contributions. This suggests that allowing for time variation in the mapping does not overturn the basic conclusion that AI has made a positive contribution to aggregate inflation when measured broadly.

The third design is a local-projection approach that explicitly incorporates the dynamic propagation of AI movements over a fixed horizon. Rather than focusing only on the contemporaneous response, this design accumulates the effect of a given AI movement over several quarters and translates that accumulation into a contribution series. Because the construction embeds propagation, the resulting average contributions are naturally larger in level than those from the more impact-oriented benchmark for some variants, though not uniformly for all of them. For the GDP deflator, the dynamic local-projection design implies that the investment measure excluding data centers contributes on average about 0.49 annualized percentage points, while the comprehensive composite including data centers contributes about 0.17 percentage points. The capital-services variant is much larger, at about 3.52 percentage points. What matters for robustness is that the sign, the ranking across AI variants, and the high positive-share statistics are preserved.

The robustness analysis is not confined to the GDP deflator. Table 4 reports the same contribution statistics for two alternative inflation concepts. The first is the Personal Consumption Expenditures (PCE) Price Index, a broad national-accounts measure of consumer prices that incorporates imputed and third-party expenditures and allows expenditure weights to adjust as households substitute across goods and services. The second is the Consumer Price Index (CPI), which tracks out-of-pocket consumer prices using a detailed consumption basket and distinct weighting and housing conventions. Because these indices differ in coverage and construction, they provide a strong test of whether the AI contribution is a narrow artifact of a particular measurement system or a more general macroeconomic phenomenon.

For PCE inflation, the baseline environment already yields a positive contribution from the comprehensive AI composite including data centers, with an average around 0.21 annualized percentage points and a positive share near three quarters of the sample. The corresponding composite in the benchmark CPI table is somewhat larger, around 0.29 percentage points, with a similar positive share. These levels are only modestly higher than the 0.20 percentage point contribution for the GDP deflator, indicating that the basic magnitude of the comprehensive contribution is relatively stable across the three broad indices. Adding observed controls has little effect on this conclusion: the contribution of the comprehensive composite remains positive for both PCE and CPI, with means only slightly below the benchmark and positive shares still well above one half. This suggests that the positive AI contribution is not being driven by omitted short-run covariates that could be captured by simple control sets.

The comparison between PCE and CPI is informative for understanding incidence and transmission. PCE, with its national-accounts coverage and substitution-sensitive weighting, tends to smooth category-level spikes and emphasize broad expenditure-weighted price movements. CPI, by contrast, is more sensitive to goods prices and to the details of housing and other consumer categories. The fact that the comprehensive AI composite generates positive contributions in both indices suggests that the AI–inflation link is not confined to a narrow set of categories or to peculiarities of national-accounts measurement. At the same time, the somewhat larger CPI contributions are plausible given CPI's greater sensitivity to the goods and supply-chain components through which AI-related investment and infrastructure may exert short-run pricing pressure.

Personal Consumption Expenditures (PCE)

Case	Variant	Mean	CI	Median	Cum.	Pos. share	N
Benchmark	Investment: Excluding Data Centers	0.873	(0.710, 1.036)	0.782	17.526	89.2	74
	Investment: Including Data Centers	1.061	(0.865, 1.258)	0.948	12.981	95.7	46
	Capital Services	1.019	(0.910, 1.128)	0.884	20.740	100.0	74
	Patents	-0.015	(-0.039, 0.010)	-0.004	-0.245	49.3	67
	Capital Services & Patents	-0.015	(-0.026, -0.005)	-0.016	-0.279	36.5	74
	Composite: Excluding Capital Services	0.018	(0.008, 0.028)	0.018	0.332	66.2	74
	Composite: Excluding Data Centers	0.016	(0.008, 0.023)	0.014	0.292	66.2	74
	Composite: Including Data Centers	0.214	(0.128, 0.300)	0.264	2.491	76.1	46
Controls	Investment: Excluding Data Centers	0.260	(0.212, 0.309)	0.233	4.934	89.2	74
	Investment: Including Data Centers	0.371	(0.303, 0.440)	0.332	4.362	95.7	46
	Capital Services	0.884	(0.789, 0.978)	0.767	17.760	100.0	74
	Patents	-0.004	(-0.010, 0.002)	-0.001	-0.061	49.3	67
	Capital Services & Patents	0.007	(0.002, 0.013)	0.008	0.137	63.5	74
	Composite: Excluding Capital Services	0.025	(0.011, 0.039)	0.026	0.465	66.2	74
	Composite: Excluding Data Centers	0.016	(0.008, 0.024)	0.014	0.298	66.2	74
	Composite: Including Data Centers	0.184	(0.110, 0.258)	0.226	2.135	76.1	46

Consumer Price Index (CPI)

Case	Variant	Mean	CI	Median	Cum.	Pos. share	N
Benchmark	Investment: Excluding Data Centers	1.095	(0.891, 1.300)	0.981	22.461	89.2	74
	Investment: Including Data Centers	1.353	(1.103, 1.604)	1.209	16.839	95.7	46
	Capital Services	1.075	(0.960, 1.190)	0.933	22.006	100.0	74
	Patents	-0.017	(-0.046, 0.011)	-0.004	-0.287	49.3	67
	Capital Services & Patents	-0.021	(-0.036, -0.007)	-0.022	-0.389	36.5	74
	Composite: Excluding Capital Services	0.005	(0.002, 0.008)	0.005	0.093	66.2	74
	Composite: Excluding Data Centers	0.017	(0.009, 0.025)	0.015	0.309	66.2	74
	Composite: Including Data Centers	0.294	(0.175, 0.412)	0.362	3.436	76.1	46
Controls	Investment: Excluding Data Centers	0.272	(0.221, 0.323)	0.244	5.160	89.2	74
	Investment: Including Data Centers	0.396	(0.323, 0.469)	0.354	4.661	95.7	46
	Capital Services	0.865	(0.772, 0.958)	0.751	17.354	100.0	74
	Patents	-0.003	(-0.008, 0.002)	-0.001	-0.051	49.3	67
	Capital Services & Patents	0.008	(0.002, 0.013)	0.008	0.146	63.5	74
	Composite: Excluding Capital Services	0.018	(0.008, 0.027)	0.018	0.328	66.2	74
	Composite: Excluding Data Centers	0.017	(0.009, 0.025)	0.015	0.319	66.2	74
	Composite: Including Data Centers	0.227	(0.136, 0.319)	0.280	2.651	76.1	46

Table 4: Summary of AI contribution series across cases and AI variants (windowed mean and 90% CI; median reported). Mean/median are in annualized percentage points. Cumulative is log-compounded: $100\,(\exp(\sum_t c_t/400) - 1)$. Pos. share: the share (in %) of quarters with positive contributions to inflation; $\widehat{c}^{AI} > 0$ within the window.

Taken together, the GDP deflator evidence across benchmark and alternative empirical designs, together with the PCE and CPI robustness results, points to a coherent conclusion. Once AI is measured in a comprehensive way that includes data centers and related investment, its contribution to aggregate inflation is consistently positive, with average magnitudes on the order of a few tenths of a percentage point per year in the impact-style designs and larger dynamic contributions once propagation is explicitly embedded. The ranking across AI variants is stable and economically interpretable, and the basic qualitative message survives substantial changes in econometric architecture, time-variation assumptions, and the underlying inflation concept.

4 Exposure-Based Sectoral Evidence

This section presents sector-level evidence on whether industries with greater ex ante exposure to AI-relevant task content experienced differential producer-price inflation after the 2022Q4 generative-AI shock. The empirical analysis uses three complementary estimators: a two-way fixed-effects difference-in-differences specification, an instrumental-variables difference-in-differences specification based on a Bartik-style shift-share instrument, and a Callaway–Sant'Anna group-time average treatment effect, following Callaway and Sant'Anna (2021). The estimators differ in the source of identifying variation and in how they treat dynamic and heterogeneous effects, but all are designed to ask the same basic question: whether producer-price inflation rose more in industries that were more exposed, before the shock, to AI-relevant tasks.

The dependent variable is sectoral producer-price inflation, denoted by π_{it}, where i indexes industries and t indexes quarters. It is constructed at the 6-digit North American Industry Classification System (NAICS) industry level.

4.1 Sectoral AI exposure

The treatment-intensity measure is a task-based AI exposure score constructed at the industry level. The construction proceeds in two steps. First, occupation-level exposure scores are built from O*NET Work Activities data. Second, these occupation-level scores are aggregated to the industry level using pre-period occupational employment shares.

We begin with four O*NET work-activity elements that are closely related to the cognitive and analytical tasks most directly affected by modern AI systems: *Thinking Creatively* (4.A.2.a.4), *Getting Information* (4.A.2.a.1), *Processing Information* (4.A.2.b.2), and *Analyzing Data or Information* (4.A.4.a.8). For each detailed O*NET occupation, we use the O*NET importance rating for each element. These ratings range from 1 to 5, with higher values indicating that the activity is more important for the occupation.

The industry employment data are organized by Standard Occupational Classification (SOC) occupation groups. The baseline specification uses 2-digit SOC groups. This level of aggregation is useful because it captures broad occupational task content while avoiding excessive noise from very fine occupation-level matching. Since the OES/OEWS employment shares must be linked to O*NET task scores, the 2-digit SOC aggregation provides a stable mapping between detailed task information and industry employment structure. Later robustness checks vary the SOC aggregation level to verify that the results are not driven by this baseline choice.

Let $\mathcal{K}(o)$ denote the set of detailed O*NET occupations mapped to SOC group o, and let $IM_{k,e}$ denote the O*NET importance rating for detailed occupation k and work-activity element e. The raw task-based AI score for SOC group o, denoted by AI_o^r, is computed as the average importance rating across the selected work-activity elements and across the detailed O*NET occupations mapped into that SOC group:

$$\mathrm{AI}_o^r = \frac{1}{|\mathcal{K}(o)|} \sum_{k \in \mathcal{K}(o)} \frac{1}{4} \left(IM_{k,4.A.2.a.4} + IM_{k,4.A.2.a.1} + IM_{k,4.A.2.b.2} + IM_{k,4.A.4.a.8} \right).$$

This score is high for occupation groups in which creative thinking, information acquisition, information processing, and data analysis are important work activities. The interpretation is that occupation groups relying more heavily on these activities are more exposed to AI because such tasks are precisely the types of cognitive and analytical activities that generative AI systems can assist, automate, or transform.

The raw SOC-level scores are then min–max normalized across SOC groups:

$$\mathrm{AI}_o = \frac{\mathrm{AI}_o^r - \min_{o'} \mathrm{AI}_{o'}^r}{\max_{o'} \mathrm{AI}_{o'}^r - \min_{o'} \mathrm{AI}_{o'}^r}.$$

Here, o is the SOC group whose exposure score is being normalized, while o' indexes the SOC groups over which the cross-occupation minimum and maximum are computed. The normalization maps the SOC-level exposure scores into the unit interval, so that $\mathrm{AI}_o \in [0, 1]$. A value closer to one indicates high AI-relevant task content relative to other occupation groups, while a value closer to zero indicates lower exposure.

The SOC-level task scores are then aggregated to the industry level using 2019 OES/OEWS employment shares as weights. Denoting by s_{io} the share of total employment in industry i accounted for by SOC group o, the industry-level exposure measure is

$$\mathrm{AIExp}_i = \sum_o s_{io}\, \mathrm{AI}_o.$$

This variable is time-invariant by construction. It captures the occupational composition of each industry at the 2019 baseline and measures how exposed that industry was, before the 2022Q4 AI shock, to tasks most susceptible to AI augmentation. The 2019 employment shares are useful because they predate the 2022Q4 event and avoid contamination from post-event AI-induced restructuring. They also predate the peak pandemic period, reducing concerns that the exposure measure reflects temporary COVID-era labor-market dislocations rather than underlying task composition.

4.2 The AI exposure distribution and treatment groups

Figure 4 displays the distribution of AIExp_i across the $N = 380$ 6-digit NAICS sectors in the estimation sample, sorted in ascending order. The bottom 20% of the distribution serves as the low-exposure control group, and the top 20% serves as the high-exposure treated group; the middle 60% is excluded from the baseline symmetric comparison. The horizontal dashed lines indicate the 20th-percentile threshold, at approximately an exposure score of 0.13, and the 80th-percentile threshold, at approximately 0.34.

Several features of the figure are economically informative. First, the distribution is right-skewed: most sectors cluster below a score of 0.25, while a smaller group reaches scores above 0.40. Second, there is meaningful cross-sectional variation, with exposure scores spanning roughly 0.07 to 0.63. Third, the gap between the top-20% treated group and the bottom-20% control group is economically substantial, approximately 0.20 score units on a unit-interval scale. This dispersion provides the identifying variation for the difference-in-differences estimators.

4.3 Treatment definition

Let D_i denote a binary indicator equal to one if sector i belongs to the top 20% of the AIExp distribution and zero if it belongs to the bottom 20%, so that $D_i \in \{0, 1\}$ for sectors in the baseline estimation sample. The post-event indicator is

$$\mathrm{Post}_t = \mathbf{1}\{t \geq t_0\},$$

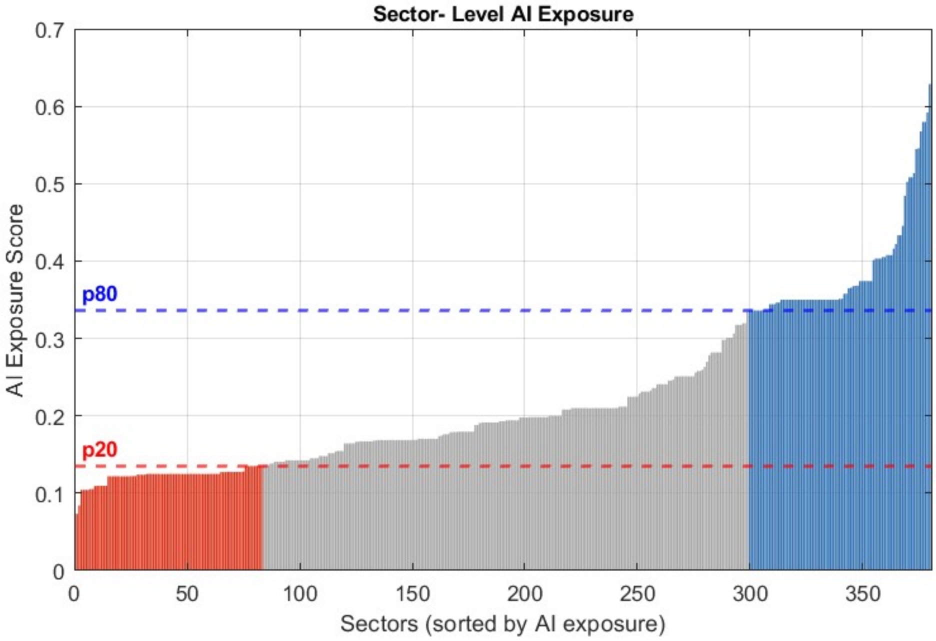


Figure 4: Distribution of sectoral AI task exposure scores. The figure displays the task-based AI exposure score AIExp_i for all $N = 380$ sectors in the estimation sample, sorted in ascending order. Red bars indicate sectors in the bottom 20% of the exposure distribution; blue bars indicate sectors in the top 20%; grey bars indicate the excluded middle 60%. The red and blue dashed horizontal lines mark the 20th- and 80th-percentile thresholds, respectively. Scores are constructed from O*NET Work Activities data aggregated to the 6-digit NAICS level using 2019 OES/OEWS employment weights.

where t_0 denotes the event quarter, here 2022Q4. The standard static treatment interaction is $D_i \times \mathrm{Post}_t$.

4.4 Two-way fixed-effects specification

The baseline specification is a two-way fixed-effects model that interacts the high-AI-exposure treatment indicator with the post-2022Q4 indicator:

$$\pi_{it} = \beta^{\mathrm{TWFE}}\big(D_i \times \mathrm{Post}_t\big) + \boldsymbol{\delta}'\mathbf{X}_{it} + \alpha_i + \lambda_t + f_i(t) + \varepsilon_{it},$$

where α_i are sector fixed effects, λ_t are quarter fixed effects, $f_i(t)$ are sector-specific linear time trends, and $\mathbf{X}_{it}$ is a vector of time-varying controls. The controls include one lag of π_{it}, upstream PPI inflation, an energy-price interaction, an aggregate demand interaction, and a digital/robots/ICT intensity proxy constructed from O*NET Knowledge data.

The coefficient β^{TWFE} is measured in percentage points of year-over-year PPI inflation. It compares the average change in post-event inflation for high-exposure sectors relative to low-exposure sectors after removing permanent sector differences, common quarterly shocks, sector-specific linear trends, and the included controls. A positive estimate means that high-AI-exposure sectors experienced higher PPI inflation after 2022Q4 than low-exposure sectors, relative to their pre-event difference. Under the parallel trends assumption—that, absent the AI shock, the treated and control sectors would have followed similar inflation trajectories after conditioning on the fixed effects, trends, and controls—β^{TWFE} estimates the average post-event inflation differential associated with belonging to the high-AI-exposure group. Standard errors are clustered at the sector level.[2]

4.5 IV-DiD specification

The two-way fixed-effects specification may still be biased if high-exposure sectors experience correlated shocks beyond AI. For example, task exposure may be correlated with unobserved demand conditions, sectoral markups, supply-chain pressures, or other industry characteristics that also affect producer-price inflation. To address this concern, we estimate an IV-style Bartik specification that uses the interaction between predetermined industry AI exposure and aggregate AI progress as a source of variation in effective AI treatment.

A time-varying aggregate AI series $\Delta \mathrm{AI}_t$ is constructed from the selected quarterly AI composite index and cumulated over time:

$$\mathrm{CumAI}_t = \sum_{\tau \leq t} \Delta \mathrm{AI}_\tau.$$

This series plays the role of the aggregate shift in the Bartik design. In the baseline implementation, we use the AI composite that excludes data centers, described in Section 2.

The predetermined exposure component is the industry-level AI exposure measure, AIExp_i. The sector-time Bartik variable is

$$\mathrm{Bartik}_{it} = \mathrm{AIExp}_i \times \mathrm{CumAI}_t.$$

To avoid confusion with the AI measures denoted elsewhere by Z, the instrument is denoted by S_{it}:

$$S_{it} = \mathrm{Bartik}_{it} \times \mathrm{Post}_t = \big(\mathrm{AIExp}_i \times \mathrm{CumAI}_t\big) \times \mathrm{Post}_t.$$

The instrument is therefore zero before the event and varies after the event according to both the aggregate AI trajectory and the degree to which each industry was exposed ex ante to AI-relevant tasks.

The first stage is

$$D_i \times \mathrm{Post}_t = \eta_S S_{it} + \mathbf{\Gamma}' \mathbf{X}_{it} + \alpha_i + \lambda_t + f_i(t) + \nu_{it}.$$

The coefficient η_S measures whether the shift-share variable predicts the post-period treatment interaction. A strong first stage means that industries with greater pre-period AI exposure receive a larger predicted treatment when aggregate AI progress rises.

The second stage is

$$\pi_{it} = \beta^{\mathrm{IV}} \, (\widehat{D_i \times \mathrm{Post}_t}) + \boldsymbol{\delta}' \mathbf{X}_{it} + \alpha_i + \lambda_t + f_i(t) + \varepsilon_{it}.$$

The coefficient β^{IV} measures the inflation differential associated with the component of post-period high-exposure treatment predicted by the interaction between predetermined sectoral exposure and aggregate AI progress. Because the endogenous variable remains the post-period treatment interaction, the coefficient is still interpretable in percentage points of year-over-year PPI inflation for the high-exposure treatment margin. The difference from TWFE is the source

[2] Using heteroskedasticity- and autocorrelation-consistent standard errors with four Bartlett lags yields very similar conclusions. The point estimates are unchanged, and the statistical significance of the main TWFE and IV-DiD estimates is preserved.

of identifying variation: IV-DiD uses only the part of the treatment contrast induced by the shift-share interaction between ex ante task exposure and aggregate AI progress.

The identifying assumption is an exclusion restriction. Conditional on sector fixed effects, quarter fixed effects, sector-specific trends, and controls, S_{it} should affect sectoral PPI inflation only through effective AI treatment intensity. Quarter fixed effects absorb the aggregate AI trajectory itself, together with other common macroeconomic shocks. The identifying variation therefore comes from the differential exposure of industries to the common aggregate AI movement, allocated according to predetermined task exposure. Inference for the IV-DiD estimates is based on sector-clustered standard errors.

4.6 Callaway–Sant'Anna group-time ATT

As an additional robustness exercise, we compute group-time average treatment effects using the logic of Callaway and Sant'Anna (2021). The Callaway–Sant'Anna approach is designed for difference-in-differences settings in which treatment effects may vary across groups and over time. Rather than imposing a single constant post-treatment effect, it constructs group-time treatment effects by comparing treated units to an appropriate comparison group at each event time and then aggregates those effects into an overall ATT.

In the present application, there is a single treated cohort, industries assigned $G_i = t_0$, corresponding to the top-20% AI-exposure group, and an untreated comparison group, $G_i = \infty$, corresponding to the bottom-20% group. The general Callaway–Sant'Anna framework therefore simplifies to a cohort-specific comparison between treated and never-treated industries at each event time.

Prior to computing the group-time ATT, the outcome is residualised with respect to the selected controls, fixed effects, and trends. Denote the residualised outcome by π_{it}^{CS}:

$$\pi_{it}^{CS} = \pi_{it} - \widehat{\boldsymbol{\delta}'\mathbf{X}_{it}} - \widehat{\alpha}_i - \widehat{\lambda}_t - \widehat{f_i(t)}.$$

This residualisation removes the component of sectoral inflation explained by observed controls, permanent industry differences, common time effects, and industry-specific trends before treated and control industries are compared.

Let the average residualised inflation rate of the treated group and the control group in quarter t be

$$\bar{\pi}_{g,t}^{CS} = \frac{1}{N_g} \sum_{i:G_i=t_0} \pi_{it}^{CS}, \qquad \bar{\pi}_{c,t}^{CS} = \frac{1}{N_c} \sum_{i:G_i=\infty} \pi_{it}^{CS},$$

with g denoting the treated group and c denoting the control group. The pre-event baseline window is

$$\mathcal{B} = [t_0 + h_{\min},\, t_0 + h_{\max}],$$

which in the baseline implementation spans quarters -6 through -1 relative to the event. Baseline means for the treated and control groups are

$$\bar{\pi}_{g,\mathcal{B}}^{CS} = \frac{1}{|\mathcal{B}|} \sum_{t\in\mathcal{B}} \bar{\pi}_{g,t}^{CS}, \qquad \bar{\pi}_{c,\mathcal{B}}^{CS} = \frac{1}{|\mathcal{B}|} \sum_{t\in\mathcal{B}} \bar{\pi}_{c,t}^{CS}.$$

For each quarter t, the group-time ATT is

$$\mathrm{ATT}(t) = \left(\bar{\pi}_{g,t}^{CS} - \bar{\pi}_{g,\mathcal{B}}^{CS}\right) - \left(\bar{\pi}_{c,t}^{CS} - \bar{\pi}_{c,\mathcal{B}}^{CS}\right).$$

This estimand compares the change in residualised inflation for treated industries from the pre-event baseline to period t with the corresponding change for never-treated industries. It therefore measures the event-time differential in sectoral PPI inflation associated with high AI exposure after removing the controls, fixed effects, and trends included in the residualisation step.

The sequence of group-time estimates $\{\mathrm{ATT}(h)\}$ is mapped to relative event time and then averaged over post-event quarters:

$$\mathrm{ATT}^{CS}_{\text{overall}} = \frac{1}{H_{\text{post}}} \sum_{h>0} \mathrm{ATT}(h),$$

where H_{post} is the number of post-event quarters included in the average. The overall CS ATT is therefore measured in percentage points of year-over-year PPI inflation and reports the average post-event differential inflation response of high-AI-exposure industries relative to low-AI-exposure industries, evaluated relative to the six-quarter pre-event baseline. Standard errors are based on sector-level variation in the treated and control group changes between the current and baseline periods.

4.7 Main results

Table 5 reports estimates from all three estimators for the 2022Q4 event. The sample comprises 165 sectors observed over 11,573 sector-quarter observations, with the treatment contrast defined as the top 20% versus the bottom 20% of the AI task exposure distribution.

Across all three specifications, the estimated effect of AI task exposure on producer-price inflation is positive, economically meaningful, and statistically significant. Before discussing magnitudes, it is worth emphasising what a positive coefficient means in this setting. It means that industries in which workers spend a disproportionate share of their time on analytical, information-processing, and reasoning tasks—the activities most directly augmented by large language models and generative AI—saw their output prices rise more rapidly in the quarters following the ChatGPT shock than otherwise comparable low-exposure industries. These patterns are consistent with a supply-side interpretation: the estimated inflation differential is obtained after controlling for demand-output interactions, upstream price pressures, energy interactions, sector fixed effects, quarter fixed effects, and sector-specific trends. The evidence is therefore difficult to attribute solely to differential demand. It is most naturally read as reflecting cost and adjustment channels associated with AI adoption, although the empirical design does not separately identify every mechanism.

TWFE. The TWFE estimate is 0.718 percentage points, significant at the 5% level. Conditioning on sector fixed effects, quarter fixed effects, sector-specific linear time trends, lagged inflation, upstream PPI, energy price interactions, demand-output interactions, and a digital and ICT intensity proxy, high-AI-exposure sectors experienced approximately three-quarters of a percentage point more annual PPI inflation than low-exposure sectors in the post-2022Q4 period. To put this in context, the average annual PPI inflation rate in the estimation sample over the post-event window is of the order of 2–4 percentage points, so the estimated differential represents a meaningful share—roughly 20–35%—of typical sectoral inflation variation. This is not a marginal effect: it corresponds to a substantively different inflation trajectory for industries at the top of the AI task exposure distribution relative to those at the bottom.

The two-way fixed-effects structure is important for the credibility of this estimate. Sector fixed effects ensure that the comparison is not contaminated by persistent differences in price-setting behaviour, markup levels, or structural cost differentials across industries. Quarter fixed effects absorb any aggregate macro shock—monetary policy tightening, global commodity price movements, supply-chain disruptions—that affects all sectors simultaneously. The sector-specific linear trends further control for the possibility that high-exposure sectors were already on a diverging inflation trajectory before the AI shock, a concern that is particularly acute in a period that includes the post-COVID inflationary surge of 2021–2022.

IV-DiD. The IV-DiD estimate of 0.765 percentage points, significant at the 5% level, is slightly larger than the TWFE counterpart. The Bartik shift-share instrument provides strong first-stage leverage, with a sector-clustered first-stage F-statistic of 427.24, far above conventional weak-instrument thresholds. This makes weak-instrument bias unlikely to be the main driver of the IV estimate, though the validity of the IV estimate still depends on the exclusion restriction.

The IV design also mitigates a subtler identification concern: the possibility that the sector-level task composition is correlated with unobserved industry characteristics that independently predict post-2022 inflation. For instance, if high-task-exposure industries systematically overlap with sectors experiencing idiosyncratic demand booms or import competition relief in the post-ChatGPT period, TWFE would conflate the AI effect with these confounders. The Bartik instrument uses the differential interaction of predetermined task shares with the aggregate AI trajectory, thereby isolating the component of treatment variation that is driven by the combination of ex ante industry structure and the economy-wide AI shock rather than by sector-specific shocks. The near-equality of the TWFE and IV estimates thus also speaks to the robustness of the TWFE specification: the controls and fixed effects appear to absorb much of the potential confounding variation, leaving the Bartik variation and the TWFE variation telling a consistent story.

Callaway–Sant'Anna ATT. The Callaway–Sant'Anna overall average treatment effect is 1.085 percentage points and is statistically significant at the 10% level. Its being larger than the corresponding TWFE estimate is consistent with treatment-effect heterogeneity and with the possibility that the inflation response to AI exposure strengthened over the post period. When treatment effects evolve over time rather than remaining constant, TWFE can aggregate group-time effects in ways that attenuate the average post-treatment effect, especially when dynamic effects differ across groups or horizons. By contrast, the Callaway–Sant'Anna estimator constructs group-time average treatment effects directly and then aggregates them transparently, making it better suited to settings in which post-treatment effects vary over time. The larger CS estimate is therefore consistent with a setting in which the inflationary impact of AI exposure intensified over the quarters following the shock rather than remaining constant throughout the post period. This interpretation is economically plausible, given the gradual diffusion of AI technologies and the time required for firms to adjust production processes, organizational structures, and complementary inputs.

The CS estimate is obtained on an outcome that has been residualised with respect to the same controls, fixed effects, and trends as the TWFE and IV specifications, so it is not a comparison of raw inflation levels. It reflects the average post-event inflation differential between the top-20% and bottom-20% groups after removing the shared confounders, evaluated relative to the six-quarter pre-event baseline window of quarters -6 through -1. Also, as

	TWFE	IV-DiD	CS Overall ATT
Estimate	0.718**	0.765**	1.085*
	(0.283)	(0.329)	(0.653)
First-stage F	–	427.24	–
Sector FE	Yes	Yes	Residualised
Quarter FE	Yes	Yes	Residualised
Controls	Yes	Yes	Yes
Sector trends	Sector poly. (1)	Sector poly. (1)	Sector poly. (1)
Lagged inflation	Yes	Yes	Yes
Dummies	No	No	No
Observations	11,573	11,573	11,573
Sectors	165	165	165
Event quarter		2022Q4	
Treatment definition		Top 20% vs. Bottom 20% AI task exposure	
CS baseline window		$[-6, -1]$ quarters relative to event	

Table 5: AI task exposure and sectoral PPI inflation: main estimates. This table reports estimates from three complementary estimators. TWFE is a two-way fixed-effects difference-in-differences regression. IV-DiD instruments the post-period treatment interaction with a Bartik shift-share variable constructed as the industry's AI task exposure score interacted with a cumulative aggregate AI index. CS Overall ATT is the Callaway–Sant'Anna (2021) average treatment effect averaged over all post-event quarters, computed on an outcome residualised with respect to the fixed effects, trends, and controls listed above. AI task exposure is constructed at the NAICS 6-digit industry level using 2019 OES/OEWS employment weights and O*NET Work Activities importance scores for four analytical and information-processing task elements. Standard errors in parentheses are clustered at the sector level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

shown in Appendix Figure C.1, the positive overall ATT is not driven by a small number of extreme positive observations. Instead, the distribution is shifted to the right, with the vast majority of treated high-exposure sectors displaying positive differential post-event effects relative to their pre-event baseline. This pattern strengthens the interpretation that the CS estimate reflects a broad-based inflationary response among high-exposure sectors rather than an average dominated by a few outliers.

Perhaps the most important finding in Table 5 is the consistency of the evidence across the three approaches. TWFE, IV-DiD, and Callaway–Sant'Anna use different identifying variation and handle treatment-effect heterogeneity differently. The fact that all three converge to a range of roughly 0.7–1.1 percentage points, with the same sign and qualitatively similar magnitudes, indicates that the finding is not an artefact of any single estimator.

4.8 Asymmetric treatment design: robustness and broader inference

Table 6 reports estimates under an asymmetric treatment definition, in which the control group is expanded from the bottom 20% to the full bottom 80% of the AI task exposure distribution. This expands the estimation sample from 165 to 382 sectors and from 11,573 to 26,871 sector-quarter observations. The comparison changes from a high-versus-low extreme contrast to a high-versus-typical contrast. This design is useful both as a robustness check and as a broader policy-relevant comparison between the most AI-exposed industries and the rest of the economy.

	TWFE	IV-DiD	CS Overall ATT
Estimate	0.541**	0.901***	0.967*
	(0.240)	(0.345)	(0.611)
First-stage F	–	535.79	–
Sector FE	Yes	Yes	Residualised
Quarter FE	Yes	Yes	Residualised
Controls	Yes	Yes	Yes
Sector trends	Sector poly. (1)	Sector poly. (1)	Sector poly. (1)
Lagged inflation	Yes	Yes	Yes
Dummies	No	No	No
Observations	26,871	26,871	26,871
Sectors	382	382	382
Event quarter	2022Q4		
Treatment definition	Top 20% vs. Bottom 80% AI task exposure		
CS baseline window	$[-6, -1]$ quarters relative to event		

Table 6: AI task exposure and sectoral PPI inflation: top 20% vs. bottom 80%. This table reports estimates from three complementary estimators under the asymmetric treatment definition, in which all sectors below the 80th percentile of the AI task exposure distribution serve as controls for the top-20% treated group. TWFE is a two-way fixed-effects difference-in-differences regression. IV-DiD instruments the post-period treatment interaction with a Bartik shift-share variable constructed as the industry's AI task exposure score interacted with a cumulative aggregate AI index. CS Overall ATT is the Callaway–Sant'Anna (2021) average treatment effect averaged over all post-event quarters, computed on an outcome residualised with respect to the fixed effects, trends, and controls listed above. AI task exposure is constructed at the 6-digit NAICS industry level using 2019 OES/OEWS employment weights and O*NET Work Activities importance scores for four analytical and information-processing task elements. Standard errors in parentheses are clustered at the sector level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The asymmetric estimates remain positive and statistically significant across all three estimators. The TWFE estimate is 0.541 percentage points, smaller than the symmetric estimate of 0.718 percentage points. This attenuation is expected. In the symmetric design, the control group consists only of the bottom-20% sectors, which provide the sharpest contrast with the high-exposure group. In the asymmetric design, the control group also includes the middle 60% of the exposure distribution. Since many middle-exposure sectors may themselves experience some AI-related cost pressure, the estimated contrast narrows. The positive and significant TWFE coefficient therefore indicates that the top-20% sectors exhibit higher post-event inflation even relative to a broad control group that includes many moderately exposed industries.

The IV-DiD estimate is 0.901 percentage points, significant at the 1% level, with a sector-clustered first-stage F-statistic of 535.79. The IV estimate is larger than the TWFE estimate, suggesting that the shift-share component of treatment variation isolates a stronger AI-related inflation differential than the average post-period high-exposure contrast. The instrument remains highly relevant, and the identifying logic is the same as in the symmetric design: sectors with higher predetermined exposure to analytical and information-processing tasks should respond more strongly to aggregate AI progress.

The Callaway–Sant'Anna overall ATT is 0.967 percentage points and is significant at the 10% level. Its proximity to the IV estimate suggests that allowing treatment effects to vary over time recovers a stronger post-event effect than the static TWFE coefficient. The distributional diagnostic discussed above is also relevant in this design: the positive CS estimate reflects a broad rightward shift among treated high-exposure sectors rather than a mean dominated by a small number of extreme values.

Taken together, Tables 5 and 6 tell a consistent story. The symmetric design delivers the sharpest treatment-control contrast by comparing the top and bottom tails of the exposure distribution. The asymmetric design delivers broader inference by comparing the high-exposure frontier to the rest of the economy. Both designs imply that AI task exposure is associated with higher producer-price inflation in the post-2022Q4 period, with estimates generally ranging from about one-half to just over one percentage point of annual PPI inflation.

4.9 Robustness

The baseline results are subjected to three further robustness checks, with supporting tables and figures collected in Appendix C. The first uses a median-split treatment definition. The second examines sensitivity to the level of industry and task aggregation. The third re-estimates the baseline specification using occupational employment shares from 2007.

4.9.1 Median-split treatment definition

Table C.1 reports estimates under a median-split treatment definition, in which the above-median half of the AI task exposure distribution is compared against the below-median half. This specification uses the full sample of 382 sectors and 26,871 sector-quarter observations. It asks whether the inflation differential remains detectable even when the treated and control groups are separated only by the median of the exposure distribution.

The estimates pass this test. The TWFE coefficient is 0.767 percentage points and is significant at the 1% level; the IV-DiD estimate is 0.832 percentage points and is significant at the 5% level; and the CS estimate is 0.892 percentage points and is significant at the 10% level. The magnitudes are strikingly consistent with those in Tables 5 and 6, despite the different construction of the treatment and control groups. This stability indicates that the estimated effect reflects a broader gradient in inflation dynamics along the AI exposure distribution rather than a contrast confined to the extreme tails.

4.9.2 Sensitivity to industry and task aggregation level

Table C.2 reports estimates across nine NAICS–SOC aggregation combinations. Moving from Panel A to Panel D, both the NAICS and SOC digit levels increase together from 2 to 5, yielding progressively finer industry classifications and larger estimation samples. Panels E through I hold the NAICS digit level fixed at 6 and vary the SOC digit level from 2 to 6, isolating sensitivity to the occupational aggregation used to match O*NET task scores to OES/OEWS employment weights.

The AI task exposure effect is positive in all specifications and statistically significant in the overwhelming majority of them. At coarser aggregation levels, the estimates are somewhat larger, reflecting stronger exposure contrasts when industries are defined more broadly. At the 6-digit NAICS level, the estimates generally remain positive and economically meaningful as the SOC digit level varies. The results therefore do not depend on a single NAICS–SOC matching level, although very fine occupation-level matching can introduce additional noise.

4.9.3 Occupational employment shares from 2007

The baseline exposure measure AIExp_i is constructed using 2019 OES/OEWS occupational employment shares. These shares are predetermined with respect to the 2022Q4 ChatGPT event, but they postdate earlier AI milestones considered elsewhere in the analysis. To address the concern that occupational composition had already adjusted in response to earlier AI waves by 2019, the exposure measure is recomputed using occupational employment shares from the 2007 OES/OEWS survey, the opening year of the estimation sample.

Figure C.2 displays the resulting distribution of AIExp_i. Compared with the 2019-based distribution, the 2007 distribution is lower and more compressed: the 20th-percentile threshold falls from approximately 0.14 to 0.11 and the 80th-percentile threshold falls from approximately 0.34 to 0.22. This is consistent with a progressive concentration of AI-relevant task content in high-exposure industries over the 2007–2019 period. The 2007 shares therefore provide a more conservative and more predetermined treatment contrast.

Tables C.3 and C.4 report the symmetric and asymmetric estimates obtained with 2007 shares. The results are qualitatively identical to the baseline and stronger in magnitude. Under the symmetric design, the TWFE estimate rises to 1.538 percentage points, the IV-DiD estimate rises to 1.411 percentage points, and the Callaway–Sant'Anna overall ATT rises to 2.181 percentage points, with all three estimates significant at the 1% level. The asymmetric design yields a TWFE estimate of 1.113 percentage points, an IV-DiD estimate of 1.015 percentage points, and a CS ATT of 1.906 percentage points, again all significant at the 1% level. The sector-clustered first-stage F-statistics of 278.78 and 390.30 confirm that the Bartik instrument remains strong under the 2007 shares.

The amplification of the estimates under 2007 shares has a natural interpretation. The 2007 shares capture each industry's task structure before any AI-related adjustment during the sample. By 2019, some of the cross-sectional variation in AI task content may already have been partially incorporated into industry outcomes, including prices, as firms in high-exposure industries adapted to earlier AI waves. To the extent this occurred, the 2019 exposure measure may attenuate the treatment contrast. The stronger estimates based on 2007 shares therefore reinforce the conclusion that the baseline results are, if anything, conservative.

The empirical results in this section deliver a consistent message. Industries whose task structure is most heavily weighted toward analytical, reasoning, and information-processing activities experienced meaningfully higher producer-price inflation after the 2022Q4 generative-AI shock. The estimated differential ranges from roughly 0.5 to 1.1 percentage points of annual PPI inflation across the main 2019-share treatment designs and estimators, and is larger when exposure is constructed from 2007 occupational employment shares. The results are robust to alternative identification strategies, alternative control group definitions, alternative approaches to treatment-effect heterogeneity, alternative levels of industry and task aggregation, and the use of pre-2007 occupational employment shares.

Economically, the finding is consistent with an adoption-cost and deployment-pressure mechanism. As firms in high-exposure industries integrate AI tools, they may face transitional costs related to software adoption, worker retraining, organizational restructuring, complementary capital investment, and input bottlenecks. These costs can be reflected in producer prices during the deployment phase. The evidence therefore points to a positive inflationary footprint of AI adoption, even if the longer-run productivity effects of AI may eventually operate in a disinflationary direction.

5 Static Analytical Model

We start from a static environment in which AI affects (i) firm-level productivity, (ii) market power, and (iii) the level of demand faced by the firm. Consider a monopolistically-competitive firm that faces a downward-sloping demand curve

$$y = \bar{y} - \sigma\,(p - \bar{p}) + \chi z, \qquad \sigma > 0, \tag{8}$$

where y is the log of output Y, p is the log of the price P, z is the log of AI intensity Z, and $\bar{p}$ and $\bar{y}$ are reference levels. The term χz captures a *demand channel* of AI: for $\chi > 0$, a higher level of AI intensity shifts the demand schedule outward at any given relative price. The parameter σ governs how sensitive demand is to relative price changes.

Output is produced using labor L and AI-enhanced productivity: $Y = \bar{A}\,Z^{\psi}L$, with $\psi > 0$ and $\bar{A}$ denoting a reference level of productivity (e.g. non-AI TFP). In logs,

$$y = \bar{a} + \psi z + \ell, \tag{9}$$

where $\ell = \log L$ and $\bar{a} = \log \bar{A}$.

With a competitive labor market and real wage W, total cost for producing Y units is $TC = WL$. Using the production function, $TC = W \cdot \frac{Y}{\bar{A}Z^{\psi}}$, and the baseline marginal cost $(\partial TC/\partial Y)$ is

$$MC_0(z) = \frac{W}{\bar{A}Z^{\psi}}. \tag{10}$$

Taking logs,

$$mc_0(z) = w - \bar{a} - \psi z. \tag{11}$$

Treating w and $\bar{a}$ as constants for this comparative-statics exercise, we can write

$$mc_0 = c_{mc} - \psi z, \tag{12}$$

where $c_{mc} \equiv w - \bar{a}$. This expression captures the *productivity channel*: higher AI raises productivity and thus lowers marginal cost.

To allow for congestion or capacity constraints, we let marginal cost rise with output. In log terms we write

$$mc = c_{mc} - \psi z + \lambda\,(y - \bar{y}), \qquad \lambda \geq 0, \tag{13}$$

so that higher sales push up marginal cost when $\lambda > 0$.

AI also affects market power. We let the desired markup to depend on AI:

$$\mu = c_{\mu} + \gamma z, \qquad \gamma \geq 0, \tag{14}$$

where μ is the markup and c_{μ} is a constant. The parameter γ captures a *direct market-power channel* of AI: for example, AI-enabled pricing or more precise screening of customers can raise desired markups even holding demand fixed.

The firm sets its price as a markup over marginal cost:

$$p = \mu + mc. \tag{15}$$

Substituting the expressions for μ and mc,

$$p = c + (\gamma - \psi)z + \lambda(y - \bar{y}), \tag{16}$$

where $c \equiv c_\mu + c_{mc}$. Using the demand schedule, we get

$$p = c + (\gamma - \psi + \lambda\chi)z - \lambda\sigma(p - \bar{p}). \tag{17}$$

Hence the equilibrium price level as a function of AI intensity can be written as

$$p = \tilde{c} + \frac{\gamma - \psi + \lambda\chi}{1 + \lambda\sigma}\, z, \tag{18}$$

where $\tilde{c} \equiv \big(c + \lambda\sigma\bar{p}\big)/(1 + \lambda\sigma)$ collects constants.

The effect of AI on the price level is therefore

$$\frac{\partial p}{\partial z} = \frac{\gamma - \psi + \lambda\chi}{1 + \lambda\sigma}. \tag{19}$$

Likewise, substituting p back into the demand curve, the response of output to AI is

$$\frac{\partial y}{\partial z} = \frac{\chi - \sigma\gamma + \sigma\psi}{1 + \lambda\sigma}. \tag{20}$$

These expressions make transparent how the different channels operate. Through the productivity effect, higher AI lowers marginal cost. Holding markups and demand conditions fixed, this pushes prices down and output up. On the other hand, AI can raise desired markups (e.g. via AI-enabled pricing, data-driven screening, or platform power), pushing prices up and output down. If $\gamma > \psi$, the net supply-side effect on the price level is positive.

The demand channel now works through the interaction of χ and λ. A higher level of AI shifts demand outward, which raises output; when marginal cost is increasing in output ($\lambda > 0$), this tighter capacity feeds back into higher marginal cost and thus higher prices, so that the term $\lambda\chi$ appears in the numerator of $\partial p/\partial z$. If $\lambda = 0$, marginal cost is flat and the pure demand shift affects quantities but not prices, so the demand channel drops out of the price equation.

The denominator $(1 + \lambda\sigma)$ reflects general-equilibrium dampening: when demand is very price-sensitive (σ large) or marginal cost is very responsive to output (λ large), changes in AI have more muted effects on prices because part of the pressure is absorbed by quantity adjustments and by the feedback between demand and costs.

Several limiting cases are informative. If the demand channel is shut down ($\chi = 0$), the impact of AI on prices simplifies to

$$\frac{\partial p}{\partial z} = \frac{\gamma - \psi}{1 + \lambda\sigma}, \tag{21}$$

which suggests that, in this case, AI is disinflationary if productivity gains dominate ($\psi > \gamma$) and inflationary if market-power effects dominate ($\gamma > \psi$). Similar analysis apply if the marginal cost is completely flat ($\lambda = 0$). Generally, even if the direct markup effect γ is modest, a strong demand shift combined with convex costs (large λ) can make $\gamma - \psi + \lambda\chi$ positive, so that AI becomes inflationary in spite of its cost-reducing effect. Conversely, if productivity gains are sufficiently large relative to both market power and demand-induced cost pressure, the net effect of AI on prices remains disinflationary. With all channels present, the productivity impact would need to be sufficiently strong to offset the myriad of inflationary channels.

6 Dynamic Framework: AI, Marginal Cost, and Inflation

This section develops a simple New Keynesian framework to illustrate how artificial intelligence affects inflation through five structural channels: (i) a *productivity* effect that lowers marginal cost, (ii) an *AI input* or bottleneck effect that raises marginal cost via an additional factor of production, (iii) *endogenous market power* in product markets that increases desired markups, (iv) *monopsony power* in labor markets that raises effective labor costs by increasing firms' marginal cost of hiring, and (v) a *demand* effect operating through AI-induced investment and spending. The framework extends the canonical New Keynesian system by allowing AI intensity z_t to enter marginal cost and desired markups in a disciplined way.

In this section, AI intensity z_t is introduced as an exogenous process. This reduced-form treatment is useful as a starting point because it isolates the main inflationary and disinflationary mechanisms associated with AI in a transparent way, keeps the mapping from AI to marginal cost and inflation analytically tractable, and helps fix ideas before moving to a richer environment. It also makes the model directly estimable: the exogenous AI state can be taken to the data, allowing us to recover the composite supply-side wedge, the demand-channel term, and the implied impact elasticities within a simple New Keynesian system. In the following section, we relax this simplifying assumption and allow AI to enter endogenously through investment.

6.1 Households and the IS Curve

A representative household maximizes expected lifetime utility and faces standard intertemporal budget constraints, which yield the usual Euler equation. Log-linearizing around a zero-inflation steady state, the output gap x_t satisfies:

$$x_t = \mathbb{E}_t x_{t+1} - \frac{1}{\sigma}\left(i_t - \mathbb{E}_t \pi_{t+1} - r_t^n\right) + \lambda_d d_t, \tag{22}$$

where i_t is the nominal policy rate, π_t denotes inflation, r_t^n is the natural real rate, and $\sigma > 0$ is the inverse of the intertemporal elasticity of substitution. The term d_t captures AI-driven demand shifts—for instance, investment in computing capacity, data centers, or complementary capital—with $\lambda_d > 0$ measuring their impact on the output gap. We assume

$$d_t = \phi_D z_t, \tag{23}$$

so that an AI shock z_t has a direct expansionary effect on demand.

6.2 Technology, AI Input, and Productivity

On the production side, a continuum of intermediate firms $i \in [0, 1]$ combines labor $L_t(i)$ and an AI-intensive input $M_t(i)$ via a Cobb–Douglas technology,

$$Y_t(i) = A_t\, L_t(i)^{1-\alpha} M_t(i)^{\alpha}, \qquad \alpha \in (0, 1), \tag{24}$$

where A_t denotes total factor productivity. AI affects technology in two ways. First, it raises productivity:

$$a_t \equiv \log A_t = \phi_A z_t, \qquad \phi_A > 0. \tag{25}$$

Second, higher AI intensity requires more specialized AI input M_t (e.g. computing, data, specialized energy), which may be subject to capacity constraints and rising relative cost. Let $p_{M,t}$

be the real price of M_t; then log-linearization around the steady state yields:

$$\hat{p}_{M,t} = \phi_{pM} z_t, \qquad \phi_{pM} > 0, \tag{26}$$

reflecting the idea that surges in AI adoption raise the relative price of AI-related inputs. In our framework, we capture the possibility that scaling AI capability draws on scarce complementary inputs (compute, energy, data infrastructure) by allowing the relative price of the AI-intensive input to rise with AI intensity. Cost minimization implies that real marginal cost depends on the output gap, the real price of the AI input, and productivity. In particular, the AI shock contributes to marginal cost through two offsetting components: a *productivity* channel $(-a_t)$ and an *AI-input bottleneck* channel that scales with the cost share of M, $\alpha \in (0,1)$, via $\alpha \hat{p}_{M,t}$. Using equations (25–26), these contributions can be summarized by the composite term $\big(\alpha\phi_{pM} - \phi_A\big)z_t$, thus extending the familiar marginal-cost decomposition to include AI as both productivity-enhancing and input-cost increasing through scarce complements.

6.3 AI and Monopsony Power in the Labor Market

We next allow AI to affect labor costs through monopsony power. See Appendix E for more details. Suppose each firm faces an upward-sloping labor supply curve,

$$L_t(i) = \left(\frac{W_t(i)}{W_t}\right)^{\varphi_t} L_t, \qquad \varphi_t > 0, \tag{27}$$

where φ_t is the wage elasticity of labor supply to the firm. Marginal expenditure on labor is then:

$$ME_{L,t}(i) = W_t(i)\left(1 + \frac{1}{\varphi_t}\right), \tag{28}$$

so that the effective wage entering marginal cost is

$$W_t^{\text{eff}}(i) = W_t(i)\left(1 + \frac{1}{\varphi_t}\right). \tag{29}$$

We assume that AI raises monopsony power by improving worker screening, monitoring, and algorithmic scheduling, which reduces the effective elasticity of labor supply to the firm. Let

$$\varphi_t = \bar{\varphi} - \eta_M z_t, \qquad \eta_M > 0, \tag{30}$$

so that higher AI intensity reduces φ_t and raises $\big(1 + 1/\varphi_t\big)$. Log-linearizing the effective wage around the steady state gives:

$$\widehat{W_t^{\text{eff}}} \approx \widehat{W}_t + \phi_M z_t, \qquad \phi_M \equiv \frac{\eta_M}{\bar{\varphi}(\bar{\varphi}+1)} > 0. \tag{31}$$

This wedge $\phi_M z_t$ provides an additional AI-driven force raising labor costs and, hence, marginal cost.

Empirical work documents pervasive monopsony power in labor markets (e.g. (Dube et al., 2020)) and shows that digital and algorithmic management tools can further reduce workers' effective outside options and mobility. Algorithmic scheduling, monitoring, and performance tracking increase firms' ability to exploit local labor-market frictions and customize wage offers. Emerging evidence suggests that automation and AI-intensive firms exhibit stronger monopsony power and a larger wedge between marginal product and wage, consistent with $\phi_M > 0$. This provides a structural rationale for letting AI raise marginal cost through labor-market power.

6.4 AI and Endogenous Markups in Product Markets

As surveyed above, several empirical studies find that AI adoption is associated with higher markups and increased industry concentration. These results support modeling AI as lowering effective product substitutability and raising desired markups. To this end, we model the impact of AI on markups. On the product-market side, we retain the Dixit–Stiglitz demand structure:

$$Y_t(i) = \left(\frac{P_t(i)}{P_t}\right)^{-\varepsilon_t} Y_t, \tag{32}$$

where ε_t is the elasticity of substitution across varieties. A monopolistically competitive firm chooses $P_t(i)$ to maximize static profits,

$$\max_{P_t(i)} \; P_t(i)Y_t(i) - MC_t\, Y_t(i), \tag{33}$$

leading to the optimal price:

$$P_t(i) = \mu_t MC_t, \qquad \mu_t = \frac{\varepsilon_t}{\varepsilon_t - 1}. \tag{34}$$

AI may increase firms' market power by enabling better demand prediction, personalized pricing, and network-based lock-in, which effectively reduce ε_t. We model this as:

$$\varepsilon_t = \bar{\varepsilon} - \gamma z_t, \qquad \gamma > 0. \tag{35}$$

Log-linearizing around a steady state with $\varepsilon_t = \bar{\varepsilon}$ yields

$$\hat{\mu}_t \approx \frac{\gamma}{\bar{\varepsilon}(\bar{\varepsilon} - 1)} z_t \equiv \phi_\mu z_t, \qquad \phi_\mu > 0. \tag{36}$$

with $\hat{\mu}_t$ being the deviation of desired markups from steady state. The time-varying price markup is, thus, increasing in the intensity of AI.

Real marginal cost. From firms' cost minimization for the Cobb–Douglas technology, real marginal cost is a weighted average of factor prices net of productivity. Aggregating across firms and log-linearizing around the steady state yields

$$mc_t = \lambda x_t + \alpha \hat{p}_{M,t} - a_t + \phi_M z_t, \tag{37}$$

where $\lambda > 0$ maps the output gap into marginal cost, $\alpha \in (0, 1)$ captures the cost share of the AI input, and the term $\phi_M z_t$ reflects the AI-induced increase in the effective wage due to monopsony power. Using equations (25) and (26) to substitute for a_t and $\hat{p}_{M,t}$, we obtain the reduced-form representation

$$mc_t = \lambda x_t + \big(-\phi_A + \alpha\phi_{pM} + \phi_M\big) z_t, \tag{38}$$

which makes explicit the three AI-related components: productivity ($-\phi_A$), AI-input bottlenecks ($\alpha\phi_{pM}$), and labor-market power (ϕ_M).

6.5 Price Setting and the New Keynesian Phillips Curve

Firms adjust prices à la Calvo: each period, a fraction $(1-\theta)$ of firms can reset their price; the remainder index by steady-state inflation. Solving the Calvo problem, the optimal reset price depends on current and expected future marginal costs and desired markups. Log-linearization around the steady state yields the NKPC:

$$\pi_t = \beta \mathbb{E}_t \pi_{t+1} + \kappa mc_t + \kappa \hat{\mu}_t, \tag{39}$$

where $\beta \in (0,1)$ is the discount factor and $\kappa = (1-\theta)(1-\beta\theta)/\theta$ is the slope of the Phillips curve. Substituting Equation (38) and $\hat{\mu}_t = \phi_\mu z_t$:

$$\pi_t = \beta \mathbb{E}_t \pi_{t+1} + \kappa \lambda x_t + \kappa\big(-\phi_A + \alpha\phi_{pM} + \phi_M + \phi_\mu\big) z_t. \tag{40}$$

Equation (40) shows that AI affects inflation through (i) marginal-cost effects summarized by $-\phi_A + \alpha\phi_{pM} + \phi_M$ and (ii) desired markup variation ϕ_μ.

According to Korinek (2024), monetary policy may face a targeting problem if labor-market quantities become less central for marginal cost and pricing, so that Phillips-curve indicators based on unemployment and conventional slack lose predictive power and need to be supplemented by new notions of capacity utilization (e.g. compute, robots) tied to reproducible capital and other bottleneck inputs. Our framework aligns with this logic in three ways. First, it explicitly shifts inflation determination away from a purely labor-based marginal-cost object by allowing AI to move marginal cost directly through an AI-specific input bottleneck and wedge terms, so that inflation can rise even when labor-market slack appears stable. Second, the reduced-form relationship in which the relative price of the AI-intensive input rises with AI intensity provides a tractable proxy for "compute-and-energy tightness," capturing the idea that capacity constraints in scalable capital and complementary infrastructure can act like a new source of slack relevant for inflation dynamics. Third, by embedding these non-labor cost and markup components directly in the New Keynesian Phillips curve, the model clarifies why conventional activity measures may be insufficient in AI-driven episodes: inflation can be driven by bottleneck pressures and shifts in pricing power even absent large movements in standard labor-market indicators, precisely the kind of environment in which policymakers would be pushed toward broader capacity-utilization concepts.

6.6 Monetary Policy

Monetary policy follows a Taylor-type rule:

$$i_t = r_t^n + \phi_\pi \pi_t + \phi_x x_t, \tag{41}$$

with $\phi_\pi > 1$ ensuring the Taylor principle and $\phi_x \geq 0$ allowing for output stabilization.

6.7 Closed-form Solutions

Using $d_t = \phi_D z_t$ and Equation (40), the equilibrium dynamics of (x_t, π_t) in response to AI shocks are governed by the coupled IS–NKPC system:

$$x_t = \mathbb{E}_t x_{t+1} - \frac{1}{\sigma}\left(i_t - \mathbb{E}_t \pi_{t+1} - r_t^n\right) + \lambda_d \phi_D z_t, \tag{42}$$

$$\pi_t = \beta \mathbb{E}_t \pi_{t+1} + \kappa \lambda x_t + \kappa\big(-\phi_A + \alpha\phi_{pM} + \phi_M + \phi_\mu\big) z_t. \tag{43}$$

Substituting the Taylor rule gives a two-equation system linking (x_t, π_t) to z_t, from which one can derive the impact and dynamic responses of output and inflation to AI shocks.

Artificial Intelligence follows an AR(1) process,

$$z_t = \rho_z z_{t-1} + \varepsilon_t^z, \qquad |\rho_z| < 1, \tag{44}$$

so that $\mathbb{E}_t z_{t+1} = \rho_z z_t$.

Conjecture a linear equilibrium of the form

$$x_t = A^x z_t, \qquad \pi_t = A^\pi z_t, \tag{45}$$

where A^x and A^π are the elasticities of the output gap and inflation with respect to the AI shock. Then, as shown in Appendix F, we obtain:

$$x_t = \underbrace{\frac{(1-\beta\rho_z)\,\lambda_d\phi_D}{\Omega}\,z_t}_{\text{Demand}} + \frac{\kappa(\phi_\pi - \rho_z)}{\sigma\,\Omega}\left[\underbrace{\phi_A}_{\text{Productivity}} - \underbrace{\alpha\phi_{pM}}_{\text{AI-input cost}} - \underbrace{\phi_\mu}_{\text{Markup}} - \underbrace{\phi_M}_{\text{Monopsony}}\right] z_t \tag{46}$$

$$\pi_t = \underbrace{\frac{\kappa\lambda\lambda_d\phi_D}{\Omega} z_t}_{\text{Demand}} + \frac{\kappa\,(1-\rho_z+\phi_x/\sigma)}{\Omega}\left[\underbrace{-\phi_A}_{\text{Productivity}} + \underbrace{\alpha\phi_{pM}}_{\text{AI-input cost}} + \underbrace{\phi_\mu}_{\text{Markup}} + \underbrace{\phi_M}_{\text{Monopsony}}\right] z_t. \tag{47}$$

with $\Omega \equiv \left(1-\rho_z+\frac{\phi_x}{\sigma}\right)(1-\beta\rho_z)+\frac{\kappa\lambda}{\sigma}(\phi_\pi-\rho_z)$. Equations (46)–(47) show explicitly how the five AI channels—productivity, AI input cost, endogenous price markups, monopsony power, and demand—shape the responses of the output gap and inflation. Specifically, the demand effect raises the output gap and inflation, whereas the four supply-side effects operate on inflation and the output gap in opposite directions. The productivity effect is the *only* mechanism in the model that unambiguously works to lower inflation. A crucial question regarding the impact on both variables is, then, the overall impact of the supply-side effects, which, for simplicity, can be regrouped as $\Delta \equiv -\phi_A + \alpha\phi_{pM} + \phi_\mu + \phi_M$. If Δ is non-negative then inflation will unambiguously rise. If Δ is negative, the supply-side block is disinflationary, and the overall impact on inflation depends on whether the demand term dominates in Equation (47).

A further implication of the closed-form expression for the output gap is that x_t can be used as an empirical diagnostic for the relative strength of the competing AI channels. In the analytical model, both the productivity enhancement and the demand channel push the output gap upward. In the absence of other distortions, an increase in AI intensity should therefore expand economic activity. Formally, this occurs because the productivity term ($\phi_A > 0$) and the demand term ($\lambda_d\phi_D > 0$) appear with positive coefficients in the closed-form solutions. Thus, a necessary condition for (x_t) to fail to rise is that the remaining components of the supply-side block—the markup channel, the monopsony channel, and the AI-input cost channel—are sufficiently strong to offset these expansionary forces.

Consequently, observing a muted or negative response of the output gap to rising AI adoption is informative about the structural composition of AI shocks. If productivity and demand are both expansionary but measured activity does not increase, the implication is that the cost-raising channels must dominate. For example, a strong endogenous markup response would induce firms to restrict output in order to sustain higher markups, counteracting the productivity gains. Similarly, monopsonistic wage setting may raise marginal costs through wage markdown dynamics, reducing labor input and suppressing output. Costly AI inputs (M_t) may also absorb resources and tighten production constraints, further limiting expansion. Thus, the sign and magnitude of the output response serve as a useful statistic for whether the inflationary or disinflationary components of AI adoption prevail in equilibrium.

Furthermore, while attention has been given to the demand/investment effect of AI, the analysis demonstrates that the demand channel is *neither necessary nor sufficient* for AI to generate higher inflation. Even with $\lambda_d \phi_D = 0$, inflation can rise if the supply-side block Δ is positive (for example, when AI-driven markups, monopsony, and AI-input costs dominate the productivity gains). Conversely, a positive demand effect need not lead to higher inflation if the productivity channel is sufficiently strong.

6.8 Numerical Illustration

This subsection provides a benchmark numerical evaluation to illustrate the quantitative implications of the analytical framework. The goal is *not* to deliver a point forecast of U.S. inflation "caused by AI," nor to claim that the parameters governing AI adoption and market structure are known with certainty. Rather, the numerical values serve two purposes. First, they discipline the discussion by ensuring that the analytical decompositions in equations (46)–(47) generate magnitudes that are economically sensible. Second, they illustrate how the sign and strength of inflation and output responses depend on the relative size of the disinflationary productivity component versus the three cost-raising components (AI-input bottlenecks, endogenous markups, and monopsony power), together with the demand channel.

6.8.1 Parameterization

The discount factor is set to $\beta = 0.99$, consistent with a quarterly frequency and a steady-state annual real interest rate close to 4 percent. The intertemporal preference parameter is set to $\sigma = 1.5$, which lies near the middle of the range commonly used in quantitative New Keynesian applications and implies an intertemporal elasticity of substitution of $1/\sigma \approx 0.67$. The slope of the Phillips curve with respect to real marginal cost is set to $\kappa = 0.085$, which corresponds to an average price duration of roughly three quarters and implies meaningful but not excessive inflation sensitivity to cost pressures. The mapping from the output gap to real marginal cost is set to $\lambda = 0.50$. In the present analytical framework, λ is interpreted as an *effective* reduced-form elasticity of marginal cost with respect to economic activity, capturing real rigidities and other non-modeled features that can attenuate the sensitivity of marginal cost to the output gap. Together, these choices imply an activity-based Phillips-curve slope of $\kappa\lambda = 0.0425$, yielding a moderately flat relationship between inflation and the output gap that keeps the model's inflation dynamics responsive to the AI-related cost and market-power wedges emphasized in the paper, rather than being driven mechanically by an overly steep baseline activity–inflation link. Monetary policy is governed by a Taylor rule with $(\phi_\pi, \phi_x) = (1.5, 0.125)$.

Calibration of the AI-input cost share α. A key parameter governing the strength of the AI-input bottleneck channel is the steady-state cost share of the AI-intensive input, α, which scales the effect of movements in the relative price of the AI input $p_{M,t}$ on real marginal cost through the term $\alpha \hat{p}_{M,t}$. We set $\alpha = 0.10$ in the baseline calibration. This value is disciplined using two empirical exercises: an annual measure based on BEA–BLS industry production-account data over 2007–2023 and an external benchmark-year measure based on BEA Use tables.

Both exercises map the AI-complementary input bundle to information-technology and software capital, data- and computing-related services, and energy inputs. The resulting empirical shares are centered around moderate but non-negligible values, with the KLEMS-based estimates and the benchmark Use-table estimates delivering overlapping ranges. We therefore choose $\alpha = 0.10$ as a conservative midpoint calibration. Appendix D provides the detailed data construction, numerator and denominator definitions, benchmark-year Use-table mapping, and robustness checks underlying this calibration.

Recovering Δ and $\lambda_d \phi_D$ from the data This subsection explains how we recover the magnitude of the direct (cost-push) AI wedge Δ and the demand-channel term $\lambda_d \phi_D$ from the data within our estimated New Keynesian system. The object Δ governs the component of inflation that responds *directly* to the AI state, over and above the indirect effect operating through the output gap. Our baseline estimate of Δ comes from estimating the New Keynesian Phillips curve (NKPC) by instrumental variables and then backing out Δ from the estimated AI coefficient given the calibrated Phillips-curve slope κ. Throughout, we use the *same* dummy variables (in particular the post-2020 "surge" dummy) and the *same* macro controls as in the section that quantifies the contributions of AI to inflation; these variables enter the NKPC regression and the auxiliary channel regressions on the same aligned quarterly sample intersection.

Step 1: Constructing the AI state variable z_t (HP-filtered level index). We begin from a quarterly *level* index of AI activity. In the empirical implementation, we remove low-frequency movements using the Hodrick–Prescott (HP) filter applied to this AI level series (with the standard quarterly smoothing parameter). We denote the resulting HP *cycle* by z_t, and treat z_t as the AI state used in the estimated NK system. We then estimate an autoregressive representation,

$$z_t = \rho_z z_{t-1} + \nu_t, \tag{48}$$

and use the bounded estimate $\widehat{\rho}_z$ as the persistence parameter that enters the equilibrium mapping of the linearized system. Finally, z_t is standardized to mean zero and unit variance within the final system estimation sample so that coefficients can be interpreted as responses to a one-standard-deviation movement in the AI state.[3]

Step 2: Expectations used in the forward-looking system. The NKPC and the New Keynesian investment–savings relationship use one-quarter-ahead expectations of inflation and the output gap. Expected inflation is constructed from the Survey of Professional Forecasters (SPF), using the SPF "Median Level" forecast of the *price level*. Let P_t^e denote the SPF forecast (made at time t) of the price level for quarter t. We form expected one-quarter-ahead inflation as

$$\pi_t^e \equiv 100\big(\log P_{t+1}^e - \log P_t^e\big). \tag{49}$$

Expected output-gap x_t^e is constructed by combining (i) the SPF forecast of real output growth with (ii) a one-step projection of the HP-filter trend of log real output used to define the realized output gap. Let y_t denote log real output and let y_t^{HP} denote the HP trend of y_t. Using the SPF real-growth forecast to form a forecasted log level y_{t+1}^e, and projecting the HP trend forward by one step via

$$y_{t+1}^{\mathrm{HP,proj}} \equiv y_t^{\mathrm{HP}} + \big(y_t^{\mathrm{HP}} - y_{t-1}^{\mathrm{HP}}\big), \tag{50}$$

[3] In the implementation, the standardization is performed on the exact system sample after intersecting all series, dummies, and controls, ensuring comparability across specifications.

we set

$$x_t^e \;\equiv\; 100\big(y_{t+1}^e - y_{t+1}^{\mathrm{HP,proj}}\big), \qquad x_t \;\equiv\; 100\big(y_t - y_t^{\mathrm{HP}}\big). \tag{51}$$

Step 3 (baseline): estimating Δ from the NKPC. Define the forward-looking inflation object

$$\pi_t - \beta\,\pi_t^e, \tag{52}$$

where β is the discount factor. The NKPC augmented with a direct AI cost-push term is written as

$$\pi_t - \beta\,\pi_t^e \;=\; \kappa\lambda\, x_t \;+\; \kappa\Delta\, z_t \;+\; u_t, \tag{53}$$

where κ is calibrated, λ maps the output gap into real marginal cost, and u_t is a residual. The parameter of interest is Δ, which scales the *direct* response of inflation to the AI state z_t holding x_t fixed.

Because x_t is potentially endogenous in (53), we estimate (53) by two-stage least squares (2SLS). The regressor vector includes x_t and z_t (and an intercept when included), along with the same controls and the same "surge" dummy used above. The instrument set contains z_t (which instruments itself), lags of inflation, the output gap, the policy rate, and the AI state up to a fixed lag length; when controls/dummies are active, they are also included among the instruments so that identification is conditional on the same observed confounders. Standard errors are computed using heteroskedasticity-and-autocorrelation-consistent (HAC) Newey–West corrections.

Let $\widehat{s}$ denote the estimated coefficient on x_t and let $\widehat{k\Delta}$ denote the estimated coefficient on z_t in (53). Given the calibrated κ, we recover

$$\widehat{\Delta} \;=\; \frac{\widehat{k\Delta}}{\kappa}, \qquad \widehat{\lambda} \;=\; \frac{\widehat{s}}{\kappa}. \tag{54}$$

Equation (54) is our primary estimate of Δ.

The structural estimation implies $\Delta = 0.374$ in the specification that excludes data centers over the sample 2007Q1–2025Q3, and $\Delta = 1.072$ in the specification that includes data centers over the sample 2014Q1–2025Q3. These estimates reinforce the conclusions from the preceding empirical sections. At the aggregate level, AI exerts a positive effect on inflation, and the magnitude of that effect is stronger in the more recent period, particularly in specifications that incorporate the infrastructure-intensive component of AI diffusion.

Recovering the demand wedge from the New Keynesian investment–savings relation (NKIS). In the model, the AI-driven demand channel enters the New Keynesian investment–savings relation (NKIS) through the term $\lambda_d\phi_D$. Because these two parameters always appear multiplied, they are not separately identified from the NKIS: only their product is pinned down by the data. We thus normalize $\lambda_d \equiv 1$ and interpret the estimated NKIS coefficient on the AI state as the implied value of ϕ_D.

Operationally, we construct the NKIS left-hand-side using one-quarter-ahead expectations (from the Survey of Professional Forecasters) and the calibrated intertemporal elasticity parameter σ. We then regress this constructed NKIS residual on the AI state z_t, including the same "surge" dummy and the same set of macro controls used in the contribution accounting section, on the same quarterly sample intersection. Under the normalization $\lambda_d = 1$, the coefficient on z_t delivers $\widehat{\phi}_D$, and equivalently it is the empirical estimate of the demand wedge $\widehat{\lambda_d\phi_D}$. The estimates yield $\phi_D = 0.469$ for the specification with no data centers (2007Q1-2025Q3), and $\phi_D = 0.931$ for the specification with data centers (2014Q1-2025Q3).

6.9 Impact elasticities of inflation to AI innovations

Table 7 reports the estimated impact elasticities of inflation, A_π, to an identified AI innovation for nine alternative AI indices. For each index, the table reports the immediate response of inflation, in percentage points, to a one-standard-deviation AI shock within the estimated New Keynesian system. The reported confidence intervals are obtained by a delta method that propagates estimation uncertainty from the structural coefficients into the implied inflation elasticity. We present three related objects: the full-system response when both the IS and Phillips-curve blocks are active, the component transmitted through the IS block alone, and the component transmitted through the Phillips curve alone.

Demand and Supply transmission (NKIS & NKPC)

Variant	Mean	CI	N
Investment: Excluding Data Centers	0.339	(0.325, 0.353)	71
Investment: Including Data Centers	0.433	(0.418, 0.447)	45
Capital Services	0.339	(0.326, 0.352)	71
Patents	0.019	(-0.001, 0.038)	64
Data Centers	0.529	(0.512, 0.546)	45
Capital Services & Patents	0.278	(0.262, 0.294)	71
Composite: Excluding Data Centers & Capital Services	0.273	(0.257, 0.290)	71
Composite: Excluding Data Centers	0.303	(0.287, 0.318)	71
Composite: Including Data Centers	0.557	(0.539, 0.575)	45

Demand-side transmission only (NKIS)

Variant	Mean	CI	N
Investment: Excluding Data Centers	0.223	(0.216, 0.229)	71
Investment: Including Data Centers	0.256	(0.251, 0.261)	45
Capital Services	0.219	(0.212, 0.226)	71
Patents	0.025	(0.015, 0.035)	64
Data Centers	0.334	(0.320, 0.348)	45
Capital Services & Patents	0.165	(0.154, 0.176)	71
Composite: Excluding Data Centers & Capital Services	0.163	(0.152, 0.174)	71
Composite: Excluding Data Centers	0.187	(0.178, 0.197)	71
Composite: Including Data Centers	0.377	(0.367, 0.387)	45

Supply-side transmission only (NKPC)

Variant	Mean	CI	N
Investment: Excluding Data Centers	0.116	(0.104, 0.128)	71
Investment: Including Data Centers	0.177	(0.163, 0.191)	45
Capital Services	0.120	(0.109, 0.131)	71
Patents	-0.006	(-0.023, 0.010)	64
Data Centers	0.195	(0.185, 0.205)	45
Capital Services & Patents	0.113	(0.101, 0.125)	71
Composite: Excluding Data Centers & Capital Services	0.110	(0.098, 0.123)	71
Composite: Excluding Data Centers	0.115	(0.103, 0.127)	71
Composite: Including Data Centers	0.180	(0.165, 0.195)	45

Table 7: Impact elasticities of inflation, A_π, to an AI innovation across nine alternative AI measures. The first panel reports the total inflation response when both the IS and Phillips-curve blocks are active. The second panel isolates the component operating through the demand side of the model, while the third panel isolates the component operating through the direct inflation block. Mean reports the point estimate of the impact elasticity, CI reports the 90% confidence interval obtained by a delta method, and N denotes the effective estimation sample used for the corresponding specification.

The first result that emerges from Table 7 is simple and important: the estimated response of inflation to an AI innovation is positive for almost every measure considered. This holds most clearly in the full-system estimates, where all nine point estimates are positive, and it remains true in the decomposition across channels, with the only notable exception being the patent-based measure in the supply-side block, where the estimate is slightly negative and close to zero. Looking at the full-system estimates, the strongest responses are obtained for measures that incorporate data centers, especially the broad composite that includes data centers, the standalone data-center measure, and investment including data centers. This pattern suggests that AI measures tied more directly to realized deployment, infrastructure expansion, and compute capacity are the ones most strongly associated with inflationary pressure. Other measures, including investment excluding data centers, capital services, and the broader composites that exclude data centers, also imply clearly positive responses, though of somewhat smaller magnitude. By contrast, the patents measure remains the weakest case, with an estimated effect that is close to zero.

The decomposition into channels reinforces this message. The demand-side block shows positive responses for all measures, and in most cases these responses are larger than the corresponding direct inflation responses in the Phillips-curve block. This indicates that a substantial part of the inflationary effect of AI operates through expansionary demand forces. At the same time, the supply-side block also yields positive estimates for nearly all measures, especially for those related to data centers and AI investment, implying that direct pricing or marginal-cost effects also contribute to the overall inflation response. The patent-based measure again stands apart, showing only a very weak demand-side effect and essentially no positive direct inflation effect.

Overall, the table connects closely to the contribution exercises presented earlier in the paper. Those exercises trace how the realized sequence of AI movements maps into observed inflation over time, whereas Table 7 reports the underlying impact elasticities that govern that mapping within the estimated New Keynesian system. In that sense, the table provides the structural counterpart to the contribution results (Section 3). The positive contribution of AI to inflation documented elsewhere is not appearing mechanically or by construction; rather, it reflects the fact that, across most of the AI measures considered here, the estimated structural response of inflation to an AI innovation is itself positive. This is especially true for measures linked to realized investment, data-center expansion, and broader deployment-oriented composites, which consistently generate positive responses in the full specification and in the channel decompositions. The results therefore strengthen the interpretation that AI has operated, over the sample considered, as a force putting upward pressure on inflation through a combination of demand expansion and direct cost or pricing effects. At the same time, the smaller estimate for patents underscores that not all AI-related indicators capture the same economic margin and that not all forms of exposure to AI have the same macroeconomic implications, including the impact on the inflation rate.

6.10 Reconstructing Δ from Estimated Channels

The NKPC estimate summarizes the supply-side effect of AI on inflation through the composite term Δ. This coefficient is useful because it measures the net AI-related supply-side wedge in the Phillips curve, but it also compresses several mechanisms into a single number. To connect the aggregate estimate more directly to the channels in the model, we decompose Δ into four components:

$$\Delta = -\phi_A + \alpha\phi_{p_M} + \phi_M + \phi_\mu. \tag{55}$$

The term $-\phi_A$ captures the productivity channel and enters negatively because higher productivity lowers marginal cost. The term $\alpha\phi_{p_M}$ captures the AI-input-price channel, where ϕ_{p_M} is the response of AI-related input prices to AI intensity and α is the marginal-cost weight of these inputs. The term ϕ_M captures the labor-market power or wage-markdown channel, while ϕ_μ captures the product-market markup channel.

The objective of this exercise is to assess whether the model's supply-side channels are visible separately in the data. Each component is estimated by relating an empirical proxy for the corresponding channel to the same AI-intensity measure used in the NKPC analysis, conditional on the macroeconomic information set used elsewhere in the paper. Productivity is measured in two ways: using labor productivity and using a residual TFP measure. The product-market markup is also measured in two ways: using a unit-labor-cost markup and using a broader marginal-cost markup. The AI-input-price term is based on an AI-related input-price proxy, scaled by the input-cost share parameter α, and the labor-market power term is measured using a wage-markdown proxy based on labor productivity relative to real hourly compensation. Appendix G provides the details on the construction of each series.

Table 8 reports the resulting decomposition for the longer sample that excludes data centers from the AI measure. The productivity component is negative in all cases, confirming the presence of the disinflationary productivity mechanism. The AI-input-price, wage-markdown, and markup components are positive. The estimated values of Δ are therefore positive across all four measurement combinations, ranging from 0.348 to 0.459, with confidence intervals above zero. The implied values are also close to the directly estimated NKPC coefficient, $\Delta_z = 0.374$. This provides an important consistency check: the Phillips-curve estimate of the AI supply-side wedge is broadly aligned with the sum of independently measured channel components.

Approach	Statistic	$-\phi_A$	$\alpha\phi_{p_M}$	ϕ_M	ϕ_μ	Δ
Labor productivity; ULC markup	Estimate	-0.183	0.011	0.253	0.281	0.363
	SE	0.015	0.001	0.016	0.017	0.028
	90% CI	[-0.207, -0.158]	[0.010, 0.013]	[0.227, 0.280]	[0.254, 0.308]	[0.318, 0.409]
Labor productivity; broad MC markup	Estimate	-0.183	0.011	0.253	0.376	0.459
	SE	0.015	0.001	0.016	0.010	0.024
	90% CI	[-0.207, -0.158]	[0.010, 0.013]	[0.227, 0.280]	[0.359, 0.393]	[0.419, 0.499]
Residual TFP; ULC markup	Estimate	-0.197	0.011	0.253	0.281	0.348
	SE	0.011	0.001	0.016	0.017	0.026
	90% CI	[-0.216, -0.179]	[0.010, 0.013]	[0.227, 0.280]	[0.254, 0.308]	[0.306, 0.391]
Residual TFP; broad MC markup	Estimate	-0.197	0.011	0.253	0.376	0.444
	SE	0.011	0.001	0.016	0.010	0.022
	90% CI	[-0.216, -0.179]	[0.010, 0.013]	[0.227, 0.280]	[0.359, 0.393]	[0.407, 0.481]

Table 8: Decomposition of the AI supply-side inflation term across alternative productivity and markup measurements. The table reports the components of $\Delta = -\phi_A + \alpha\phi_{p_M} + \phi_M + \phi_\mu$, where $-\phi_A$ is the disinflationary productivity component, $\alpha\phi_{p_M}$ is the AI-input-price component, ϕ_M is the wage-markdown/monopsony component, and ϕ_μ is the product-market markup component. Standard errors are reported below point estimates, and 90% confidence intervals are reported below standard errors. For the broad marginal-cost markup, the table reports the changing-share KLEMS estimate; the corresponding fixed-share estimate is very similar, with $\phi_\mu = 0.379$, $\Delta = 0.461$ under labor productivity, and $\Delta = 0.447$ under residual TFP. The directly estimated NKPC value is $\Delta_z = 0.374$.

The main message from Table 8 is that the direct NKPC estimate of a positive AI supply-side wedge can be rationalized by the separate channel estimates. With labor productivity and the ULC markup, the implied value is $\Delta = 0.363$, almost identical to the direct estimate $\Delta_z = 0.374$. With residual TFP and the ULC markup, the implied value is $\Delta = 0.348$, again very close. When the broader marginal-cost markup is used, the implied values rise to 0.459 under labor productivity and 0.444 under residual TFP. These estimates are somewhat above the direct NKPC estimate, but remain in the same quantitative range.

This comparison shows that the aggregate supply-side effect of AI is not coming only from the Phillips-curve regression. The disinflationary productivity channel is present in the longer sample, and it is accompanied by positive input-cost, wage-markdown, and markup channels. The net effect is therefore positive in all specifications. In this sense, the data support the central mechanism of the model: AI can raise productivity while still generating inflationary supply-side pressure when its effects on input costs, labor-market wedges, and product-market markups are taken into account.

Table 9 repeats the decomposition using the AI measure that includes data centers. Because the data-center series begins later, the sample is shorter, covering 2014Q1–2025Q3. This exercise is useful because the data-center-inclusive AI measure captures the more recent infrastructure and compute-intensive phase of AI diffusion. The results show a stronger positive supply-side wedge. The implied values of Δ range from 0.802 to 0.954, depending on the markup measure. These values are below the directly estimated NKPC coefficient for the same shorter sample, $\Delta_z = 1.072$, but they remain close in magnitude. Thus, even in the data-center sample, the channel-by-channel decomposition delivers a supply-side AI effect that is in the same range as the direct Phillips-curve estimate.

Approach	Statistic	$-\phi_A$	$\alpha\phi_{p_M}$	ϕ_M	ϕ_μ	Δ
Labor productivity; ULC markup	Estimate	0.044	0.021	0.450	0.439	0.954
	SE	0.028	0.003	0.037	0.033	0.057
	90% CI	[-0.003, 0.090]	[0.017, 0.026]	[0.390, 0.511]	[0.385, 0.493]	[0.860, 1.048]
Labor productivity; broad MC markup	Estimate	0.044	0.021	0.450	0.287	0.802
	SE	0.028	0.003	0.037	0.030	0.055
	90% CI	[-0.003, 0.090]	[0.017, 0.026]	[0.390, 0.511]	[0.237, 0.336]	[0.711, 0.893]
Residual TFP; ULC markup	Estimate	0.044	0.021	0.450	0.439	0.954
	SE	0.028	0.003	0.037	0.033	0.057
	90% CI	[-0.003, 0.090]	[0.017, 0.026]	[0.390, 0.511]	[0.385, 0.493]	[0.860, 1.048]
Residual TFP; broad MC markup	Estimate	0.044	0.021	0.450	0.287	0.802
	SE	0.028	0.003	0.037	0.030	0.055
	90% CI	[-0.003, 0.090]	[0.017, 0.026]	[0.390, 0.511]	[0.237, 0.336]	[0.711, 0.893]

Table 9: Decomposition of the AI supply-side inflation term across alternative productivity and markup measurements, using the AI measure that includes data centers over the shorter sample 2014Q1–2025Q3. The table reports the components of $\Delta = -\phi_A + \alpha\phi_{p_M} + \phi_M + \phi_\mu$, where $-\phi_A$ is the productivity component, $\alpha\phi_{p_M}$ is the AI-input-price component, ϕ_M is the wage-markdown/monopsony component, and ϕ_μ is the product-market markup component. Standard errors are reported below point estimates, and 90% confidence intervals are reported below standard errors. For the broad marginal-cost markup, the table reports the changing-share KLEMS estimate; the corresponding fixed-share estimate is very similar, with $\phi_\mu = 0.279$, $\Delta = 0.794$ under labor productivity, and $\Delta = 0.794$ under residual TFP. The directly estimated NKPC value over this shorter data-center sample is $\Delta_z = 1.072$.

The decomposition with data centers differs from the longer-sample decomposition in two ways. First, the productivity component is no longer negative. The point estimate of $-\phi_A$ is positive, and its 90% confidence interval includes zero. Thus, over the shorter data-center sample, the productivity channel does not provide a statistically clear disinflationary force. This is important because the issue is not only that productivity gains may be too small to offset the inflationary supply-side channels; rather, in this sample, the measured productivity channel does not appear to operate as a clear offsetting force at all. This finding relates to the productivity-paradox mechanism emphasized by Brynjolfsson et al. (2017): even when digital technologies advance rapidly, their measured aggregate productivity effects may emerge only gradually because diffusion requires complementary investment, organizational adjustment, learning, and new forms of intangible capital. During this transition, the cost-raising components of AI adoption can be more visible in inflation than the eventual productivity gains.

Second, the supply-side channels that exert inflationary pressure become substantially larger. The AI-input-price component rises from about 0.011 in the longer sample to about 0.021 in the data-center sample. The wage-markdown component rises from about 0.253 to about 0.450, and the ULC-based markup component rises from about 0.281 to about 0.439. These larger inflationary components are consistent with the idea that the data-center-intensive phase of AI diffusion is associated with stronger infrastructure, input-cost, labor-market, and pricing pressures. The data-center results therefore sharpen the broader interpretation of the paper: the early macroeconomic footprint of AI may be dominated by deployment costs and market-structure effects, even if productivity gains materialize more fully over a longer horizon.

The channel decomposition also remains quantitatively close to the direct NKPC estimate. The direct NKPC estimate is larger in the shorter sample, and the independent channel estimates imply a much larger value of Δ than in the longer sample. With the ULC markup, the implied value is $\Delta = 0.954$, close to $\Delta_z = 1.072$. With the broad marginal-cost markup, the implied value is $\Delta = 0.802$, somewhat lower but still in the same range. The main conclusion is that once data centers are included, the positive AI supply-side wedge becomes stronger, and this stronger effect is visible not only in the aggregate NKPC coefficient but also in the separately measured input-cost, wage-markdown, and markup channels.

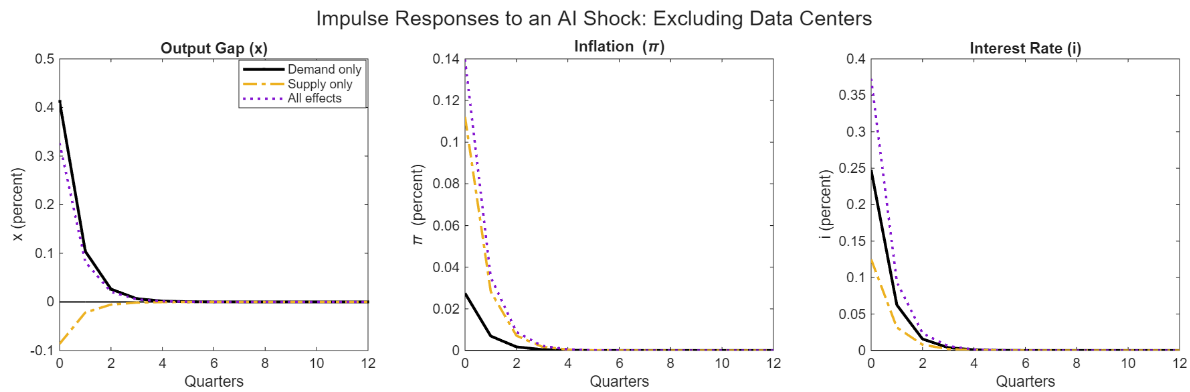
As a further robustness check, Appendix G also reports decompositions in which the productivity component is measured using quarterly production-function-based TFP proxies rather than labor productivity or the residual TFP measure from the marginal-cost accounting exercise. These proxies are constructed by subtracting a KLEMS-weighted bundle of quarterly input growth from quarterly output growth, using either nonfarm-business real value-added output or real GDP as the output measure. The resulting productivity estimates are then inserted into the same decomposition formula, while the AI-input-price, wage-markdown, and markup channels are estimated using the same conditioning information set as in the baseline exercise.

The results are strongly consistent with the main decomposition. In the longer sample that excludes data centers, the implied values of Δ remain positive across all production-function TFP specifications, ranging from 0.461 to 0.582. These values are at least as large as those obtained in Table 8, indicating that the positive AI supply-side wedge is not driven by the particular productivity measure used in the baseline decomposition. The productivity component remains disinflationary, but it is not large enough to offset the positive AI-input-price, wage-markdown, and markup components.

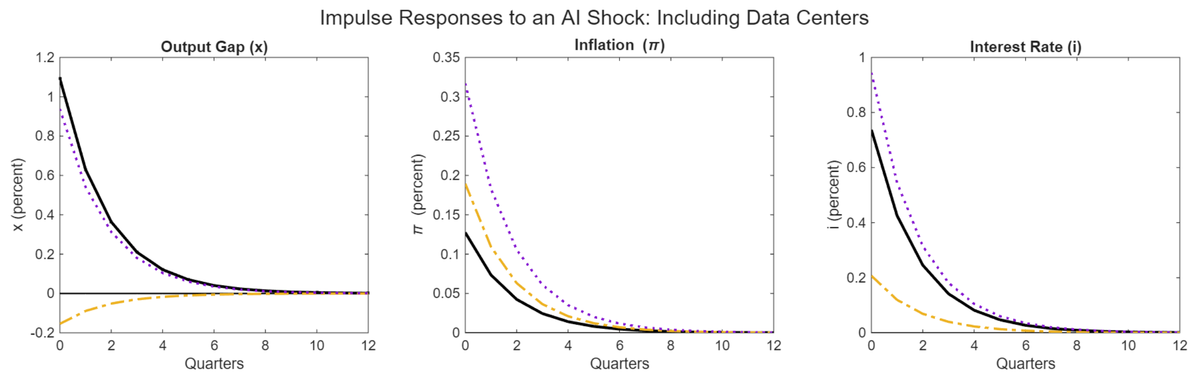
When the AI measure includes data centers, the implied values of Δ remain strongly positive, ranging from 0.650 to 0.821. This is larger than the corresponding production-function TFP estimates in the longer sample, reinforcing the conclusion that the data-center-intensive phase of AI diffusion is associated with stronger inflationary supply-side pressures. In this specification, the productivity channel continues to operate in the disinflationary direction, unlike in the baseline decomposition with data centers where the productivity component is not negative. However, the central result is unchanged: all three inflationary channels emphasized by the model, AI-input costs, wage-markdown pressures, and product-market markups, remain present and quantitatively important, and their combined effect dominates the productivity effect.

6.10.1 Impulse Responses in the One-Sector Model

We close this section by using impulse responses to summarize the model's transmission mechanism visually. Figure 5 reports responses to an AI-intensity shock in the one-sector model under three configurations: a case in which the shock operates only through the demand shifter in the IS and Phillips curves ("demand only"), a case in which it operates only through the supply-side block that alters marginal costs and markups ("supply only"), and the full model in which both demand and supply channels are active ("all effects"). Because the model is log-linearized, the paths for the output gap x_t, inflation π_t, and the policy rate i_t can be read directly as percent deviations from steady state. In all experiments, the responses decay smoothly, reflecting the AR(1) persistence of the AI process and the forward-looking structure of the IS curve and New Keynesian Phillips curve.



(a) Excluding data centers.



(b) Including data centers.

Figure 5: Impulse responses to an AI-intensity shock in the one-sector model. Each panel compares three experiments: (i) the demand channel only, (ii) all supply-side channels only (productivity, AI input costs, markups, and monopsony), and (iii) the full model with both demand and supply channels active. All series are reported as percent deviations from steady state. The top panel excludes the data-center channel from the AI supply block; the bottom panel includes data centers as an additional AI-intensive input, which amplifies the inflation response. Top panel: $\Delta = 0.374$, $\phi_D = 0.469$. Bottom panel: $\Delta = 1.072$, $\phi_D = 0.931$.

In the demand-only experiment, the AI shock enters solely as an expansionary disturbance in the IS and Phillips curves. The output gap x_t rises on impact and then decays monotonically back to zero, while inflation π_t increases modestly above its steady state. The mechanism is standard: higher AI intensity shifts desired spending forward, relaxing real rates and raising current demand; via the Phillips curve, the positive output gap translates into temporarily higher inflation. Monetary policy follows the Taylor rule and therefore raises the nominal rate, but the increase in i_t is limited because the inflation impulse itself is small.

The supply-only experiment activates the full supply block in Equations (46)–(47) while shutting down the direct demand shifter. In this configuration the AI shock affects the economy only by changing productivity, the cost of AI-intensive inputs, desired markups, and labor-market rents. The figure shows a clearly positive inflation response and a contraction in the output gap: π_t jumps up on impact and then declines gradually, while x_t becomes negative. This sign pattern is exactly what the decomposition with the composite supply term Δ implies when $\Delta > 0$: the AI shock raises marginal costs and effective markups enough to push inflation up, and the same supply-side wedge with opposite sign in the IS equation depresses real activity. The policy rate rises more strongly than in the demand-only case, as the central bank leans against the cost-push inflation generated by the supply block.

The full-model experiment (all effects) combines the inflationary supply block with the expansionary demand shifter. When data centers are *excluded* (top panel), the inflation path in the all-effects case lies above the demand-only path and not far below the supply-only path: the AI shock is clearly inflationary, but the peak increase in π_t is moderate. At the same time, the output gap response is much closer to zero than in either pure experiment. Positive demand pressure pushes x_t up, while higher markups and AI-input costs pull it down; the resulting output response is small, even though inflation is decisively positive. This "high π_t with muted x_t" configuration matches the logic of Equations (46)–(47): once $\Delta > 0$, the same AI shock that raises inflation need not generate a large expansion in activity because the markup and input-cost channels offset the demand-driven output boom.

Introducing data centers strengthens these patterns by making the AI shock more intensive in a scarce, specialized input. In the bottom panel, both the supply-only and all-effects experiments display a noticeably larger inflation spike and a somewhat steeper initial contraction in the output gap. The supply block now includes the data-center channel, which raises the relative price of AI infrastructure and tightens capacity constraints; this increases Δ and thereby amplifies the cost-push component of inflation. The all-effects inflation response becomes substantially higher than in the top panel while the output gap remains relatively small in absolute value because the stronger supply contraction is largely offset by the demand expansion. Correspondingly, the policy-rate response is also stronger when data centers are included, as the Taylor rule reacts to the larger and more persistent inflation impulse.

These impulse responses are qualitatively consistent with the empirical contribution estimates reported above. There, AI measures that include data centers generate larger average contributions to inflation than composites that exclude data centers, especially in the recent period. The model captures exactly this pattern: when AI is tied more tightly to data-center investment and AI-specific capital services, the same underlying AI-intensity shock produces a bigger and more persistent inflation response.

7 A nonlinear New Keynesian model with AI, robot substitution, monopsony, and endogenous markup dynamics

This section presents a nonlinear New Keynesian model designed to study how shocks to artificial intelligence (AI) propagate to real activity and, especially, to inflation. The model retains the key mechanisms needed for that question while remaining sufficiently disciplined to connect AI disturbances to observable macroeconomic outcomes. In particular, AI affects the economy through productivity, through the relative efficiency of robot services versus labor, through endogenous AI investment, through a bottleneck input whose scarcity can raise costs even when AI also improves efficiency, through firms' wage-setting power in labor markets, and through endogenous desired markups in product markets. These features make it possible for an AI expansion to generate either disinflationary or inflationary pressure depending on which channels dominate.

The economy is populated by households, monopolistically competitive firms, and a government. Households consume, work, save in one-period nominal bonds, and own the claims associated with robot capital and AI capability. Firms face Calvo price rigidity and therefore adjust prices only intermittently. The government consumes goods, makes transfers, raises lump-sum taxes, and issues one-period nominal debt. Monetary policy follows a standard Taylor rule. The resulting equilibrium is a nonlinear system of equations in levels, suitable for quantitative analysis of the inflationary consequences of AI shocks.

Two additional features are particularly important for the paper. First, the labor market is not perfectly competitive. Firms face an upward-sloping labor supply schedule, so the wage they pay affects the quantity of labor they can hire. This gives rise to a monopsony wedge between the wage and the marginal expenditure on labor, and AI can affect that wedge by changing effective labor market power. Second, desired markups are not imposed as an ad hoc function of AI. Instead, they are generated endogenously through a deep-habits demand structure. AI can then affect pricing power by altering customer retention and the elasticity of demand faced by firms. These two mechanisms are important because they allow AI to affect inflation not only through direct costs and demand pressure, but also through changes in labor-market power and product-market power.

A central feature of the model is that AI disturbances operate through the *investment margin*. In particular, the benchmark AI adoption shock is not a direct exogenous innovation to the stock of AI capability itself. Rather, it is a shock to the incentive or efficiency of current AI investment. Economically, this means that the economy becomes more AI-intensive only if agents devote current resources to adoption. This preserves the demand-and-investment channel emphasized in the paper and avoids the counterfactual implication that the AI stock can jump upward without requiring current expenditure.

7.1 Households

A representative household chooses consumption, aggregate labor supply, and bond holdings to maximize expected lifetime utility:

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\frac{(C_t - hC_{t-1})^{1-\sigma}}{1-\sigma} - \frac{\chi N_t^{1+\varphi}}{1+\varphi} \right], \qquad \beta \in (0,1),\ h \in [0,1). \tag{56}$$

The parameter h captures external habit formation in consumption. When $h = 0$, habits are shut down and utility becomes standard CRRA in current consumption. Keeping habits in the baseline is useful because they introduce inertia in aggregate demand and make the dynamic response of inflation more realistic.

The household faces the nominal budget constraint

$$P_t C_t + B_t = W_t N_t + \Pi_t^f + \Pi_t^k + R_{t-1} B_{t-1} + P_t T_t, \tag{57}$$

where B_t denotes one-period nominal bond holdings chosen at time t, R_t is the gross nominal interest rate between t and $t+1$, W_t is the aggregate nominal wage index, Π_t^f denotes profits rebated by intermediate firms, Π_t^k denotes profits or returns associated with ownership of robot capital and AI capability, T_t are lump-sum transfers.

The marginal utility of consumption is

$$U_{C,t} = (C_t - hC_{t-1})^{-\sigma} - \beta h(C_{t+1} - hC_t)^{-\sigma}. \tag{58}$$

Letting $\Lambda_t \equiv U_{C,t}/P_t$ denote the marginal utility of nominal income, the household's Euler equation for the one-period bond is

$$1 = \beta \mathbb{E}_t \left[\frac{\Lambda_{t+1}}{\Lambda_t} R_t \right]. \tag{59}$$

This condition governs intertemporal substitution in the usual way: a higher expected real return depresses current consumption and therefore shifts aggregate demand.

7.2 Labor aggregation and monopsony

To introduce monopsony in a parsimonious but micro-founded way, the household supplies a continuum of differentiated labor services $N_t(j)$, aggregated into total labor through

$$N_t = \left(\int_0^1 N_t(j)^{\frac{\varphi_t - 1}{\varphi_t}} \, dj \right)^{\frac{\varphi_t}{\varphi_t - 1}}, \qquad \varphi_t > 1. \tag{60}$$

Cost minimization on the part of firms implies demand for labor variety j:

$$N_t(j) = \left(\frac{W_t(j)}{W_t} \right)^{\varphi_t} N_t, \tag{61}$$

where $W_t(j)$ is the wage paid for type-j labor and W_t is the aggregate wage index:

$$W_t = \left(\int_0^1 W_t(j)^{1+\varphi_t} \, dj \right)^{\frac{1}{1+\varphi_t}}. \tag{62}$$

This structure implies that each firm faces an upward-sloping labor supply curve. Hiring more labor requires raising the wage on the inframarginal workforce as well. As a result, the relevant object for production decisions is not simply the wage, but the *marginal expenditure on labor*:

$$ME_t(j) = W_t(j) \left(1 + \frac{1}{\varphi_t} \right). \tag{63}$$

The ratio

$$\mu_t^N \equiv \frac{ME_t(j)}{W_t(j)} = 1 + \frac{1}{\varphi_t} \tag{64}$$

is the monopsony wedge. A lower labor-supply elasticity φ_t implies greater labor-market power and a larger gap between the wage and the true marginal labor cost.

The household's aggregate labor supply condition is

$$\chi N_t^{\varphi} = \frac{W_t}{P_t} U_{C,t}. \tag{65}$$

Firms take the wedge (64) into account when choosing labor input. This matters for inflation because the marginal cost relevant for price setting depends on ME_t, not just on the observed wage.

AI affects monopsony power through the labor-supply elasticity faced by firms. Let

$$\varphi_t = \bar{\varphi}_w \exp(-\psi_\varphi z_t), \qquad z_t \equiv \log\left(\frac{Z_t}{\bar{Z}}\right). \tag{66}$$

When $\psi_\varphi > 0$, a rise in AI capability lowers the effective labor-supply elasticity and raises firms' labor-market power. Economically, this can be interpreted as AI improving monitoring, scheduling, screening, or matching technologies in ways that reduce workers' outside options and make labor supply to each firm less elastic. The result is a higher marginal expenditure on labor and, potentially, more inflationary pressure even if measured wages do not rise one-for-one.

7.3 Production and AI technology

There is a representative final-good producer that aggregates a continuum of differentiated intermediate varieties into a final good. The final good is used for private consumption, government purchases, and investment in robot capital and AI capability. The final-good aggregator is standard Dixit–Stiglitz:

$$Y_t = \left(\int_0^1 Y_t(j)^{\frac{\varepsilon-1}{\varepsilon}} dj\right)^{\frac{\varepsilon}{\varepsilon-1}}, \qquad \varepsilon > 1. \tag{67}$$

In the benchmark constant-elasticity case,

$$Y_t(j) = \left(\frac{P_t(j)}{P_t}\right)^{-\varepsilon} Y_t, \tag{68}$$

with price index

$$P_t = \left(\int_0^1 P_t(j)^{1-\varepsilon} dj\right)^{\frac{1}{1-\varepsilon}}. \tag{69}$$

To generate endogenous desired markups, this demand system will later be modified using deep habits.

Each intermediate producer j operates the same production technology. Output is produced using a Cobb–Douglas combination of value added and a bottleneck input:

$$Y_t(j) = A_t\, V_t(j)^{1-\alpha} M_t(j)^{\alpha}, \qquad \alpha \in (0,1), \tag{70}$$

where A_t is total factor productivity, $V_t(j)$ is value added, and $M_t(j)$ is a bottleneck input that captures scarce AI-related infrastructure or intermediate services, such as computing capacity, data-center services, specialized chips, or energy-intensive AI-processing needs.

Value added is itself a CES aggregate of effective labor and robot services:

$$V_t(j) = \left[(1-\nu)\big(\Xi_t L_t(j)\big)^{\frac{\rho-1}{\rho}} + \nu\big(\Gamma_t R_t(j)\big)^{\frac{\rho-1}{\rho}}\right]^{\frac{\rho}{\rho-1}}, \qquad \rho > 0. \tag{71}$$

In this expression, $L_t(j)$ is labor employed by firm j, $R_t(j)$ is robot services, ν is the share parameter, and ρ is the elasticity of substitution between effective labor and effective robot services. The terms Ξ_t and Γ_t are labor-augmenting and robot-augmenting efficiency shifters, respectively. This specification is central to the model's economics. A rise in AI capability can increase overall efficiency, but it can also change the relative productivity of labor and robots, thereby altering the cost-minimizing input mix and the evolution of marginal cost.

AI capability enters production through three objects:

$$z_t \equiv \log\left(\frac{Z_t}{\bar{Z}}\right), \tag{72}$$

$$A_t = \bar{A}\exp(\phi_A z_t)\exp(\varepsilon_t^A), \tag{73}$$

$$\Gamma_t = \bar{\Gamma}\exp(\phi_\Gamma z_t), \qquad \Xi_t = \bar{\Xi}\exp(\phi_\Xi z_t). \tag{74}$$

The stock Z_t is the economy's AI capability. A rise in Z_t can increase total factor productivity directly through A_t, can make robots more effective through Γ_t, and may also improve labor efficiency through Ξ_t.

The bottleneck input has an endogenous real price. The benchmark specification links that price not only to current production use of the bottleneck input, but also to contemporaneous deployment of AI and robots:

$$p_{M,t} \equiv \frac{P_{M,t}}{P_t} = \bar{p}_M\left(\frac{M_t + \omega_Z I_t^Z + \omega_R I_t^R}{\bar{D}_M}\right)^{\upsilon}, \qquad \upsilon > 0. \tag{75}$$

Here $\omega_Z \geq 0$ and $\omega_R \geq 0$ measure the extent to which AI investment and robot investment directly load onto bottleneck demand, and $\bar{D}_M$ is steady-state bottleneck capacity. This specification captures the idea that bottleneck scarcity may arise not only because firms use more compute or intermediate infrastructure in current production, but also because rapid AI adoption and robot deployment themselves draw on scarce infrastructure, chips, energy, and installation capacity. When AI investment rises quickly, the relative price of the bottleneck input can therefore increase even before the productive gains from AI are fully realized.

Cost minimization implies a nominal marginal cost $MC_t(j)$ for firms. Since the exact expression is lengthy under the nested technology, it is convenient to summarize the real marginal cost as the unit-cost function associated with (70)–(71):

$$\frac{MC_t(j)}{P_t} = \mathcal{MC}\left(\frac{ME_t(j)}{P_t}, \frac{r_t^R}{P_t}, p_{M,t}, A_t, \Xi_t, \Gamma_t\right), \tag{76}$$

where r_t^R is the nominal rental rate of robot services and $ME_t(j)$ is marginal expenditure on labor. The important point is that the labor-market block enters marginal cost through the monopsony wedge. Inflation can therefore move because AI increases labor productivity, because it raises bottleneck scarcity, or because it changes labor-market power and thereby raises effective labor cost.

7.4 Robot capital and AI capability accumulation

Robot services are generated from a stock of robot capital, K_t^R, according to

$$R_t = u_t^R K_t^R, \tag{77}$$

where u_t^R is a utilization margin. To keep the model parsimonious, one may set $u_t^R = 1$ in the quantitative implementation.

Robot capital evolves according to

$$K_{t+1}^R = (1-\delta_R)K_t^R + I_t^R \left[1 - \frac{\phi_R}{2}\left(\frac{I_t^R}{I_{t-1}^R} - 1\right)^2\right]. \tag{78}$$

The adjustment-cost term slows down robot adoption and prevents unrealistically abrupt jumps in the capital stock after an AI shock.

AI capability evolves according to

$$Z_{t+1} = (1-\delta_Z)Z_t + I_t^Z \left[1 - \frac{\phi_Z}{2}\left(\frac{I_t^Z}{I_{t-1}^Z} - 1\right)^2\right]. \tag{79}$$

In this formulation, AI shocks do not enter the law of motion for the stock directly. Instead, they operate through the investment decision itself. This ensures that favorable AI disturbances increase the stock of AI capability only to the extent that firms choose to devote current resources to AI adoption.

Let q_t^R and q_t^Z denote the shadow values of robot capital and AI capability. The shadow value of robot capital satisfies

$$q_t^R = \beta \mathbb{E}_t \left[\frac{\Lambda_{t+1}}{\Lambda_t}\left(\frac{r_{t+1}^R}{P_{t+1}} + (1-\delta_R)q_{t+1}^R\right)\right]. \tag{80}$$

The shadow value of AI capability satisfies

$$q_t^Z = \beta \mathbb{E}_t \left[\frac{\Lambda_{t+1}}{\Lambda_t}\left(\frac{\Pi_{t+1}^Z}{P_{t+1}} + (1-\delta_Z)q_{t+1}^Z\right)\right]. \tag{81}$$

The payoff to AI capability should capture the marginal productive value of AI in equilibrium. A clean baseline specification is

$$\Pi_t^Z = \Big[\phi_A + (1-\alpha)\big((1-\nu)\phi_\Xi + \nu\phi_\Gamma\big)\Big] MC_t Y_t. \tag{82}$$

This specification ties the return to AI capability to the productive surplus created by the technology block. In words, AI earns a payoff because it raises total factor productivity, increases labor efficiency, and increases robot efficiency.

A broader alternative, if one wishes to attribute part of the markup and monopsony gains directly to AI capability, is

$$\Pi_t^Z = \Big[\phi_A + (1-\alpha)\big((1-\nu)\phi_\Xi + \nu\phi_\Gamma\big)\Big] MC_t Y_t + \omega_\varkappa\big(\mu_t^p - \bar{\mu}^p\big)Y_t + \omega_\varphi\big(\mu_t^N - \bar{\mu}^N\big)W_t N_t, \tag{83}$$

where $\omega_\varkappa$ and ω_φ determine how much of the product-market and labor-market surplus is imputed to AI capability. The benchmark quantitative exercises can use (82), while (83) is useful for robustness.

The key modification is that the AI adoption shock enters the *investment condition* rather than the stock accumulation equation. Let $\mathcal{A}_t^Z$ denote an AI investment wedge, with stochastic process

$$\log \mathcal{A}_t^Z = \rho_A^Z \log \mathcal{A}_{t-1}^Z + \varepsilon_t^{IZ}, \tag{84}$$

where ε_t^{IZ} is an innovation to the incentive or efficiency of AI investment. A convenient nonlinear investment condition is

$$1 = q_t^Z \left[1 - \frac{\phi_Z}{2}\left(\frac{I_t^Z}{I_{t-1}^Z} - 1\right)^2 - \phi_Z\left(\frac{I_t^Z}{I_{t-1}^Z} - 1\right)\left(\frac{I_t^Z}{I_{t-1}^Z}\right)\right] + \beta \mathbb{E}_t\left[\frac{\Lambda_{t+1}}{\Lambda_t} q_{t+1}^Z \phi_Z \left(\frac{I_{t+1}^Z}{I_t^Z} - 1\right)\left(\frac{I_{t+1}^Z}{I_t^Z}\right)^2\right] + \mathcal{X}_t^Z, \tag{85}$$

where $\mathcal{X}_t^Z$ is an investment wedge induced by $\mathcal{A}_t^Z$. A favorable realization of the shock lowers the effective marginal cost of current AI investment or raises its effective return, thereby increasing I_t^Z directly. The exact mapping between $\mathcal{A}_t^Z$ and $\mathcal{X}_t^Z$ is left flexible because alternative normalizations are possible. What matters economically is that the shock acts on the adoption decision itself. This keeps the demand-and-investment channel active and ensures that larger future AI capability must be financed by larger current AI expenditure.

The exact investment Euler equations with adjustment costs can be placed in an appendix. In the main text, the key intuition is sufficient: AI and robot accumulation are forward-looking and therefore sensitive to expected future profitability, inflation, and interest rates.

7.5 Endogenous markups through deep habits

To introduce endogenous markups without imposing an ad hoc reduced-form markup process, the model adopts a deep-habits demand structure. Consumers form variety-specific consumption habits, so current demand depends not only on current prices and total demand, but also on the stock of past purchases of each variety. This makes the elasticity of demand faced by a firm state dependent and therefore makes desired markups endogenous.

Let $S_t(j)$ denote the habit stock attached to variety j:

$$S_t(j) = (1 - \rho_S)C_t(j) + \rho_S S_{t-1}(j), \qquad \rho_S \in (0, 1). \tag{86}$$

Consumption of the final good is generated by the habit-adjusted aggregator

$$C_t = \left(\int_0^1 [C_t(j) - \varkappa_t S_{t-1}(j)]^{\frac{\varepsilon-1}{\varepsilon}} dj\right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{87}$$

where $\varkappa_t \in [0, 1)$ governs the strength of customer retention. When $\varkappa_t = 0$, the model collapses to the standard Dixit–Stiglitz case with constant elasticity and a constant desired markup.

The resulting firm-level demand curve is no longer characterized by a constant elasticity. Instead, firms that have accumulated a large habit stock face more captive demand and therefore a lower effective demand elasticity. Denote that effective elasticity by $\varepsilon_t^d(j)$, with associated desired markup

$$\mu_t^p(j) \equiv \frac{\varepsilon_t^d(j)}{\varepsilon_t^d(j) - 1}. \tag{88}$$

The important point is that $\mu_t^p(j)$ is endogenous: it varies with the state of demand, habit stocks, and the retention parameter.

AI affects this product-market channel through the retention technology:

$$\varkappa_t = \bar{\varkappa} \exp(\psi_\varkappa z_t), \tag{89}$$

with $\psi_\varkappa > 0$ implying that greater AI capability strengthens customer retention, personalization, targeted pricing, or product lock-in. Economically, this means that AI can increase firms' desired markups even without any exogenous markup shock.

7.6 Price setting and inflation

Intermediate firms are monopolistically competitive and face Calvo price rigidity. In each period, a firm can reset its price with probability $1-\theta$; otherwise, it keeps its price unchanged. For simplicity, the baseline omits mechanical price indexation.

Let P_t^* denote the common reset price chosen by firms that can reoptimize. The aggregate price index evolves according to

$$P_t^{1-\varepsilon} = (1-\theta)(P_t^*)^{1-\varepsilon} + \theta P_{t-1}^{1-\varepsilon}. \tag{90}$$

Aggregate inflation is defined by

$$\Pi_t \equiv \frac{P_t}{P_{t-1}}. \tag{91}$$

A firm that resets at time t chooses P_t^* to maximize the expected discounted value of profits while its price remains in effect:

$$\max_{P_t^*} \mathbb{E}_t \sum_{k=0}^{\infty} (\beta\theta)^k \frac{\Lambda_{t+k}}{\Lambda_t} \left[\left(\frac{P_t^*}{P_{t+k}} \right) Y_{t+k|t} - \left(\frac{MC_{t+k}}{P_{t+k}} \right) Y_{t+k|t} \right], \tag{92}$$

where

$$Y_{t+k|t} = \mathcal{D}\left(\frac{P_t^*}{P_{t+k}},\, Y_{t+k},\, S_{t+k-1},\, \varkappa_{t+k} \right) \tag{93}$$

is the demand faced by a firm that last reset at time t.

The associated first-order condition implies a nonlinear reset-price formula of the form

$$P_t^* = \mu_t^{p,*} \frac{\mathbb{E}_t \sum_{k=0}^{\infty} (\beta\theta)^k \Lambda_{t+k}\, \Omega_{t+k|t}\, MC_{t+k} Y_{t+k|t}}{\mathbb{E}_t \sum_{k=0}^{\infty} (\beta\theta)^k \Lambda_{t+k}\, \Omega_{t+k|t}\, Y_{t+k|t}}, \tag{94}$$

where $\mu_t^{p,*}$ is the endogenous desired markup for a re-optimizing firm and $\Omega_{t+k|t}$ collects the terms induced by the deep-habits demand schedule.

The inflation mechanism in the model can now be stated more completely. AI affects inflation through both marginal cost and desired markups. A positive AI adoption shock may lower marginal cost if productivity rises quickly or if robot substitution reduces the effective cost of production. But the same shock may raise inflation if it increases bottleneck scarcity, strengthens labor-market power through monopsony, or raises firms' pricing power through deep habits and endogenous customer retention. Since firms adjust prices only gradually, the timing and persistence of these forces matter for the path of inflation.

7.7 Monetary policy

Monetary policy follows a standard nonlinear Taylor rule in gross terms:

$$\log\left(\frac{R_t}{\bar{R}}\right) = \rho_R \log\left(\frac{R_{t-1}}{\bar{R}}\right) + \phi_\pi \log\left(\frac{\Pi_t}{\bar{\Pi}}\right) + \phi_y \log\left(\frac{Y_t}{\bar{Y}}\right) + \varepsilon_t^R. \tag{95}$$

The central bank raises the nominal rate when inflation exceeds target or when output rises above its steady-state level.

7.8 Resource constraint and market clearing

The aggregate resource constraint is

$$Y_t = C_t + G_t + I_t^R + I_t^Z + \mathcal{AC}_t^R + \mathcal{AC}_t^Z, \tag{96}$$

where the adjustment-cost terms are

$$\mathcal{AC}_t^R = \frac{\phi_R}{2}\left(\frac{I_t^R}{I_{t-1}^R} - 1\right)^2 I_t^R, \qquad \mathcal{AC}_t^Z = \frac{\phi_Z}{2}\left(\frac{I_t^Z}{I_{t-1}^Z} - 1\right)^2 I_t^Z. \tag{97}$$

Labor market clearing requires

$$N_t = \int_0^1 L_t(j)\, dj, \tag{98}$$

robot-service market clearing implies

$$R_t = u_t^R K_t^R, \tag{99}$$

and the bottleneck-input market clears at

$$M_t = \int_0^1 M_t(j)\, dj. \tag{100}$$

These conditions complete the model's real side. They also clarify why AI shocks need not be uniformly disinflationary. Higher AI adoption pushes on both sides of the resource constraint: it raises future productive capacity, but it also stimulates current investment in AI and robots. At the same time, it can tighten the bottleneck-input market, alter labor-market power, and strengthen product-market power. The net effect on inflation is therefore a quantitative matter.

7.9 AI shocks and the transmission to inflation

The model is built so that an AI shock can be interpreted in a disciplined way. A natural benchmark is a positive innovation to AI investment conditions, as in (84), or to expected AI profitability through (82) or (83). Such a shock raises the expected return to AI investment, stimulates AI accumulation, and through (73) and (74) improves production possibilities. However, it also raises demand for bottleneck inputs, can alter the composition of factor demand toward robot services, can change firms' effective power in labor markets, and can affect pricing power in product markets.

Five broad forces then determine the inflation response. First, there is a *productivity channel.* Higher A_t, Ξ_t, or Γ_t lowers unit costs and tends to reduce inflation, especially if these gains occur rapidly. Second, there is a *demand-and-investment channel.* Since AI and robot accumulation are endogenous, a favorable AI adoption shock raises current investment demand. This tends to increase output and can push inflation upward, particularly when monetary policy is gradual. Third, there is a *scarcity-cost channel.* If the AI shock raises demand for the bottleneck input in (75) faster than effective capacity expands, then the relative price of that input rises. This raises marginal cost directly and can generate inflation even in the presence of productivity growth.

Fourth, there is a *monopsony channel.* If AI lowers the labor-supply elasticity faced by firms, then the marginal expenditure on labor rises relative to the wage. This increases effective labor cost and therefore marginal cost, even if measured wage inflation alone appears muted. Fifth, there is an *endogenous-markup channel.* If AI strengthens retention and lowers the effective elasticity of demand through deep habits, then firms optimally choose higher desired markups. Inflation can then rise even without a proportionate increase in physical production costs.

The model is therefore well suited to study the empirical question of whether AI has recently acted as a disinflationary force, an inflationary force, or both at different horizons. In the short run, inflation may rise if the demand, scarcity, monopsony, or markup channels dominate. Over longer horizons, inflation may fall if productivity gains diffuse broadly enough.

7.10 Numerical Evaluation

7.10.1 Calibration

The impulse responses reported below are generated using a benchmark calibration designed to keep the model quantitatively disciplined while allowing the AI shock to operate through the main mechanisms emphasized in the paper: productivity, endogenous AI and robot accumulation, bottleneck scarcity, monopsony power in labor markets, and endogenous markups through deep habits.

On the production side, the bottleneck-input share is set to $\alpha = 0.10$. This choice is consistent with the empirical estimates reported earlier in the paper and places the benchmark near the middle of the estimated range, so that the constrained AI-related input is quantitatively relevant without mechanically dominating marginal-cost dynamics. Within value added, the robot-services share is set to $\nu = 0.30$, implying that automation capital plays a meaningful but not overwhelming role relative to effective labor. For the elasticity of substitution between effective labor and robot services, the benchmark sets $\rho = 1$. This corresponds to the Cobb–Douglas case and provides a natural neutral benchmark: it avoids taking a strong ex ante stand on whether labor and automation are gross complements or gross substitutes, while still delivering a transparent baseline against which alternative values can be evaluated. Since the degree of substitutability is both economically important and empirically uncertain, robustness exercises below vary ρ over a broad range.

AI capability affects production through three channels. The neutral productivity elasticity is calibrated at $\phi_A = 0.02578$, while the robot-augmenting and labor-augmenting elasticities are set to $\phi_\Gamma = 0.02063$ and $\phi_\Xi = 0.00516$, respectively. These values are chosen so that the AI Euler equation holds at the steady state, with $\phi_A : \phi_\Gamma : \phi_\Xi = 5 : 4 : 1$. This calibration preserves the idea that AI raises productivity broadly, but does so more strongly through neutral and automation-related channels than through labor augmentation alone. In economic terms, the benchmark reflects a view of recent AI progress in which machine-oriented processes, algorithmic capital, and automated tasks respond more strongly than labor efficiency in a narrow sense.

The bottleneck block is calibrated so that the steady-state relative price of the constrained input is $\bar{p}_M = 0.05$, while the scarcity elasticity is set to $\upsilon = 0.60$. This makes the bottleneck channel operative in the short run without turning it into an implausibly explosive source of inflation. Bottleneck demand also depends directly on AI and robot investment through $\omega_Z = 1$ and $\omega_R = 1$, which ensures that an AI expansion can tighten the bottleneck market not only because current production rises, but also because technology adoption itself requires scarce infrastructure, installation capacity, energy, and other complementary inputs.

For capital accumulation, robot capital depreciates at rate $\delta_R = 0.025$ and AI capability at rate $\delta_Z = 0.015$. The lower depreciation rate for AI capability reflects the idea that AI-related knowledge capital is not as short-lived as some forms of physical equipment, even though it still becomes obsolete over time. Convex adjustment costs are set to $\phi_R = 4.0$ for robot investment and $\phi_Z = 5.0$ for AI investment. This makes AI accumulation somewhat harder to scale rapidly than robot accumulation, which is consistent with the view that building AI capability requires not only expenditure on hardware and software, but also organizational adaptation, data curation, deployment, and complementary restructuring. These adjustment costs are important for generating gradual diffusion and hump-shaped responses rather than implausibly instantaneous jumps in installed technology.

The persistence of the AI investment-efficiency shock is set to $\rho_{IZ} = 0.79$. This value is estimated from the AI-investment data used in the paper using the HP-filtered series discussed earlier, and implies that favorable AI investment conditions dissipate only gradually. As a result, AI adoption is not driven by a one-period impulse, but by a shock persistent enough to generate a meaningful dynamic accumulation process.

In the labor-market block, the steady-state labor-supply elasticity faced by firms is set to $\bar{\varphi}_w = 4.0$, which implies a non-negligible monopsony wedge in steady state while remaining within a plausible range for imperfectly competitive labor markets. The AI sensitivity of labor-market power is set to $\psi_\varphi = 0.30$, so that higher AI capability lowers the effective labor-supply elasticity faced by firms and therefore raises marginal expenditure on labor relative to the wage. This choice makes the monopsony channel quantitatively relevant without allowing it to dominate the entire model.

In the product-market block, the deep-habits retention parameter is set to $\bar{\kappa} = 0.40$, while the AI sensitivity of retention is set to $\psi_\kappa = 0.30$. This implies substantial but not extreme customer retention, and allows greater AI capability to strengthen product-market lock-in and raise desired markups endogenously. Habit persistence is set at $\rho_S = 1.00$, which makes the habit stock highly persistent and therefore allows the markup channel to remain operative over horizons relevant for inflation dynamics. This persistence is helpful in a setting where AI adoption is itself gradual and where pricing-power effects need time to interact with the rest of the model.

Finally, nominal rigidity is governed by a Calvo parameter $\theta = 0.733$, while the baseline Dixit–Stiglitz elasticity is $\varepsilon = 6$. Monetary policy follows a Taylor rule with interest-rate smoothing $\rho_R = 0.65$, an inflation coefficient $\phi_\pi = 2.50$, and an output coefficient $\phi_y = 0.125$. Taken together, this calibration is intended to provide a benchmark in which AI shocks can plausibly generate either disinflationary or inflationary responses depending on horizon and on the relative strength of productivity gains, investment demand, scarcity effects, labor-market power, and product-market power. The values used for the impulse responses are therefore chosen not to force a particular sign of the inflation response, but to keep all of the model's main channels quantitatively active and empirically plausible.

7.10.2 Impulse responses in the one-sector DSGE model

Figure 6 shows that the one-sector DSGE model continues to deliver an inflationary response to an AI investment-efficiency shock even when the mechanism is embedded in a richer general-equilibrium environment with endogenous AI investment, robot substitution, habit persistence, nominal rigidities, and monetary-policy feedback. Output rises on impact, inflation increases immediately, AI investment jumps sharply, and the nominal interest rate responds endogenously through the Taylor rule. The main quantitative lesson is that allowing the shock to propagate through a complete DSGE structure does not overturn the paper's central message. Instead, it reinforces it: AI can be inflationary even in a model that explicitly incorporates productivity improvements and forward-looking investment behavior, because the disinflationary force from efficiency gains is only one part of the transmission mechanism.

A useful feature of the figure is that it compares the "productivity plus accumulation only" configuration with the full model. In both cases, the shock generates a strong increase in AI investment and a positive response of output on impact, which confirms that the endogenous adoption margin is quantitatively important. But the full specification generates a stronger and more persistent response of inflation than the version that retains only productivity and investment channels. This comparison is economically informative. It implies that the demand-and-investment channel is clearly part of the story, but it is not sufficient to summarize the model's inflation dynamics. Once the additional channels are activated—namely bottleneck scarcity, monopsony distortions, and endogenous markup movements through deep habits—the inflation response is amplified rather than dampened.

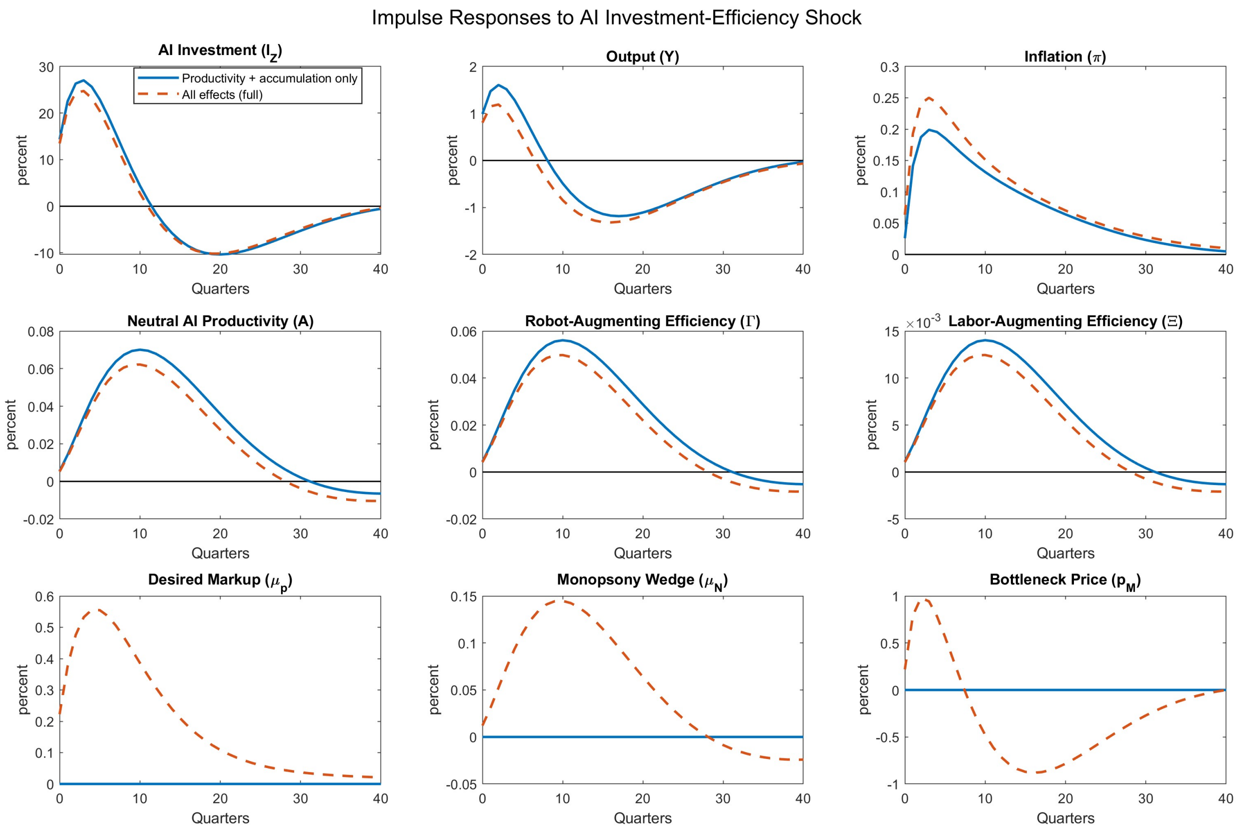


Figure 6: Impulse responses in the one-sector DSGE model following an AI investment-efficiency shock.

The productivity block shown in the middle row clarifies why the comparison is nontrivial. Following the AI investment-efficiency shock, the stock of AI capability rises only because firms increase current investment, and this gradual accumulation raises neutral productivity, robot-augmenting efficiency, and labor-augmenting efficiency over time. These are disinflationary forces in isolation. However, the figure shows that they coexist with a stronger inflation response in the full model. The implication is that the richer set of cost-push and pricing-power channels is strong enough to dominate the purely disinflationary side of the technology block, at least over the horizons relevant for the inflation response.

The bottom row makes those additional channels visible. In the benchmark without bottlenecks, monopsony, and endogenous markups, the desired-markup response and the monopsony wedge remain shut down, and the bottleneck price does not move. In the full model, by contrast, the desired markup rises substantially, the monopsony wedge becomes positive, and the bottleneck price initially jumps sharply. These responses are exactly the mechanisms through which AI can raise inflation even when it also raises efficiency. A more intense use of scarce complementary inputs pushes up marginal costs, greater labor-market power raises the effective cost of labor, and stronger product-market lock-in increases firms' desired pricing margins. The figure therefore shows directly that the inflationary consequences of AI in the full model are not a residual artifact; they are tied to distinct and economically interpretable channels.

The behavior of output is consistent with this interpretation. Output initially increases, reflecting the expansionary effect of higher AI investment and stronger demand for complementary activity, but the subsequent reversal indicates that the same shock also brings forces that restrain real activity over time. In the full model, these restraining forces operate through higher effective costs, tighter scarcity in AI-related inputs, and stronger pricing power. As a result, the economy does not display a pure productivity boom in which higher efficiency mechanically lowers inflation and supports a permanently stronger expansion. Instead, the shock produces a configuration in which inflation rises decisively while the output response is more limited and eventually turns negative. That pattern is exactly what one would expect when cost-push and markup channels materially offset the purely expansionary side of the AI shock.

The nominal interest-rate response fits naturally into this interpretation. Monetary policy reacts to the higher inflation environment by raising the policy rate substantially on impact. This endogenous tightening helps bring both output and inflation back toward steady state over time, but it does not prevent the short-run inflationary effects of the AI shock from appearing. The figure therefore highlights a general-equilibrium mechanism in which AI can generate an inflationary episode even under an active Taylor rule, with policy eventually leaning against the shock rather than eliminating its short-run effects.

7.10.3 Policy functions in the one-sector DSGE model

Overall, the DSGE impulse responses line up closely with the earlier one-sector analysis in the paper. The productivity channel remains disinflationary in isolation, the demand-and-investment channel remains inflationary, and the broader supply-side wedges can overturn or dominate the pure productivity effect when they are sufficiently strong. The DSGE environment strengthens that conclusion because it shows that the inflationary consequences of AI are not an artifact of a stripped-down setup. They survive once one introduces capital accumulation, substitution between labor and automation, habits, policy feedback, and endogenous propagation. The sign of the inflation response is therefore not pinned down by productivity alone. Even if AI raises efficiency, it also stimulates investment, intensifies the use of complementary scarce inputs, affects effective labor-market wedges, and strengthens firms' pricing power. In the benchmark calibration, those additional channels are quantitatively important enough to make the full specification more inflationary than the productivity-plus-accumulation benchmark.

Figures 7 and 8 report policy-function exercises for inflation in the one-sector DSGE model. Each panel varies one parameter around the benchmark while holding the others fixed, and traces the inflation response under both the "productivity plus accumulation only" specification and the full model. The impact figure isolates the immediate response of inflation, while the

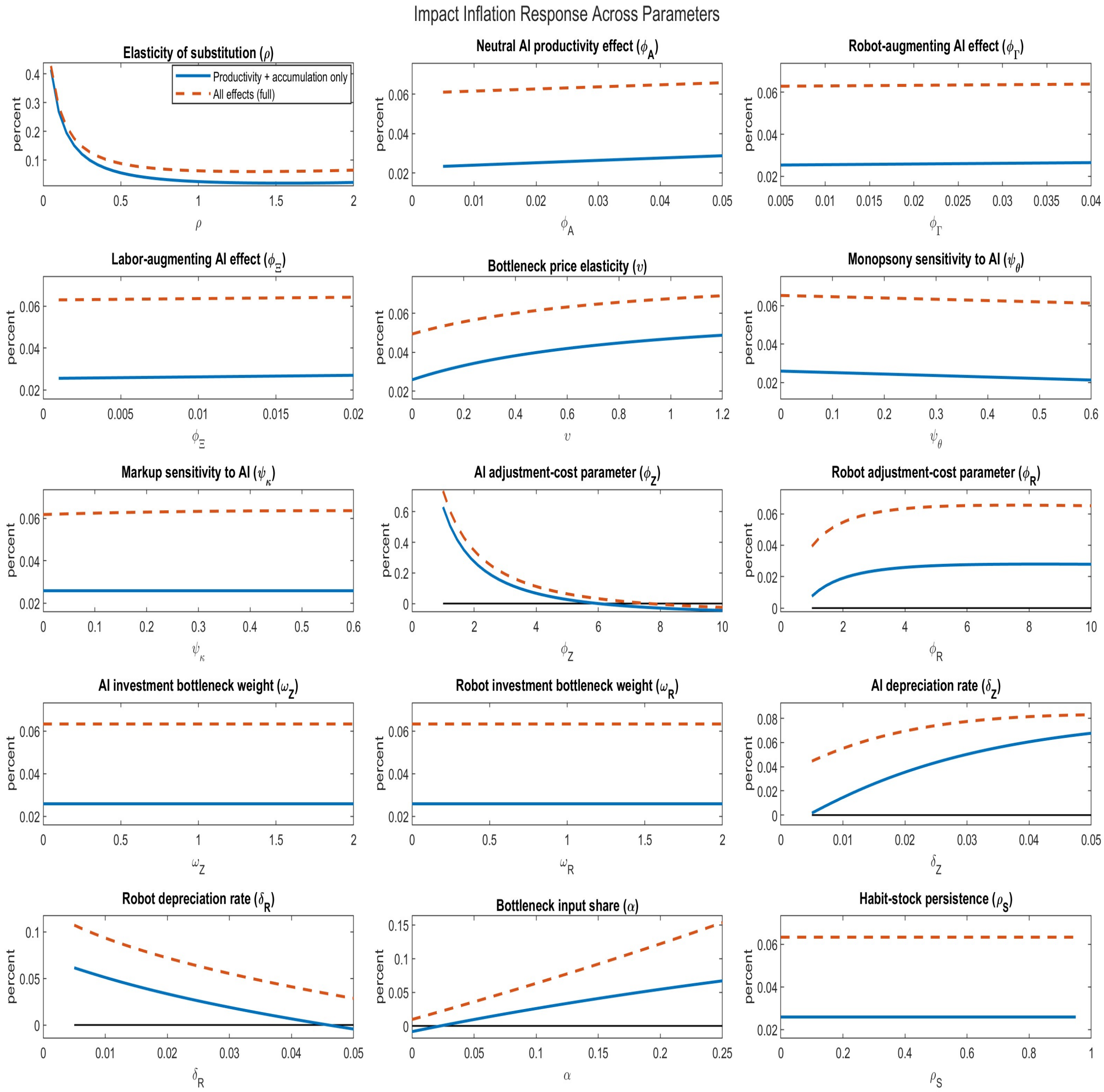


Figure 7: Impact inflation policy functions in the one-sector DSGE model. Each panel varies one parameter while holding the others at their benchmark values.

maximum-response figure is especially useful because several responses in the model are hump-shaped rather than purely front-loaded. The mean and median responses, not shown, are very similar to the maximum-response results, which indicates that the ranking of parameters is not being driven by a single transitory spike but instead reflects the broader inflationary episode generated by the shock.

The elasticity of substitution, ρ, is the most informative place to begin. The positive response of inflation to an increase in the efficiency of investment in AI (hence in ai itself) holds for all values that are considered. In addition, the policy functions show a clear and economically intuitive pattern: inflation is higher when ρ is low and is lower for higher values of ρ. In the model, lower values of ρ correspond to stronger complementarity between labor and robot services, whereas higher values correspond to greater substitutability. When labor and automation are more complementary, an AI investment-efficiency shock raises the joint demand for labor, robot services, and scarce complementary inputs rather than allowing firms to substitute cheaply away from labor. The productive benefits of AI therefore diffuse less through immediate cost-saving substitution and more through a transition that intensifies demand for

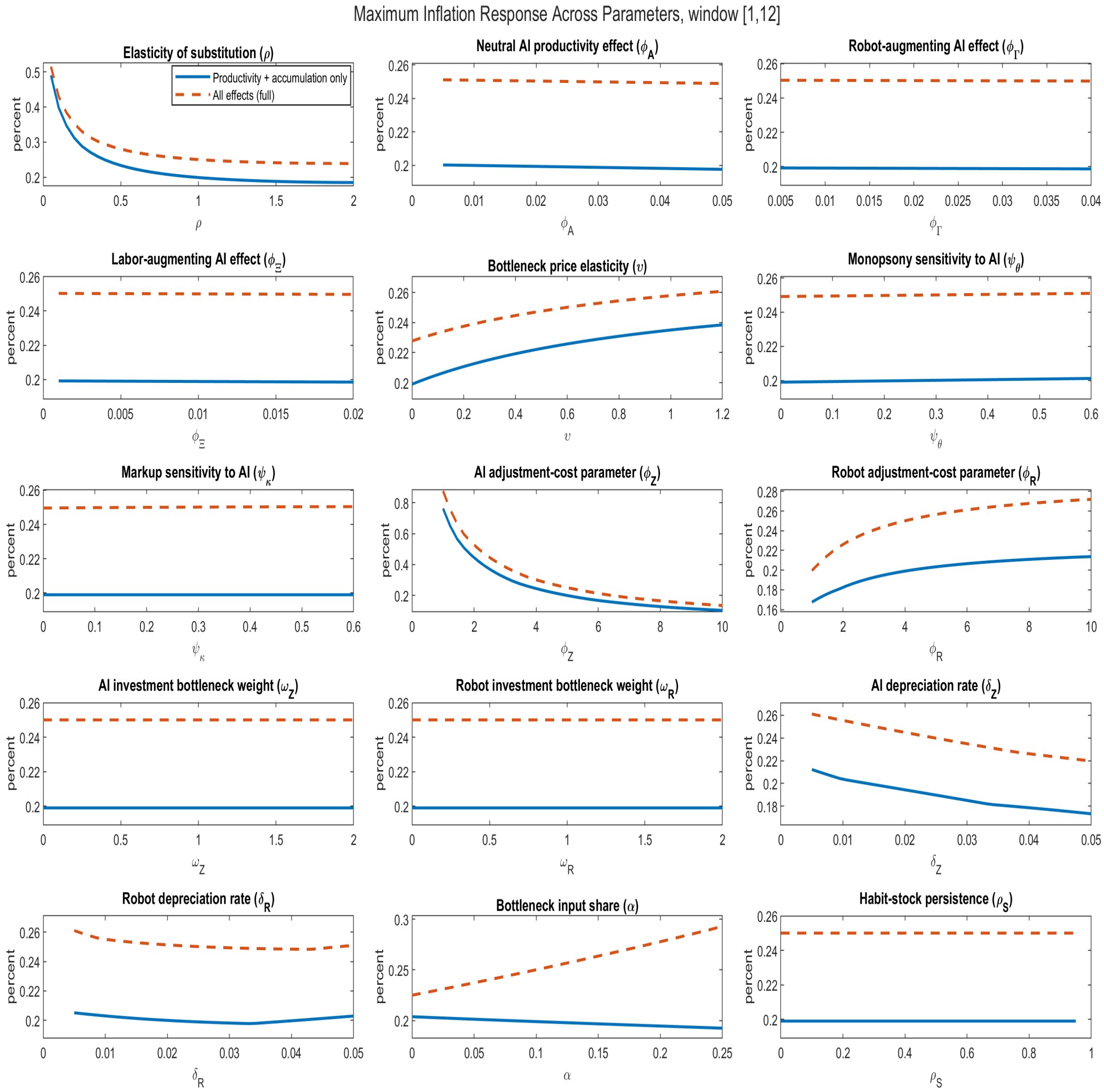


Figure 8: Maximum inflation response policy functions in the one-sector DSGE model. Each panel varies one parameter while holding the others at their benchmark values.

complementary factors and bottleneck inputs, generating inflationary pressures. That is precisely the environment in which inflation responds more strongly. By contrast, when ρ is larger, firms can reorganize production more easily toward the relatively more efficient input mix, so that the shock looks more like a supply expansion and less like a source of short-run cost pressure. This is why the benchmark sets $\rho = 1$: it provides a neutral Cobb–Douglas reference point that does not prejudge whether AI-enabled automation mainly substitutes for labor or instead raises the demand for labor and other complementary inputs. The policy functions then show that moving away from that benchmark in the direction of stronger complementarity makes the shock markedly more inflationary.

The figures also show that the inflation response is not equally sensitive to every productivity or efficiency parameter. Within the ranges considered here, varying ϕ_A, ϕ_Γ, and ϕ_Ξ changes inflation only modestly. On impact, the response is nearly flat in ϕ_Γ and ϕ_Ξ, and increases only slightly with ϕ_A. For the maximum response, the curves are even flatter, especially in the full model. This suggests that, locally around the benchmark calibration, the main qualitative inflation result is not being driven by knife-edge choices about the exact size of the direct efficiency elasticities. Those parameters matter for the broader transmission of the shock, but the policy functions indicate that inflation is considerably more sensitive to other margins.

The bottleneck channel, by contrast, is clearly important. The response of inflation rises with the bottleneck price elasticity υ in both figures, and the increase is stronger in the full model than in the productivity-plus-accumulation case. A higher υ makes the price of the constrained AI-related input more responsive to adoption and production pressure, which feeds more forcefully into marginal cost. The bottleneck input share α is also strongly influential. Both the impact response and the maximum response increase materially with α, especially in the full model. This means that when the scarce AI-related input plays a larger role in production, the inflationary effects of AI adoption are amplified in a systematic way. Together, these two panels provide strong support for the idea that the scarcity mechanism is a quantitatively central part of the model rather than a secondary embellishment.

The labor-market and markup sensitivities play a more nuanced role in these local exercises. The impact response declines slightly as ψ_φ rises, and the maximum response is nearly flat, with only very small changes across the admissible range. Likewise, the inflation response is almost completely flat in ψ_κ in both the impact and maximum figures. This does not mean that monopsony or endogenous markups are irrelevant in the model; the impulse responses show that these channels are active in the full specification. Rather, the policy-function evidence suggests that, around the benchmark calibration and for the inflation statistics plotted here, local variation in these parameters has much smaller quantitative effects than variation in the substitution parameter, the bottleneck parameters, or some of the accumulation parameters. In other words, these channels help distinguish the full model from the stripped-down case, but small changes in their intensity do not substantially alter the inflation response in the neighborhood of the benchmark.

The accumulation side of the model is especially revealing. The AI adjustment-cost parameter ϕ_Z has a large negative effect on both impact inflation and maximum inflation. When AI adjustment costs are high, firms cannot scale up AI adoption rapidly, so the demand-and-deployment surge associated with the shock is damped and inflation falls sharply. The robot adjustment-cost parameter ϕ_R works in the opposite direction: inflation rises with ϕ_R, particularly in the full model. Economically, making robot investment harder to adjust limits the economy's ability to expand automation services quickly in response to the shock. That weakens the immediate supply-side relief that robot deepening could otherwise provide, and therefore leaves more of the short-run inflationary pressure intact. The policy functions therefore show that the inflationary effect of AI depends not only on how easily firms can build AI capability, but also on how easily they can adjust the complementary automation capital needed to translate that capability into production.

The depreciation rates also matter, but in different ways at different horizons. A higher AI depreciation rate, δ_Z, raises the impact response of inflation while lowering the maximum response. This pattern indicates that faster AI depreciation strengthens the short-run adoption impulse, but weakens the medium-run propagation of the shock by making installed AI capability less durable. Robot depreciation, δ_R, has the opposite broad tendency on impact: inflation falls as δ_R rises, because faster robot depreciation weakens the effective accumulation of complementary automation capital. For the maximum response, the effect of δ_R is comparatively modest and somewhat non-monotonic, especially in the full model, which suggests that robot durability matters for inflation, but less cleanly than AI durability once one moves beyond the very short run.

Some parameters turn out to have surprisingly little influence on the plotted inflation statistics. The bottleneck-demand weights on AI and robot investment, ω_Z and ω_R, are essentially flat in both figures. Habit persistence, ρ_S, is also flat. These results indicate that, over the ranges considered here, the inflation response is not especially sensitive to marginal variation in these parameters. Put differently, once the bottleneck and deep-habit mechanisms are switched on, local changes in these particular coefficients do not materially alter the size of either the impact response or the maximum response of inflation. This is itself informative, because it suggests that the model's core inflationary result is not being driven by finely tuned choices for these parameters.

The persistence of the AI investment-efficiency shock mainly governs the duration and propagation of the inflationary episode, which is why the maximum, mean, and median measures are all useful alongside the impact response. Although that parameter is not shown in the two figures reproduced here, the broader policy-function exercises indicate that greater shock persistence strengthens the overall inflationary episode by making firms more willing to expand adoption in response to a favorable change in AI investment conditions. This is consistent with the dynamic logic of the model: when the shock is more persistent, adoption, scarcity, and pricing responses have more time to interact.

Taken together, these exercises sharpen the main quantitative message of the paper. The inflationary consequences of an AI investment-efficiency shock are strongest when labor and automation are more complementary, when bottleneck scarcity is more important, when the constrained input carries more weight in production, and when AI adoption can respond forcefully while complementary robot capital adjusts only sluggishly. By contrast, the exact magnitudes of the direct productivity elasticities and small local changes in the monopsony, markup, bottleneck-demand-weight, and habit-persistence parameters matter less for the inflation statistics shown here. The full model therefore behaves as a genuinely multi-channel environment, but the policy functions make clear that the dominant quantitative drivers of inflation are the production structure, the scarcity mechanism, and the dynamics of technology accumulation.

Finally, Appendix H extends the benchmark model to a two-sector economy with a high-AI-intensive sector and a low-AI-intensive sector. The extension preserves the main mechanisms of the one-sector framework—endogenous AI capability, robot-capital accumulation, bottleneck inputs, monopsony power, endogenous markups, and Calvo price rigidity—but allows these channels to operate with different intensities across sectors. Households consume a CES aggregate of the two sectoral goods, so sectoral relative prices affect the allocation of expenditure across sectors. This creates an additional reallocation channel absent from the one-sector model: an AI shock can change not only within-sector marginal costs, markups, and inflation, but also the relative price of high- and low-AI goods and therefore the composition of aggregate demand. The two-sector impulse responses show that the main conclusion of the paper is robust to sectoral heterogeneity. When only the productivity and capital-accumulation channels are active, AI shocks have relatively muted inflationary effects and behave more like conventional efficiency improvements. Once bottleneck scarcity, monopsony, and endogenous markup channels are activated, sectoral inflation rises more strongly and more persistently, and aggregate inflation reflects both the direct response of the shocked sector and spillovers through expenditure reallocation, common consumption, and monetary policy. The two-sector model therefore reinforces the central interpretation of the paper: AI is not necessarily disinflationary in transition, because the productivity gains associated with AI adoption can be offset, and in some cases dominated, by investment demand, scarce complementary inputs, labor-market wedges, and pricing-power effects. The appendix presents the full structure, parameterization, and impulse-response analysis.

8 Impact of AI on the Long-Run Inflation Rate

This section studies the implications of artificial intelligence for the long-run inflation rate. The discussion proceeds in two steps. We begin with the empirical evidence, using long-horizon projections of GDP-deflator inflation as a function of alternative AI measures. We then introduce a parsimonious structural framework that rationalizes these patterns and clarifies the mechanisms through which AI can raise inflation over the medium to long run.

8.1 Long-run projection methodology for GDP-deflator inflation as a function of AI measures

This exercise constructs projected medium- to long-horizon paths for GDP-deflator inflation as a function of alternative AI measures. The procedure is reduced-form and predictive rather than structural. It does not estimate a full New Keynesian model, nor does it identify a causal impulse response from a complete dynamic system. Instead, it estimates horizon-specific forecasting regressions and uses those estimates to trace out scenario paths under alternative values of the AI variable.

The starting point is a quarterly panel of standardized AI measures. The broader empirical analysis considers several variants, including investment, capital services, patents, and composite measures. For the long-run projection exercise emphasized below, however, the discussion focuses on measures that exclude data centers. This restriction is useful for two reasons. First, data-center series begin much later than the other AI measures, which shortens the usable sample and weakens long-horizon estimation. Second, data centers are best interpreted as a specific infrastructure-intensive phase of AI deployment rather than as a general long-run measure of AI diffusion. Excluding them therefore yields a more conservative and cleaner benchmark for studying whether the inflationary relationship between AI and prices remains present even outside that infrastructure channel.

These AI series are matched to quarterly GDP-deflator inflation, measured at annualized quarterly rates. After constructing the inflation series, the AI panel and the inflation series are aligned on their common quarterly support so that all regressions are estimated on a consistent sample. Additional conditioning variables are formed from lagged GDP-deflator inflation.

For each AI variant, the methodology estimates a separate predictive regression at each horizon $h = 1, \dots, H$. The dependent variable is the forward average of GDP-deflator inflation over the next h quarters. Thus, letting π_t^{GDP} denote GDP-deflator inflation, the horizon-h outcome is

$$\bar{\pi}_{t,h}^{GDP} \equiv \frac{1}{h} \sum_{j=1}^{h} \pi_{t+j}^{GDP}.$$

The corresponding regression takes the form

$$\bar{\pi}_{t,h}^{GDP} = a_h + b_h z_t + c_h' X_t + u_{t,h},$$

where z_t is the AI measure at time t, X_t is the vector of lagged controls, and $u_{t,h}$ is an error term. The coefficient b_h is the key reduced-form object of interest: it measures how the future average of GDP-deflator inflation over horizon h is associated with the current AI state, conditional on the included controls. Standard errors are computed using HAC corrections.

Once the sequence of coefficients $\{b_h\}_{h=1}^{H}$ has been estimated, three counterfactual scenario paths are constructed for each AI measure. Let z_{t_0} denote the baseline AI level at a chosen starting date t_0. The three scenarios are

$$\text{baseline: } z_{t_0}, \qquad \text{high-AI: } z_{t_0}+1, \qquad \text{low-AI: } z_{t_0}-1,$$

where the shifts of $+1$ and -1 are in standardized units. For each horizon h, the estimated regression coefficients are used to generate fitted values of future GDP-deflator inflation under each of these three AI states, holding the control variables fixed at their baseline values. In this sense, the procedure is a scenario-projection device: it asks how the predicted path of GDP-deflator inflation changes when the AI variable is moved up or down by one standard deviation relative to baseline.

The projected paths are then anchored to a benchmark value at horizon 0. For GDP-deflator inflation, the benchmark is set at 2 percent. Rather than allowing the projected path to jump immediately from the benchmark to the regression-implied fitted value, the projection uses a smooth anchoring rule. Let $\widehat{\pi}_h$ denote the fitted value from the regression at horizon h, and let π^* denote the benchmark. The anchored path is defined as

$$\pi_h^{\text{anchored}} = \pi^* + w_h\big(\widehat{\pi}_h - \pi^*\big),$$

where

$$w_h = 1 - e^{-h/\tau}.$$

This guarantees that the projected path begins at the benchmark at horizon 0 and gradually converges toward the regression-implied prediction as the horizon increases. The anchoring is therefore presentational: it smooths the transition from the benchmark initial condition to the estimated medium- and long-run projection.

Confidence bands are computed from the variance–covariance matrix of the estimated regression coefficients and therefore reflect uncertainty in the fitted values implied by the predictive regressions. The methodology produces two complementary graphical objects. The first is a set of anchored level paths, which display the projected trajectories of GDP-deflator inflation under the low-AI, baseline, and high-AI scenarios. The second is a set of incremental effect plots, which display the difference between the high-AI and baseline scenarios, and between the low-AI and baseline scenarios. The resulting objects should be interpreted as reduced-form predictive projections rather than structural impulse responses.

8.2 Empirical long-run inflation projections without data centers

Figure 9 reports anchored long-run projection paths for GDP-deflator inflation when the analysis is restricted to AI measures that do not include data-center-based indicators. Across all six specifications, a higher AI state is associated with a persistently higher projected inflation path, while a lower AI state is associated with a lower projected inflation path. The strongest projected responses are obtained for the investment-based and capital-services measures. In the specifications based on investment excluding data centers and on capital services, a higher-AI scenario produces a marked upward drift in inflation over the projection horizon, with the lower-AI scenario moving symmetrically below the benchmark path. This indicates that even when data centers are excluded, the capital-deepening and production-side dimensions of AI remain strongly associated with inflationary pressure.

The broader composite measures convey a similar message. Both the composite excluding data centers and the composite excluding data centers and capital services imply a clear positive

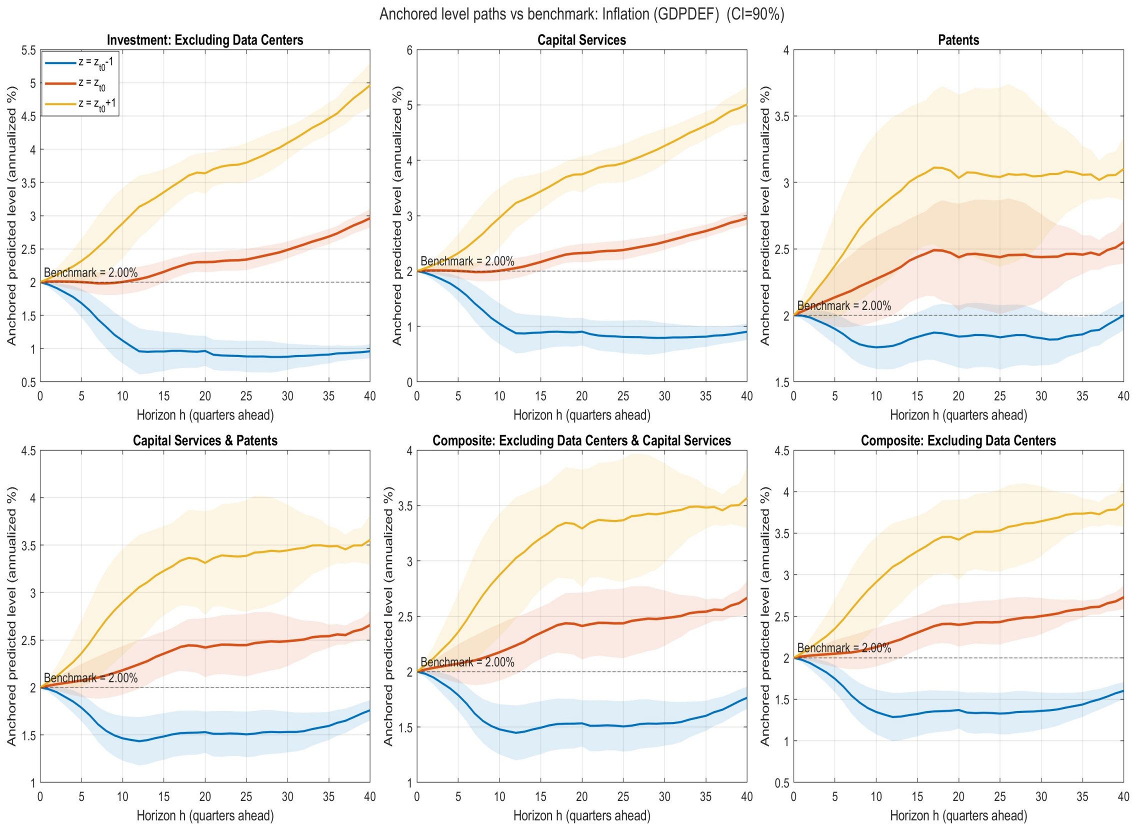


Figure 9: Anchored long-run projection paths for GDP-deflator inflation under alternative AI scenarios for the specifications that exclude data-center-based measures. In each panel, the middle path corresponds to the baseline AI level at the projection date, while the upper and lower paths correspond to one-standard-deviation increases and decreases in the AI measure, respectively. Shaded areas denote 90% confidence intervals. The dashed horizontal line marks the 2% inflation benchmark.

inflation gradient across scenarios, although the magnitude is somewhat smaller than in the pure investment and capital-services cases. The specification combining capital services and patents also shows a positive effect, again pointing to an inflationary role for AI-related accumulation and deployment even outside the data-center channel.

By contrast, the patents-only specification is noticeably flatter. Its high-AI scenario still lies above the baseline and its low-AI scenario below it, but the separation between paths is smaller and the confidence bands are wider. This suggests that patent-based measures are less tightly linked to future GDP-deflator inflation dynamics than measures based more directly on realized investment, capital services, and broader deployment-related activity.

Taken together, the empirical projections support two conclusions. First, the inflationary association of AI does not depend on the inclusion of data centers. Second, the magnitude of the projected effect varies systematically across measures, with investment- and capital-based proxies delivering the strongest projected inflation responses and patents delivering the weakest.

8.3 A structural interpretation of the long-run inflation evidence

The empirical long-run projections show that higher AI intensity is associated with persistently higher GDP-deflator inflation paths even when data-center-based measures are excluded. The purpose of this subsection is to provide a simple structural interpretation of that evidence. The model is not meant to replace the projection exercise or to turn it into a structural forecast. Rather, it clarifies how the empirical pattern can arise in a forward-looking New Keynesian environment when AI diffuses gradually into effective production.

The key distinction is between the AI frontier and effective AI use. A major AI innovation may shift the technological frontier permanently, but firms do not immediately use the full frontier technology in production. Effective AI depends on adoption, organizational adjustment, complementary capital, software integration, workforce reallocation, and infrastructure. These adjustment margins imply that the macroeconomic effects of AI should unfold gradually even if the underlying frontier improvement is permanent. This distinction is important for interpreting the empirical results: the projected inflation paths do not jump immediately to a new level; instead, they rise gradually over medium- and long-run horizons.

Let $\bar{z}_t$ denote the frontier component of AI intensity. We study a permanent step in the frontier at $t = 0$:

$$\bar{z}_t = \begin{cases} 0, & t < 0, \\ \Delta \bar{z}, & t \geq 0. \end{cases} \tag{101}$$

The variable that enters production, marginal cost, demand, and inflation is not the frontier itself, but effective AI intensity, denoted by z_t. We assume that effective AI diffuses only gradually toward the new frontier. A convenient way to write this is in terms of the adoption gap,

$$\tilde{z}_t \equiv z_t - \bar{z}_t,$$

which follows

$$\tilde{z}_t = \rho_z \tilde{z}_{t-1}, \qquad 0 < \rho_z < 1. \tag{102}$$

If the economy starts from $z_{-1} = 0$ and the frontier shifts permanently at $t = 0$, then $\tilde{z}_0 = -\Delta \bar{z}$. Hence

$$\tilde{z}_t = -\rho_z^t \Delta \bar{z},$$

and effective AI evolves according to

$$z_t = \bar{z}_t + \tilde{z}_t = \left(1 - \rho_z^t\right) \Delta \bar{z}, \qquad t \geq 0. \tag{103}$$

Thus z_t starts below the new frontier and converges gradually to $\Delta \bar{z}$. The parameter ρ_z governs the speed of diffusion. A higher value of ρ_z means slower adoption and a more persistent transition.

The macroeconomic block is the same as in Section 6, except that AI enters through effective AI intensity z_t:

$$x_t = \mathbb{E}_t x_{t+1} - \frac{1}{\sigma}\Big(i_t - \mathbb{E}_t \pi_{t+1} - r_t^n\Big) + \lambda_d \phi_D z_t, \tag{104}$$

$$\pi_t = \beta \mathbb{E}_t \pi_{t+1} + \kappa \lambda x_t + \kappa \Delta z_t, \tag{105}$$

$$i_t = r_t^n + \phi_\pi \pi_t + \phi_x x_t. \tag{106}$$

AI affects inflation through two objects. The first is the demand term $\lambda_d \phi_D z_t$, which captures AI-related investment, adoption spending, and complementary capital formation. The second is the composite supply-side wedge,

$$\Delta \equiv -\phi_A + \alpha \phi_{pM} + \phi_\mu + \phi_M. \tag{107}$$

The term $-\phi_A$ is the disinflationary productivity channel. The remaining terms capture AI-input bottlenecks, product-market markups, and labor-market wedges. Hence $\Delta > 0$ means that the cost-raising supply-side channels dominate productivity in the Phillips curve.

The long-run effect is obtained by asking where the economy settles once diffusion is complete. In the long run, $z_t = \bar{z}_t = \Delta\bar{z}$ is constant, and expectations satisfy

$$\mathbb{E}_t x_{t+1} = x_t, \qquad \mathbb{E}_t \pi_{t+1} = \pi_t.$$

The long-run deviations $(\bar{x}, \bar{\pi})$ therefore satisfy

$$(\phi_\pi - 1)\bar{\pi} + \phi_x \bar{x} = \sigma\lambda_d\phi_D\Delta\bar{z}, \tag{108}$$

$$(1-\beta)\bar{\pi} - \kappa\lambda\bar{x} = \kappa\Delta\Delta\bar{z}. \tag{109}$$

Solving these two equations gives

$$\bar{x} = A^x_{LR}\Delta\bar{z}, \qquad \bar{\pi} = A^\pi_{LR}\Delta\bar{z}, \tag{110}$$

where

$$A^\pi_{LR} = \frac{\kappa\lambda\,\sigma\,\lambda_d\phi_D + \phi_x\kappa\Delta}{(\phi_\pi - 1)\kappa\lambda + (1-\beta)\phi_x}, \qquad A^x_{LR} = \frac{(1-\beta)\sigma\,\lambda_d\phi_D - (\phi_\pi - 1)\kappa\Delta}{(\phi_\pi - 1)\kappa\lambda + (1-\beta)\phi_x}. \tag{111}$$

These expressions show how the long-run inflation effect of AI depends on two forces. The first is the demand channel, $\lambda_d\phi_D$, which raises inflation through higher desired spending and investment. The second is the supply-side wedge, Δ, which raises inflation when AI-input costs, markups, and labor-market wedges dominate the productivity effect. Importantly, the model does not require the demand channel to remain permanently large for AI to raise the long-run inflation rate. Even if the demand effect becomes small in the long run, the supply-side block alone can generate a higher long-run inflation rate provided that the inflationary components of Δ dominate the disinflationary productivity component. Monetary policy dampens these effects through ϕ_π and ϕ_x, but does not mechanically eliminate them. Thus the model can rationalize a positive long-run association between AI intensity and inflation whenever either the demand channel remains sufficiently strong or the net supply-side wedge is positive.

8.4 Transition dynamics under gradual diffusion

The previous subsection characterizes the long-run endpoint after effective AI has fully converged to the new frontier. This subsection characterizes the transition path toward that endpoint. The distinction is important. The long-run mapping in Equation (111) answers the question: what happens once diffusion is complete? The transition dynamics answer a different question: how do inflation and the output gap move while effective AI is still gradually diffusing through production?

To derive the transition path, we use the same New Keynesian block in Equations (104)–(106). Since effective AI is the state variable, we conjecture linear decision rules of the form

$$x_t = A_x z_t, \qquad \pi_t = A_\pi z_t. \tag{112}$$

These rules say that, along the transition, the output gap and inflation are proportional to the amount of effective AI that has already diffused into the economy. The coefficients A_x and A_π are impact loadings with respect to effective AI, not with respect to the frontier itself.

Using the diffusion law, the expected future AI state satisfies

$$\mathbb{E}_t z_{t+1} = \rho_z z_t + (1-\rho_z)\Delta\bar{z}. \tag{113}$$

Equivalently, if the system is written in deviations from the new long-run level, the relevant state follows an AR(1) process with persistence ρ_z. Substituting the decision rules into the IS curve, the NKPC, and the Taylor rule yields a 2×2 linear system. In deviation form, the determinant of this system is

$$\Omega = \left(1 - \rho_z + \frac{\phi_x}{\sigma}\right)(1-\beta\rho_z) + \frac{\kappa\lambda}{\sigma}\left(\phi_\pi - \rho_z\right), \tag{114}$$

and the reduced-form coefficients are

$$A_x = \frac{(1-\beta\rho_z)\lambda_d\phi_D - \frac{\phi_\pi - \rho_z}{\sigma}\kappa\Delta}{\Omega}, \tag{115}$$

$$A_\pi = \frac{\kappa\lambda\lambda_d\phi_D + \kappa\Delta\left(1-\rho_z+\frac{\phi_x}{\sigma}\right)}{\Omega}. \tag{116}$$

Equations (115)–(116) have the same economic interpretation as the long-run formulas, but they apply along the diffusion path rather than only at the endpoint. Inflation rises with effective AI when the demand term and the composite supply-side wedge are positive and sufficiently large. The output-gap response can be positive or muted depending on whether demand and productivity forces dominate the cost-raising supply-side wedges. The persistence parameter ρ_z governs how much of the future AI state is anticipated and therefore how strongly forward-looking agents respond today.

Combining the coefficient A_π with the diffusion path for z_t gives the inflation transition:

$$\pi_t = A_\pi z_t = A_\pi\left(1-\rho_z^t\right)\Delta\bar{z}, \qquad t \geq 0. \tag{117}$$

Similarly,

$$x_t = A_x z_t = A_x\left(1-\rho_z^t\right)\Delta\bar{z}. \tag{118}$$

These expressions make clear why the model generates gradual inflation adjustment. Inflation does not jump immediately to its long-run level because effective AI does not jump immediately to the frontier. Instead, inflation rises as AI becomes operational in production, investment, pricing systems, and complementary infrastructure. This is the structural counterpart of the empirical long-horizon projections: the projection exercise documents a persistent association between current AI intensity and future average inflation, while the diffusion model explains why such persistence can arise even after a permanent frontier improvement.

8.5 Quantitative illustration and connection to the empirical evidence

Figure 10 reports the annualized inflation rate implied by the benchmark calibration under gradual diffusion of effective AI. The figure compares three cases: a demand-only case, a supply-only case, and an all-effects case in which both the demand channel and the full supply-side wedge are active. The model delivers a positive medium-run effect of AI on inflation even though inflation does not jump on impact. Instead, inflation begins at the 2% benchmark and then rises gradually as AI diffuses. This mirrors the logic of the empirical projections, where higher AI scenarios imply persistently higher inflation paths over the projection horizon rather

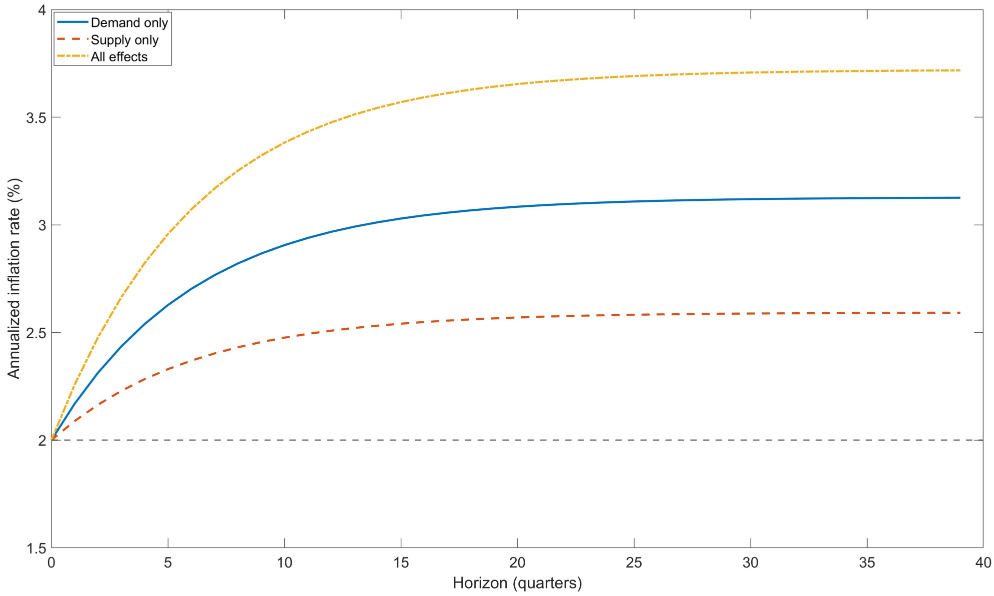


Figure 10: Annualized inflation after a permanent AI frontier step with gradual diffusion of effective AI. Inflation starts at the 2% benchmark and then evolves as effective AI rises over time. The three paths correspond to a demand-only case, a supply-only case, and an all-effects case.

than an instantaneous move to a new level.

The decomposition across cases clarifies the underlying mechanism. In the demand-only case, inflation rises because AI stimulates aggregate demand. In the supply-only case, inflation also rises because the benchmark calibration implies a positive composite supply-side wedge: the cost-raising effects associated with bottlenecks, markups, and labor-market wedges dominate the disinflationary productivity term. When all effects are combined, inflation rises the most, since demand expansion and the positive net supply-side channel reinforce one another.

This structural pattern aligns closely with the empirical evidence reported above. In the data, measures tied to investment, capital services, and broader deployment-oriented activity produce the strongest positive inflation projections, while patents produce noticeably weaker responses. The model provides a natural interpretation of that ranking. AI can be inflationary not because productivity is absent, but because productivity is only one component of a broader transition in which complementary inputs are scarce, pricing power may rise, and labor-market frictions remain relevant. In that environment, deployment-oriented AI measures are precisely the measures most likely to be associated with upward pressure on inflation.

The exclusion of data centers from the empirical benchmark fits naturally within this interpretation. Data centers capture an important infrastructure channel, but they are also a later and more specific phase of AI deployment. Because the data-center series begins much later in the sample, it is less suitable for the long-horizon projection exercise used here. The empirical benchmark therefore focuses on AI measures that exclude data centers, showing that the positive long-run inflation relationship is not driven solely by that infrastructure component. This choice can also be interpreted as a *conservative* scenario in which the pace of future data-center expansion slows relative to the recent deployment surge. However, if data centers continue to be built over the longer run, even at a slower pace than in the recent period, then the estimated long-run inflation effects reported here are likely to understate the full inflationary impact of AI diffusion. The model is consistent with this strategy: it is written in terms of a general AI frontier and its diffusion into effective production, not in terms of any single physical input. As a result, it can explain why inflation remains positively related to AI even when one abstracts from the data-center channel, while also allowing the data-center channel to amplify inflationary pressures when infrastructure expansion remains active.

Taken together, the empirical projections and the model point to the same conclusion. Over the transition considered here, AI is not simply a force that lowers marginal cost through productivity. It is also a force that expands demand and places pressure on complementary inputs and pricing margins. Consequently, the economy converges gradually to a higher long-run inflation path, with that adjustment unfolding over time as effective AI diffuses through production. The model thus provides a coherent explanation for the empirical finding that AI is associated with persistently positive inflation effects.

9 Conclusion

This paper studies the effect of artificial intelligence on inflation using a combination of new measurement, reduced-form evidence, sector-level analysis, and structural modeling. The starting point is that AI is not directly observed as a single macroeconomic series, and that its inflationary effects are unlikely to operate through only one channel. To address this, the paper constructs quarterly U.S. AI proxies that capture multiple dimensions of development and diffusion, including investment, capital services, infrastructure, and innovation. Across these measures and across estimation approaches, the empirical evidence points to a consistent conclusion: over the sample considered, AI has exerted positive pressure on inflation on average, and that effect has become more pronounced in the later years of the sample.

The aggregate evidence shows that the contribution of AI to inflation is not only positive on average, but also heterogeneous across measures in economically meaningful ways. Measures tied more closely to realized deployment, capital accumulation, and infrastructure scaling produce the strongest inflationary associations, while patent-based measures produce much weaker effects. This pattern suggests that the inflationary footprint of AI is especially visible during phases of diffusion in which complementary investment and deployment intensify. The results therefore caution against viewing AI exclusively through the lens of productivity and disinflation. Even if AI improves efficiency over time, the path through which it is scaled and deployed can generate upward pressure on prices through demand expansion, scarce complementary inputs, and related pricing frictions.

The sector-level evidence reinforces this interpretation. Industries with greater ex ante exposure to AI-relevant tasks experienced systematically higher producer-price inflation following the post-2022 acceleration in generative AI. These effects are robust across alternative exposure measures, treatment definitions, and empirical designs. This matters because it shows that the aggregate results are not driven only by a small number of macro time-series co-movements. Instead, the inflationary footprint of AI also appears in the cross-section of industries most exposed to AI-intensive production structures, providing an additional and more disaggregated layer of support for the central empirical findings of the paper.

The structural analysis helps explain why these empirical patterns arise. In the model, AI affects inflation through multiple channels: a productivity channel that lowers marginal cost, a demand channel linked to AI-related investment and complementary capital formation, and additional supply-side wedges associated with input scarcity, pricing power, and labor-market frictions. The key implication is that the net inflation effect of AI is not mechanically disinflationary even when productivity rises. Under plausible parameterizations, the demand

and cost-raising channels can dominate the productivity channel during the diffusion phase, producing inflation that rises above target and adjusts only gradually toward its long-run level. The structural estimates also support this interpretation. The estimated aggregate AI supply-side wedge is positive, and when its components are measured separately—productivity, AI-input prices, labor-market power, and product-market markups—their sum delivers the same qualitative conclusion and, to a large degree, a similar quantitative magnitude.

The long-run inflation exercises provide an additional implication of the analysis. They suggest that the inflationary effects of AI are not merely short-lived impact effects, but may persist over medium- to long-run horizons. This persistence is consistent with the model's diffusion logic: AI does not affect the macroeconomy only at the moment of innovation, but as it is gradually embedded in production, investment, infrastructure, pricing systems, and organizational practices. The projected long-run paths therefore reinforce the central message of the paper: AI can raise inflation not only through temporary deployment costs, but also through a persistent interaction between demand expansion, complementary-input bottlenecks, and market-structure changes.

The analyses of this paper imply that AI should be treated as a macroeconomic force whose inflation consequences depend on the interaction of technological progress, investment dynamics, complementary-input bottlenecks, and market structure. For monetary policy, this means that rapid AI diffusion need not translate immediately into disinflation and may instead coincide with sustained price pressures, especially during infrastructure- and deployment-intensive phases. More broadly, the paper suggests that understanding the inflationary implications of AI requires moving beyond one-dimensional narratives. AI can simultaneously raise productivity and raise inflation, depending on how it diffuses and which channels dominate at a given stage of adoption. This perspective opens several avenues for future research, including richer measurement of AI deployment, deeper identification of sectoral transmission mechanisms, and quantitative models that more fully integrate AI accumulation, infrastructure constraints, and endogenous market structure into macroeconomic analysis.

References

Acemoglu, D. (2024). The simple macroeconomics of AI. *Economic Policy 40*, 13–58.

Adams, Z., H. Fang, J. Liu, and Y. Wang (2025). The rise of AI pricing: Trends, driving forces, and implications for firm performance. Working Paper 2024-33, Federal Reserve of San Francisco.

Alderucci, D., L. Branstetter, E. Hovy, P. Tambe, S. Baviskar, N. Goldschlag, A. Runge, and N. Zolas (2024). Quantifying the impact of AI on productivity and labor demand: Evidence from u.s. census microdata. Working paper. See `https://conference.nber.org/conf_papers/f204793.pdf`.

Babina, T., A. Fedyk, A. He, and J. Hodson (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics 151*, 103745.

Basso, H. S. and O. Rachedi (2025). Robot adoption and inflation dynamics. *Journal of Monetary Economics 156*, 103856.

Borowski, J., J. Fidrmuc, and K. Jaworski (2025). Artificial intelligence adoption and producer price inflation in the EU. *Economics Letters 257*, 112638.

Brynjolfsson, E., D. Li, and L. Raymond (2025). Generative AI at work. *Quarterly Journal of Economics 140*.

Brynjolfsson, E., D. Rock, and C. Syverson (2017). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. NBER Working Paper 24001, National Bureau of Economic Research. Available at `https://www.nber.org/system/files/working_papers/w24001/w24001.pdf`.

Brynjolfsson, E. and S. Unger (2023). The macroeconomics of artificial intelligence. Working paper. Available at `https://www.imf.org/en/publications/fandd/issues/2023/12/macroeconomics-of-artificial-intelligence-brynjolfsson-unger`.

Callaway, B. and P. H. C. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Chirinko, R. S. (2008). Σ: The Long and Short of It. *Journal of Macroeconomics 30*(2), 671–686.

Crouzet, N., J. C. Eberly, A. L. Eisfeldt, and D. Papanikolaou (2022). The economics of intangible capital. *Journal of Economic Perspectives 36*(3), 29–52.

Dube, A., A. Manning, and S. Naidu (2020). Monopsony in labor markets. In *Handbook of Labor Economics*.

Faria-e Castro, M. and S. Ozkan (2026, March). Can ai optimism raise inflation? what a standard macro model says. Federal Reserve Bank of St. Louis, On the Economy Blog. Accessed May 24, 2026.

Firooz, H., Q. Liu, and N. Wang (2025). Automation and the rise of superstar firms. *Journal of Monetary Economics 151*(103733).

Korinek, A. (2024, September). Economic policy challenges for the age of AI. NBER Working Paper 32980, National Bureau of Economic Research.

Luccioni, S., Y. Jernite, and E. Strubell (2024). Power hungry processing: What's driving the cost of AI deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, New York, NY, USA, pp. 204–216. Association for Computing Machinery.

Luccioni, S., B. Trevelin, and M. Mitchell (2024). The environmental impacts of AI – policy primer. Technical report, Hugging Face. Policy primer.

Melina, G. and S. Villa (2025). From servers to rates: AI, ict capital, and the natural rate. IMF Working Paper No. 25/224, International Monetary Fund, Washington, D.C.. See `https://www.imf.org/-/media/files/publications/wp/2025/english/wpiea2025224-source-pdf.pdf`.

Noy, S. and W. Zhang (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science 381*, 187–192.

Nygaard, V. M., J. W. Welburn, and A. Nergui (2025). Macroeconomic implications of artificial intelligence. RAND Working paper. See `https://www.rand.org/pubs/perspectives/PEA3888-3.html`.

Oberfield, E. and D. Raval (2021). Micro data and macro technology. *Econometrica 89*(2), 703–732.

Tambe, P., L. M. Hitt, D. Rock, and E. Brynjolfsson (2020). The productivity effects of AI adoption: Firm-level evidence. NBER Working Paper No. 28285. See https://www.nber.org/papers/w28285.

Tinoco, G., A. Abeliuk, and J. Ruiz del Solar (2025). Impact of price inflation on algorithmic collusion through reinforcement learning agents. Working paper. Available at https://arxiv.org/abs/2504.05335.

Turegeldinova, A. et al. (2025). AI as a centripetal technology: Price compression, homogenization, and entry. Working paper. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5579910.

Appendix

A Construction and data sources for AI activity indicators

This appendix describes the data sources and transformations used to construct the AI-related indicators and composite indices. It documents the underlying national-accounts series, the construction of capital services, the definition of the data-center, patent, adoption, and capability measures, and the formulas used for standardization and aggregation.

A.1 AI-related investment indicators

The investment-based AI indicators are constructed from quarterly national-accounts series on real private fixed investment in information processing equipment, software, computers and peripheral equipment, communication equipment, and research and development. These series are aligned to quarter-end dates, restricted to the sample used in the empirical analysis, and combined into broad investment measures intended to capture AI-related capital formation.

Let

$$I_t^{\text{IPE}}, \qquad I_t^{\text{Soft}}, \qquad I_t^{\text{Comp}}, \qquad I_t^{\text{Comm}}, \qquad I_t^{\text{R\&D}}$$

denote the corresponding real investment flows. The AI-related investment measures are constructed as combinations of these components, and each resulting flow is converted into an index normalized to 100 at the first quarter in which it is observed:

$$A_t^{\text{Inv}} = 100 \times \frac{I_t^{\text{Inv}}}{I_{t_0}^{\text{Inv}}}, \qquad t_0 = \min\{t : I_t^{\text{Inv}} \text{ is observed}\}.$$

In addition to these broader investment measures, the information-processing-equipment series is also retained separately as its own index:

$$A_t^{\text{IPE}} = 100 \times \frac{I_t^{\text{IPE}}}{I_{t_0}^{\text{IPE}}}.$$

A.2 Capital-stock and capital-services index

The capital-services index is built from the same broad set of AI-relevant asset classes: software, computers and peripheral equipment, communication equipment, research and development, and nonresidential structures. Communication equipment is available at annual frequency and is converted to quarterly frequency before entering the construction.

For each asset a, a quarterly depreciation rate δ_a is obtained from an assumed annual depreciation rate δ_a^{ann} via

$$\delta_a = 1 - (1 - \delta_a^{\text{ann}})^{1/4}.$$

Let I_t^a denote the real investment flow for asset a. The corresponding capital stock K_t^a is constructed using a perpetual-inventory formula:

$$K_t^a = (1 - \delta_a)K_{t-1}^a + I_t^a.$$

The initial stock is chosen using a simple steady-state relation between the initial investment flow and depreciation.

To aggregate the asset-specific stocks into a capital-services index, we compute a Tornqvist growth rate using time-varying investment shares. Let s_t^a denote the share of asset a in total investment, and let

$$g_t^a = \log K_t^a - \log K_{t-1}^a$$

denote the growth rate of the capital stock of asset a. The aggregate growth rate of capital services is

$$g_t^{\mathrm{KS}} = \sum_a \bar{s}_t^a \, g_t^a, \qquad \bar{s}_t^a = \frac{1}{2}(s_t^a + s_{t-1}^a).$$

The capital-services index is then formed recursively and normalized to 100 at the first usable quarter:

$$A_t^{\mathrm{CapServ}} = 100 \times \exp\left(\sum_{\tau=t_0+1}^{t} g_\tau^{\mathrm{KS}} \right).$$

A.3 Data-center construction indicator

The data-center infrastructure indicator is based on a dataset on U.S. data-center construction activity. Each observation contains a date and a measure of the level of construction activity. Dates are mapped to quarter-end dates, and within each quarter the available observations are aggregated by taking their mean. This produces a quarterly raw series,

$$X_t^{\mathrm{DC}}.$$

The indexed data-center series is then defined as

$$A_t^{\mathrm{DC}} = 100 \times \frac{X_t^{\mathrm{DC}}}{X_{t_0}^{\mathrm{DC}}}, \qquad t_0 = \min\{t : X_t^{\mathrm{DC}} \text{ is observed}\}.$$

The raw data-center series is retained because it is used directly when constructing investment measures that incorporate data-center activity. The indexed version is used as a standalone indicator and as an input into the standardized composite measures. Because the underlying data only become available in the later part of the sample, the data-center series is left missing before its first observed quarter rather than backfilled.

A.4 AI-related patent indicator

The AI-patent indicator is constructed from a patent-level dataset that contains a date variable and multiple AI-related flags. A patent is classified as AI-related if at least one of the AI flags is nonzero. Each patent date is mapped to a quarter-end date, and the number of AI-related patents in each quarter is counted to form the series

$$N_t^{\mathrm{Pat}}.$$

The patent indicator is then defined as an index normalized to 100 at the first quarter in which the AI-related patent count is observed:

$$A_t^{\mathrm{Pat}} = 100 \times \frac{N_t^{\mathrm{Pat}}}{N_{t_0}^{\mathrm{Pat}}}, \qquad t_0 = \min\{t : N_t^{\mathrm{Pat}} \text{ is observed}\}.$$

The construction does not artificially backfill the patent series prior to its first observed quarter.

A.5 Adoption indicator

The adoption or diffusion indicator is based on survey evidence on firm use of AI. Let s_τ^{AI} denote the AI-use share observed at survey date τ. These observations are mapped to quarter-end dates, and within each quarter the available values are averaged:

$$\bar{s}_t^{\mathrm{AI}} = \frac{1}{n_t} \sum_{\tau \in t} s_\tau^{\mathrm{AI}},$$

where n_t is the number of survey observations in quarter t.

The adoption indicator is then defined as

$$A_t^{\mathrm{Adopt}} = 100 \times \frac{\bar{s}_t^{\mathrm{AI}}}{\bar{s}_{t_0}^{\mathrm{AI}}}, \qquad t_0 = \min\{t : \bar{s}_t^{\mathrm{AI}} \text{ is observed}\}.$$

Because this series is only available late in the sample, it is used mainly as an auxiliary indicator rather than as a core component of the main composite measures.

A.6 Capability frontier indicator

The capability frontier indicator is derived from benchmark information on frontier AI models. For each date with available benchmark data, the models are ranked by overall performance, and the average score of the top-performing models is computed. This produces a daily frontier statistic F_τ.

The daily frontier series is mapped to quarter-end dates, and within each quarter the available values are averaged:

$$\bar{F}_t = \frac{1}{m_t} \sum_{\tau \in t} F_\tau,$$

where m_t is the number of benchmark observations in quarter t.

The quarterly capability indicator is then defined as

$$A_t^{\mathrm{Cap}} = 100 \times \frac{\bar{F}_t}{\bar{F}_{t_0}}, \qquad t_0 = \min\{t : \bar{F}_t \text{ is observed}\}.$$

As with the adoption series, the capability frontier indicator is available only in the later part of the sample and is used primarily to interpret the timing of changes in the broader AI activity indicators.

A.7 Investment measures with and without data centers

The construction distinguishes between investment measures that exclude data centers and investment measures that include them. The data-center-inclusive measures are formed by adding the *raw* quarterly data-center series to the corresponding investment flow and then normalizing the resulting sum to 100 at its first observed quarter:

$$I_t^{\text{Inv, WithDC}} = I_t^{\text{Inv, NoDC}} + X_t^{\mathrm{DC}},$$

$$A_t^{\text{Inv, WithDC}} = 100 \times \frac{I_t^{\text{Inv, WithDC}}}{I_{t_0}^{\text{Inv, WithDC}}}.$$

In addition, an extended-back version is constructed by setting the raw data-center series equal to zero before the first quarter in which data-center activity is observed:

$$\tilde{X}_t^{\text{DC}} = \begin{cases} 0, & \text{before the first observed data-center quarter,} \\ X_t^{\text{DC}}, & \text{thereafter.} \end{cases}$$

This yields

$$I_t^{\text{Inv, WithDC, Early}} = I_t^{\text{Inv, NoDC}} + \tilde{X}_t^{\text{DC}},$$

and the corresponding index

$$A_t^{\text{Inv, WithDC, Early}} = 100 \times \frac{I_t^{\text{Inv, WithDC, Early}}}{I_{t_0}^{\text{Inv, WithDC, Early}}}.$$

The principal composite measures use the standard data-center-inclusive investment indices rather than the extended-back versions. The extended-back variants are retained mainly for descriptive purposes.

A.8 Standardization and composite indices

All component and investment indices that enter the composite measures are first standardized over the sample. For a generic level index $A_{j,t}$, the standardized version is

$$Z_{j,t} = \frac{A_{j,t} - \bar{A}_j}{s(A_j)},$$

where $\bar{A}_j$ and $s(A_j)$ denote the sample mean and standard deviation over the quarters in which $A_{j,t}$ is observed.

The composite indices are then formed as simple arithmetic means of selected standardized components. The broad baseline composite combines a data-center-inclusive investment measure, the capital-services index, and the patent indicator:

$$\text{AI}_t^{\text{Baseline}} = \frac{1}{3}\left(Z_t^{\text{Inv, WithDC}} + Z_t^{\text{CapServ}} + Z_t^{\text{Pat}}\right).$$

The composite that excludes data centers is

$$\text{AI}_t^{\text{NoDC}} = \frac{1}{3}\left(Z_t^{\text{Inv, NoDC}} + Z_t^{\text{CapServ}} + Z_t^{\text{Pat}}\right).$$

The composite that excludes capital services is

$$\text{AI}_t^{\text{NoCapServ}} = \frac{1}{2}\left(Z_t^{\text{Inv, WithDC}} + Z_t^{\text{Pat}}\right).$$

The composite that excludes both capital services and data centers is

$$\text{AI}_t^{\text{NoCapServ, NoDC}} = \frac{1}{2}\left(Z_t^{\text{Inv, NoDC}} + Z_t^{\text{Pat}}\right).$$

Finally, the supply-side composite averages only capital services and patents:

$$\mathrm{AI}_t^{\mathrm{Supply}} = \frac{1}{2}\left(Z_t^{\mathrm{CapServ}} + Z_t^{\mathrm{Pat}}\right).$$

In addition to these standardized composites, corresponding level composites are also formed using the same arithmetic-average formulas applied directly to the level indices before standardization.

A.9 Series retained in the quarterly panel

The quarterly panel contains the national-accounts investment series used to construct the AI-related investment indicators, together with the additional variables used to build the broader AI activity measures and composites. The underlying investment series are:

- real private fixed investment in software (FRED mnemonic `B985RX1Q020SBEA`);
- real private fixed investment in computers and peripheral equipment (FRED mnemonic `B935RX1Q020SBEA`);
- real private fixed investment in communication equipment (FRED mnemonic `C275RX1A020NBEA`, observed at annual frequency and interpolated to quarterly frequency before use);
- real private fixed investment in research and development (FRED mnemonic `Y006RX1Q020SBEA`);
- real private fixed investment in nonresidential structures (FRED mnemonic `W003RX1Q020SBEA`);
- real private fixed investment in information processing equipment (FRED mnemonic `Y034RX1Q020SBEA`).

From these series, we construct the investment-based AI indicators used in the main analysis, both excluding and including data centers, as well as an extended-back version of the data-center-inclusive investment measure that sets data-center activity equal to zero before its first observed quarter.

In addition to the investment-based measures, the panel includes the following series:

- a capital-services index constructed from software, computers and peripheral equipment, communication equipment, research and development, and nonresidential structures;
- a data-center construction indicator based on the U.S. Census Bureau's *Value of Private Construction Put in Place* release, using the seasonally adjusted annual rate series for private data-center construction;
- an AI-related patent indicator based on two complementary sources: the USPTO's *Artificial Intelligence Patent Dataset (AIPD) 2023*, which identifies AI-related patent documents through 2023, and *PatentsView*, which provides patent-level and assignee-level information used in constructing the quarterly patent-based indicator;
- an adoption indicator based on the U.S. Census Bureau's *Business Trends and Outlook Survey (BTOS)*, using the biweekly share of firms reporting AI use and aggregating those observations to the quarterly frequency;

- a capability-frontier indicator based on the *LM Arena history* dataset, using the historical model-score data reported in the public `scores.json` file; for each date, the measure is constructed from the top-performing models and then aggregated to the quarterly frequency;

All of these level indicators are retained in the merged quarterly dataset, and standardized versions are then formed by subtracting the sample mean and dividing by the sample standard deviation. The standardized panel therefore includes standardized investment indicators, standardized versions of capital services, patents, data centers, adoption, capability, and the standalone information-processing-equipment series.

The composite indices used in the analysis are then constructed from these standardized series. The panel includes:

- a baseline composite that combines a data-center-inclusive investment measure, capital services, and patents;
- a composite excluding data centers, which combines a data-center-excluding investment measure, capital services, and patents;
- a composite excluding capital services, which combines a data-center-inclusive investment measure and patents;
- a composite excluding both capital services and data centers, which combines a data-center-excluding investment measure and patents;
- a supply-side composite that averages capital services and patents.

Corresponding level composites are also retained prior to standardization, using the same arithmetic-average formulas applied directly to the level indices. Taken together, the quarterly panel therefore contains the full set of investment-based indicators, the additional auxiliary AI-related variables, their standardized counterparts, and the composite indices used throughout the empirical analysis.

B Detailed definitions and implementation of the AI contribution exercises

This appendix documents the construction of the AI contribution exercises used in Section 3. It consolidates the definitions, transformations, controls, and alternative empirical designs in one place. The purpose is to make clear how the contribution series are constructed, what variables enter each specification, and how the resulting fitted AI component should be interpreted.

B.1 Outcome series and quarterly transformations

The macro outcomes are constructed from the following U.S. macroeconomic series:

- *Real Gross Domestic Product* (FRED mnemonic `GDPC1`; source: BEA via FRED), observed quarterly;
- *Gross Domestic Product: Implicit Price Deflator* (FRED mnemonic `GDPDEF`; source: BEA via FRED), observed quarterly;
- *Consumer Price Index for All Urban Consumers: All Items* (FRED mnemonic `CPIAUCSL`; source: BLS via FRED), observed monthly;
- *Personal Consumption Expenditures: Chain-type Price Index* (FRED mnemonic `PCEPI`; source: BEA via FRED), observed monthly.

The real-activity outcome is real GDP growth. The inflation outcomes are GDP-deflator inflation, CPI inflation, and PCE inflation. The baseline sample begins in 2007Q1 and extends through 2025Q3, with the effective estimation sample determined by maximal overlap for each outcome, AI measure, and empirical specification.

All outcomes are expressed as annualized quarter-on-quarter log changes. For real GDP,

$$\Delta y_t \equiv 400\big(\log Y_t - \log Y_{t-1}\big). \tag{B.1}$$

For each price index P_t,

$$\pi_t \equiv 400\big(\log P_t - \log P_{t-1}\big). \tag{B.2}$$

This transformation places real GDP growth and inflation in annualized percentage points.

Because CPI and the PCE price index are observed monthly, quarterly price levels are first constructed by averaging monthly observations within each quarter:

$$P_t^{CPI} \equiv \frac{1}{3}\sum_{m\in t} P_m^{CPI}, \qquad P_t^{PCE} \equiv \frac{1}{3}\sum_{m\in t} P_m^{PCE}. \tag{B.3}$$

Quarterly inflation is then computed from these quarterly-average price levels as $\pi_t = 400\Delta \log P_t$. This ensures that GDP-deflator inflation, CPI inflation, and PCE inflation are evaluated on a common quarterly calendar and expressed in comparable units.

The GDP deflator is the headline inflation concept because it is a domestic-production price index aligned with the GDP concept used for real activity. It measures the price of domestically produced final output and excludes imports. CPI and PCE are used primarily as robustness outcomes and for comparability with familiar policy benchmarks.

B.2 AI indicators, variants, and quarterly regressors

The AI measures are taken from the quarterly AI panel used elsewhere in the paper. The contribution exercise uses standardized quarterly AI indicators and standardized composite indices. The variants emphasized in the contribution analysis are constructed from the following component measures:

- an AI-oriented investment measure excluding data centers;
- an AI-oriented investment measure including data centers;
- an AI-related capital-services measure;
- a patent-based measure;
- a data-center measure;
- a capital-services-and-patents composite;
- a composite excluding data centers and capital services;
- a composite excluding data centers;
- a composite including data centers.

Let Z_t denote the standardized quarterly AI index associated with a given variant. The contribution exercises use the first difference of the standardized AI measure:

$$x_t \equiv \Delta Z_t = Z_t - Z_{t-1}. \tag{B.4}$$

This transformation captures quarter-to-quarter movements in AI intensity on a common, scale-free basis. For composite indices, Z_t is the standardized composite itself. For individual components, the raw level series is first normalized and standardized, and the contribution regressor is then the first difference of that standardized series.

For data-center-inclusive variants, observations prior to the first quarter in which the data-center series is available are treated as missing. Consequently, the effective sample for data-center-inclusive variants begins later than the sample for variants that exclude data centers.

B.3 Common contribution mapping

Across all contribution exercises, the central object is a quarterly fitted AI component. For a generic inflation outcome π_t and AI regressor x_t, the simplest projection is

$$\pi_t = \alpha + \beta x_t + u_t. \tag{B.5}$$

The associated AI contribution series is

$$\widehat{c}_t^{AI} \equiv \widehat{\beta} x_t. \tag{B.6}$$

This definition follows from the fitted value of the linear projection. The fitted inflation rate is

$$\widehat{\pi}_t = \widehat{\alpha} + \widehat{\beta} x_t. \tag{B.7}$$

The term $\widehat{\beta}x_t$ is the component of fitted inflation associated with the AI regressor. Equivalently, it is the change in fitted inflation associated with the observed value of x_t, relative to a counterfactual fitted value in which the AI movement is set to zero while the remaining terms in the specification are held fixed.

Because the dependent variable is measured in annualized percentage points, $\widehat{c}_t^{AI}$ is also measured in annualized percentage points. Since x_t is a change in standardized AI intensity, $\widehat{\beta}$ measures the change in annualized inflation associated with a one-unit quarterly increase in the standardized AI measure. The contribution series is therefore a reduced-form fitted component, not a structural decomposition. Omitted variables, simultaneity, and measurement error can affect $\widehat{\beta}$, and the contribution should be interpreted as a conditional association under the maintained specification.

B.4 Benchmark contemporaneous projection

The benchmark case estimates the contemporaneous projection

$$\pi_t = \alpha + \beta x_t + \delta' D_t + \varepsilon_t, \tag{B.8}$$

where $x_t = \Delta Z_t$ is the quarterly change in standardized AI intensity and D_t denotes included time dummies. The main dummy is a post-COVID surge dummy equal to one from 2020Q1 through 2022Q4 and zero otherwise.

The AI contribution is

$$\widehat{c}_t^{AI} = \widehat{\beta}x_t. \tag{B.9}$$

This case provides the most transparent same-quarter association between AI movements and inflation. The slope coefficient is fixed over the estimation window, so all time variation in the contribution series comes from observed quarter-to-quarter movements in the AI regressor.

B.5 Controls-augmented projection

The controls case augments the benchmark projection with a vector of quarterly controls:

$$\pi_t = \alpha + \beta x_t + \theta' \mathbf{X}_t + \delta' D_t + \varepsilon_t. \tag{B.10}$$

The contribution series is again

$$\widehat{c}_t^{AI} = \widehat{\beta}x_t. \tag{B.11}$$

The control vector $\mathbf{X}_t$ is designed to absorb major non-AI drivers of inflation. When a control is observed at daily or monthly frequency, it is first aggregated to quarterly frequency by within-quarter averaging. The baseline controls are:

- the Global Supply Chain Pressure Index (GSCPI), capturing supply bottlenecks;
- real transfer payments, constructed from *Personal current transfer receipts* (FRED mnemonic `B087RC1Q027SBEA`) deflated by the GDP deflator and transformed as an annualized q/q log change;
- an energy-price control based on *Crude Oil Prices: West Texas Intermediate (WTI), Cushing, Oklahoma* (FRED mnemonic `DCOILWTICO`), averaged to quarterly frequency and transformed as $400\Delta \log(\text{WTI}_t)$;

- Federal Reserve balance-sheet size, based on *Assets: Total Assets: Total Assets (Less Eliminations from Consolidation): Wednesday Level* (FRED mnemonic `WALCL`), averaged to quarterly frequency and transformed as an annualized q/q log change;
- a tariff proxy based on *Customs Duties* (FRED mnemonic `B235RC1Q027SBEA`), transformed as an annualized q/q log change.

The controls case asks whether the AI-attributed component remains positive and economically meaningful after conditioning on salient macroeconomic forces. Relative to the benchmark, it provides a more conservative contribution measure, though at the cost of greater specification dependence and, in some cases, a shorter effective sample.

B.6 Factor-Augmented Vector Autoregression (FAVAR)-lite exercise

The Factor-Augmented Vector Autoregression (FAVAR)-lite exercise provides an alternative way to summarize the information contained in the AI measurement block. The motivation is that the individual AI proxies used in the paper capture related but distinct dimensions of AI development and diffusion. Some measures are closer to investment, others to capital services, infrastructure, or innovation. Each series may therefore contain both a common AI component and idiosyncratic measurement noise. The FAVAR-lite exercise extracts the common movement across the relevant AI variables and uses that movement to construct an alternative AI contribution to inflation.

In a full FAVAR, a large set of macroeconomic and financial variables is summarized by a small number of latent factors, and those factors are then included in a vector autoregression together with the variables of interest. The factors and the observed variables evolve jointly, and impulse responses can be computed from the estimated system. The present exercise is called "lite" because it uses only the factor-extraction part of that logic. It does not estimate a full structural FAVAR, does not identify structural shocks, and does not compute dynamic impulse responses from a VAR system. Instead, it uses the first principal component of the AI block as a compact measure of common AI intensity and then relates quarterly movements in that factor to inflation.

For each AI variant, the relevant block of AI variables is first assembled in standardized units. Let

$$\mathcal{A}_t = (A_{1t}, A_{2t}, \ldots, A_{Nt})$$

denote the standardized AI block available for a given specification. The specific variables included in the block depend on the AI measure being considered. For example, the block may include AI-related investment, AI capital-services measures, patent-based innovation measures, and, in the data-center-inclusive specification, the data-center infrastructure component. Standardizing the series before extracting the factor ensures that the principal component is not driven mechanically by variables with larger units or variances.

The first principal component of the standardized AI block is then extracted:

$$F_t \equiv \mathrm{PC1}\left(\mathcal{A}_t\right). \tag{B.12}$$

This factor is the linear combination of the underlying standardized AI variables that explains the largest share of their common variation. It can therefore be interpreted as a summary index of broad AI intensity for that measurement block. Since the empirical contribution exercise is based on quarterly changes in AI intensity rather than levels, the factor is transformed into a quarterly movement:

$$x_t^F = \Delta F_t = F_t - F_{t-1}. \tag{B.13}$$

This is directly analogous to the treatment of composite AI indices expressed in standardized units, where the contribution exercise uses first differences rather than log growth rates.

The inflation projection is then estimated as

$$\pi_t = \alpha + \beta x_t^F + \delta' D_t + \varepsilon_t, \tag{B.14}$$

where π_t is the inflation measure used in the contribution calculation and D_t denotes the set of controls or dummy variables included in the corresponding projection. The coefficient β measures the conditional association between quarterly movements in the common AI factor and inflation. The fitted AI contribution from the FAVAR-lite exercise is

$$\widehat{c}_t^{AI,F} = \widehat{\beta} x_t^F. \tag{B.15}$$

Thus, the contribution is the part of inflation mechanically attributed to the movement in the extracted common AI factor, holding fixed the other controls included in the projection.

This construction should be interpreted as a reduced-form accounting exercise rather than as a structural causal decomposition. It asks whether the common component of the AI measurement block is associated with inflation in a way similar to the individual or composite AI indices used in the baseline contribution calculations. The advantage is that it reduces reliance on any single AI proxy and gives less weight to purely idiosyncratic movements in individual series. If several AI indicators move together, the first principal component captures that common movement and provides a smoother summary of AI intensity. This is useful when the underlying AI measures are noisy, measured with error, or capture different stages of the same diffusion process.

The tradeoff is that the factor is less directly interpretable than an individual observable series. A coefficient on an investment-based AI measure, for example, can be read as the inflation association of that particular investment proxy. By contrast, the coefficient on x_t^F reflects a weighted common movement across the entire AI block. The weights are determined statistically by the covariance structure of the block, not by a theoretical model of AI diffusion. For this reason, the FAVAR-lite results are best viewed as a robustness and summarization exercise. They help verify whether the inflation contribution attributed to AI is present in the common AI signal, rather than being driven by one noisy component of the measurement system.

B.7 Rolling or time-varying contribution exercise

The rolling exercise allows the estimated sensitivity of inflation to AI to change over time. For each quarter t, a benchmark-type projection is estimated over a rolling window ending at t:

$$\pi_\tau = \alpha_t + \beta_t x_\tau + \delta_t' D_\tau + \varepsilon_\tau, \qquad \tau \in \mathcal{W}_t, \tag{B.16}$$

where $\mathcal{W}_t$ denotes the rolling estimation window. The contribution is then

$$\widehat{c}_t^{AI,\text{roll}} = \widehat{\beta}_t x_t. \tag{B.17}$$

This exercise is intended to capture the possibility that the macroeconomic relevance of AI changes over time. For example, the inflation sensitivity to AI may increase as AI moves from an innovation phase to a deployment phase, as data-center investment expands, or as firms increasingly use AI in pricing, production, and logistics. The rolling contribution therefore differs from the benchmark contribution because both the AI regressor and the estimated coefficient can vary over time.

B.8 Local-projection contribution exercises

The local-projection exercises allow the relationship between AI movements and inflation to be estimated over different horizons. At horizon h, the generic local-projection specification is

$$\pi_{t+h} = \alpha_h + \beta_h x_t + \theta_h' \mathbf{X}_t + \delta_h' D_t + u_{t+h,h}. \tag{B.18}$$

The horizon-0 case is closely related to the benchmark contemporaneous projection:

$$\pi_t = \alpha_0 + \beta_0 x_t + \theta_0' \mathbf{X}_t + \delta_0' D_t + u_{t,0}. \tag{B.19}$$

The impact local-projection contribution is

$$\widehat{c}_t^{AI,LP0} = \widehat{\beta}_0 x_t. \tag{B.20}$$

A dynamic local-projection contribution can also be constructed by accumulating the estimated horizon responses over a finite horizon H:

$$\widehat{c}_t^{AI,LP} = \sum_{h=0}^{H} \widehat{\beta}_h x_{t-h}. \tag{B.21}$$

This dynamic contribution allows inflation in quarter t to reflect both current AI movements and the propagated effects of earlier AI movements. It is useful if AI adoption, infrastructure build-out, or capital deepening affects inflation gradually rather than immediately. Because this procedure combines multiple estimated horizon coefficients, it is more sensitive to horizon choice and carries more estimation uncertainty than the benchmark contemporaneous contribution.

B.9 Cumulative contribution and summary statistics

For each contribution series $\widehat{c}_t^{AI}$, the tables report the windowed mean, median, cumulative contribution, positive share, and confidence interval for the mean. The cumulative contribution is log-compounded as

$$100 \left[\exp\left(\sum_{t\in\mathcal{T}} \frac{\widehat{c}_t^{AI}}{400} \right) - 1 \right], \tag{B.22}$$

where $\mathcal{T}$ is the relevant estimation window. This converts a sequence of annualized quarterly contribution rates into a cumulative price-level contribution over the window.

The positive share is the fraction of quarters in which

$$\widehat{c}_t^{AI} > 0. \tag{B.23}$$

This statistic is useful because a positive average contribution could, in principle, be driven by a small number of extreme positive quarters. A high positive share indicates that the estimated AI contribution is positive in a broad fraction of the sample.

The confidence interval for the windowed mean contribution is computed as

$$\bar{c} \pm t_{1-\alpha/2,N-1} \frac{s_c}{\sqrt{N}}, \tag{B.24}$$

where $\bar{c}$ is the sample mean contribution, s_c is the sample standard deviation of the contribution series, and N is the number of non-missing contribution observations in the window.

B.10 Interpretation and limitations

The contribution exercises should be interpreted as reduced-form fitted-component calculations. The term $\widehat{\beta}x_t$ is called a contribution because it is the part of the fitted value of inflation associated with the AI regressor under the maintained specification. It answers the question: given the estimated projection and the observed quarterly movement in standardized AI intensity, how much of fitted inflation in quarter t is associated with AI?

This interpretation is narrower than a structural causal decomposition. The contribution series does not prove that AI caused the corresponding amount of inflation. Rather, it quantifies the inflation component that co-moves with AI under a particular empirical mapping. The comparison across benchmark, controls, FAVAR-lite, rolling, and local-projection designs is therefore important. If the sign, timing, and ranking of the contribution remain similar across these environments, the evidence is more consistent with a stable AI-inflation association rather than a specification-specific artifact.

C Robustness: Exposure-Based Sectoral Evidence

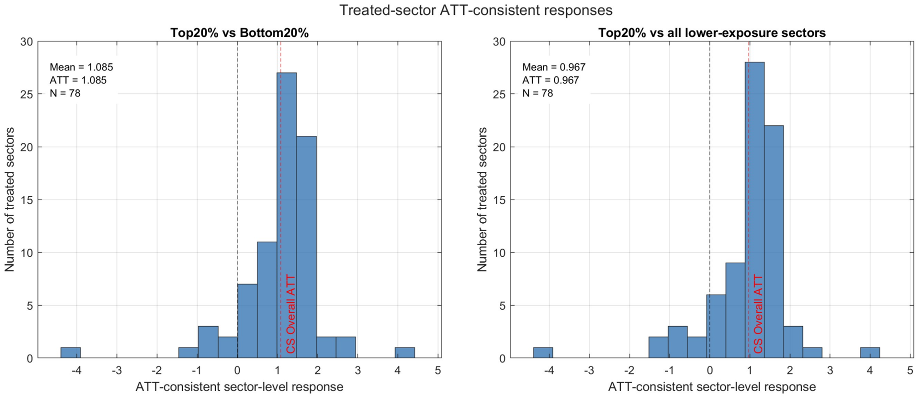


Figure C.1: Distribution of individual treated-sector effects. The figure reports the distribution of treated-sector differential effects used to diagnose the Callaway–Sant'Anna results. The distribution shows that the positive overall ATT is not driven by a small number of extreme positive observations. Instead, the treated-sector effects are broadly shifted to the right, with the vast majority of high-exposure treated sectors displaying positive differential post-event effects relative to their pre-event baseline.

	TWFE	IV-DiD	CS Overall ATT
Estimate	0.767***	0.832**	0.892*
	(0.212)	(0.400)	(0.514)
First-stage F	–	281.87	–
Sector FE	Yes	Yes	Residualised
Quarter FE	Yes	Yes	Residualised
Controls	Yes	Yes	Yes
Sector trends	Sector poly. (1)	Sector poly. (1)	Sector poly. (1)
Lagged inflation	Yes	Yes	Yes
Dummies	No	No	No
Observations	26,871	26,871	26,871
Sectors	382	382	382
Event quarter	2022Q4		
Treatment definition	Top 50% vs. Bottom 50% AI task exposure		
CS baseline window	$[-6, -1]$ quarters relative to event		

Table C.1: AI task exposure and sectoral PPI inflation: median split. This table reports estimates from three complementary estimators under a median-split treatment definition, in which sectors above the 50th percentile of the AI task exposure distribution are assigned to the treated group and sectors at or below the median serve as controls. TWFE is a two-way fixed-effects difference-in-differences regression. IV-DiD instruments the post-period treatment interaction with a Bartik shift-share variable constructed as the industry's AI task exposure score interacted with a cumulative aggregate AI index. CS Overall ATT is the Callaway–Sant'Anna (2021) average treatment effect averaged over all post-event quarters, computed on an outcome residualised with respect to the fixed effects, trends, and controls listed above. AI task exposure is constructed at the NAICS 6-digit industry level using 2019 OES/OEWS employment weights and O*NET Work Activities importance scores for four analytical and information-processing task elements. Standard errors in parentheses are clustered at the sector level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	TWFE	IV-DiD	CS Overall ATT
Panel A: NAICS 2-digit, SOC 2-digit			($N = 6$ sectors, 413 obs.)
Estimate	2.142***	1.882***	3.626***
	(0.194)	(0.517)	(1.049)
First-stage F	–	32.51	–
Panel B: NAICS 3-digit, SOC 3-digit			($N = 22$ sectors, 1,517 obs.)
Estimate	1.896**	1.453	3.418**
	(0.839)	(0.929)	(1.633)
First-stage F	–	282.37	–
Panel C: NAICS 4-digit, SOC 4-digit			($N = 68$ sectors, 4,706 obs.)
Estimate	1.064***	0.883**	1.584*
	(0.370)	(0.397)	(0.932)
First-stage F	–	1,148.39	–
Panel D: NAICS 5-digit, SOC 5-digit			($N = 125$ sectors, 8,339 obs.)
Estimate	0.906**	0.711*	1.020
	(0.367)	(0.418)	(0.783)
First-stage F	–	653.55	–
Panel E: NAICS 6-digit, SOC 2-digit (baseline)			($N = 165$ sectors, 11,573 obs.)
Estimate	0.718**	0.765**	1.085*
	(0.283)	(0.329)	(0.653)
First-stage F	–	427.24	–
Panel F: NAICS 6-digit, SOC 3-digit			($N = 162$ sectors, 11,353 obs.)
Estimate	0.743**	0.743**	1.133
	(0.311)	(0.339)	(0.740)
First-stage F	–	1,042.68	–
Panel G: NAICS 6-digit, SOC 4-digit			($N = 154$ sectors, 10,785 obs.)
Estimate	0.676**	0.697**	1.057
	(0.308)	(0.335)	(0.754)
First-stage F	–	1,057.93	–
Panel H: NAICS 6-digit, SOC 5-digit			($N = 153$ sectors, 10,748 obs.)
Estimate	0.703**	0.701*	0.293
	(0.341)	(0.394)	(0.844)
First-stage F	–	846.43	–
Panel I: NAICS 6-digit, SOC 6-digit			($N = 165$ sectors, 10,748 obs.)
Estimate	0.546**	0.193	0.847*
	(0.223)	(0.476)	(0.490)
First-stage F	–	1,606.42	–

Table C.2: Robustness to NAICS and SOC aggregation level. Each panel replicates the baseline specification of Table 5 at a different level of industry and task aggregation. Moving from Panel A to Panel D, both the NAICS and SOC digit levels increase from 2 to 5, yielding progressively finer industry classifications and larger estimation samples. Panels E–I hold the NAICS digit level fixed at 6 and vary the SOC digit level from 2 to 6, isolating the sensitivity of the results to the granularity of the intermediate classification used to match O*NET task scores to OES/OEWS employment weights before aggregating to the industry level. The AI task exposure measure is constructed from O*NET Work Activities importance scores for four analytical and information-processing task elements, aggregated to the industry level using 2019 OES/OEWS employment weights at the indicated NAICS and SOC digit levels. Standard errors in parentheses are clustered at the sector level. Sector and quarter fixed effects are included, with the outcome residualized accordingly for the Callaway–Sant'Anna estimates. All panels include controls for upstream PPI inflation, energy, demand, digital/ICT exposure, and lagged inflation, as well as sector-specific linear trends. The event quarter is 2022Q4. Treatment is defined as sectors in the top 20% versus the bottom 20% of AI task exposure. For the Callaway–Sant'Anna estimates, the baseline window is $[-6, -1]$ quarters relative to the event. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

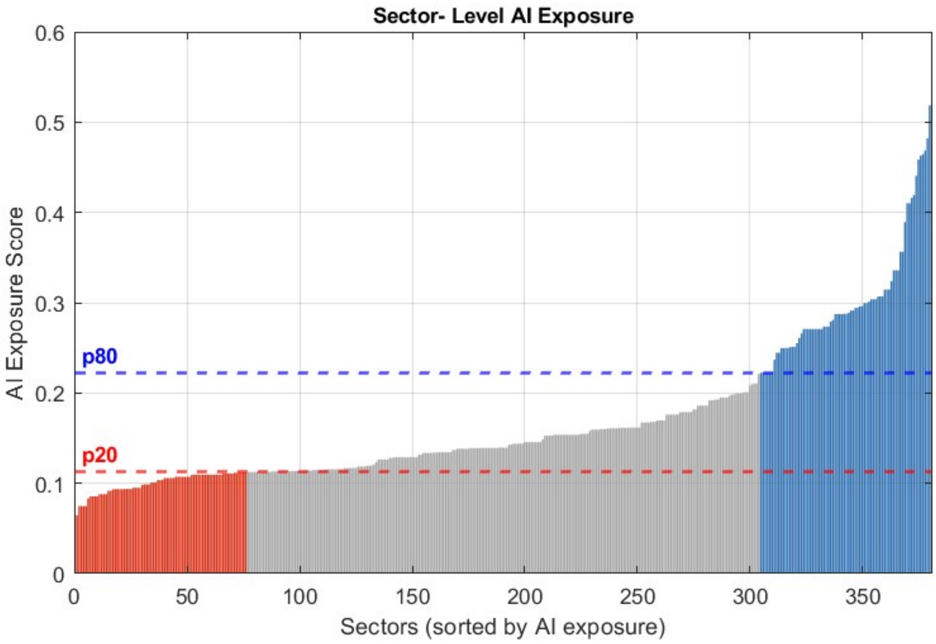


Figure C.2: Distribution of sectoral AI task exposure scores: 2007 occupational employment shares. The figure displays the task-based AI exposure score AIExp_i for all $N = 380$ sectors in the estimation sample, sorted in ascending order, constructed using 2007 OES/OEWS occupational employment shares. Red bars indicate sectors in the bottom 20% of the exposure distribution (control group); blue bars indicate sectors in the top 20% (treated group); grey bars indicate the excluded middle 60%. The red and blue dashed horizontal lines mark the 20th- and 80th-percentile thresholds, respectively. Scores are constructed from O*NET Work Activities importance scores for four analytical and information-processing task elements, aggregated to the 6-digit NAICS level using 2007 OES/OEWS occupational employment shares. Compare with Figure 4 for the 2019-based distribution.

	TWFE	IV-DiD	CS Overall ATT
Estimate	1.538***	1.411***	2.181***
	(0.299)	(0.382)	(0.687)
First-stage F	–	278.78	–
Sector FE	Yes	Yes	Residualised
Quarter FE	Yes	Yes	Residualised
Controls	Yes	Yes	Yes
Sector trends	Sector poly. (1)	Sector poly. (1)	Sector poly. (1)
Lagged inflation	Yes	Yes	Yes
Dummies	No	No	No
Observations	10,729	10,729	10,729
Sectors	153	153	153
Event quarter	2022Q4		
Treatment definition	Top 20% vs. Bottom 20% AI task exposure		
CS baseline window	$[-6, -1]$ quarters relative to event		

Table C.3: AI task exposure and sectoral PPI inflation: 2007 occupational employment shares, symmetric treatment definition. This table replicates the baseline specification of Table 5 with the AI task exposure measure constructed using 2007 OES/OEWS occupational employment shares in place of the 2019 shares used in the baseline. All other aspects of the specification are identical to Table 5. Standard errors in parentheses are clustered at the sector level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	TWFE	IV-DiD	CS Overall ATT
Estimate	1.113***	1.015***	1.906***
	(0.175)	(0.258)	(0.460)
First-stage F	–	390.30	–
Sector FE	Yes	Yes	Residualised
Quarter FE	Yes	Yes	Residualised
Controls	Yes	Yes	Yes
Sector trends	Sector poly. (1)	Sector poly. (1)	Sector poly. (1)
Lagged inflation	Yes	Yes	Yes
Dummies	No	No	No
Observations	26,871	26,871	26,871
Sectors	382	382	382
Event quarter	2022Q4		
Treatment definition	Top 20% vs. Bottom 80% AI task exposure		
CS baseline window	$[-6, -1]$ quarters relative to event		

Table C.4: AI task exposure and sectoral PPI inflation: 2007 occupational employment shares, asymmetric treatment definition. This table replicates the asymmetric specification of Table 6 with the AI task exposure measure constructed using 2007 OES/OEWS occupational employment shares in place of the 2019 shares used in the baseline. All other aspects of the specification are identical to Table 6. Standard errors in parentheses are clustered at the sector level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

D Measuring the AI-input cost share α

This appendix documents the construction of the empirical counterparts to the AI-input cost share α used to calibrate the bottleneck term $\alpha \hat{p}_{M,t}$ in the marginal-cost decomposition. We report (i) a baseline annual series from BEA–BLS industry production accounts (2007–2023) and (ii) a benchmark-year cross-section from benchmark-year BEA Use tables (2007, 2012, 2017). For transparency and replicability, we describe the precise inputs, formulas, sample restrictions, and the mapping from raw components to the "ITSE" proxy for the AI-intensive input M.

D.1 KLEMS / BEA–BLS industry production accounts (annual, 2007–2023)

Data source and sample. We use the BEA–BLS industry-level production account workbook, `BEA-BLS-industry-level-production-account-1997-2023.xlsx`, and retain annual observations for 2007–2023. The unit of observation is an *industry–year*. Industry identifiers are taken from the workbook's production-code field (stored as `IndustryKey`) and are optionally mapped to NAICS codes and industry names using the sheet `NAICS codes`.

Component series (sheet names). We read the following sheets (each contains industry rows with year-wide columns):

- `Service Compensation` (services input compensation)
- `Energy Compensation` (energy input compensation)
- `Materials Compensation` (materials input compensation)
- `Labor_Col Compensation` (college labor compensation)
- `Labor_NoCol Compensation` (non-college labor compensation)
- `Capital_IT Compensation` (IT capital compensation)
- `Capital_Software Compensation` (software capital compensation)
- `Capital_R&D Compensation` (R&D capital compensation)
- `Capital_Art Compensation` (art capital compensation)
- `Capital_Other Compensation` (other capital compensation)

Panel assembly and cleaning. Each component sheet is converted from wide (industry-by-year) to long (industry–year) format. Components are merged by `IndustryKey` and `Year` using full outer joins to preserve coverage across sheets; missing or non-finite component values are set to zero for summation when constructing denominators and bundles.[4] The resulting panel contains industry–year observations for the IT/ITS/ITSE bundles over 2007–2023 (and fewer for service-only, reflecting sheet-specific availability).

Denominators. We construct two cost denominators:

$$\text{VarCost}_t \equiv \text{Services}_t + \text{Energy}_t + \text{Materials}_t + \text{LaborCol}_t + \text{LaborNoCol}_t, \tag{D.1}$$

$$\text{TotalCost}_t \equiv \text{VarCost}_t + \text{K\_IT}_t + \text{K\_Software}_t + \text{K\_R\&D}_t + \text{K\_Art}_t + \text{K\_Other}_t. \tag{D.2}$$

[4] We impose a small minimum denominator threshold, $\texttt{minDenom} = 10^{-12}$, to avoid division by zero. Observations with non-positive or non-finite denominators are dropped for the relevant share.

Numerator bundles (ITSE mapping of M). We define four nested bundles designed to proxy for AI complements; the ITSE bundle is our preferred mapping:

$$M_t^S \equiv \text{Services}_t, \tag{D.3}$$
$$M_t^{IT} \equiv \text{K\_IT}_t + \text{K\_Software}_t, \tag{D.4}$$
$$M_t^{ITS} \equiv \text{Services}_t + M_t^{IT}, \tag{D.5}$$
$$M_t^{ITSE} \equiv M_t^{ITS} + \text{Energy}_t. \tag{D.6}$$

Shares. For each industry–year, we compute

$$\alpha^{\text{KLEMS}}(\text{ITSE, Total}) \equiv \frac{M_t^{ITSE}}{\text{TotalCost}_t}, \qquad \alpha^{\text{KLEMS}}(\text{ITSE, Var}) \equiv \frac{M_t^{ITSE}}{\text{VarCost}_t}, \tag{D.7}$$

and analogously for the S, IT, and ITS bundles. Shares that are negative, non-finite, or based on non-positive denominators are set to missing. We report pooled summary statistics across industry–year observations.

Pooled moments (2007–2023). For the ITSE mapping, the pooled distribution implies economically meaningful shares under both the total-cost and variable-cost denominators, with substantial cross-industry dispersion. These pooled summaries are computed over the 2007–2023 sample.

D.2 Benchmark-year BEA Use tables (2007, 2012, 2017)

Data source and sample. We use BEA Supply–Use detail (Use tables) in `Use_SUT_Detail.xlsx` for benchmark years 2007, 2012, and 2017. For each benchmark year, the Use table reports detailed intermediate purchases by commodity (rows) and by industry (columns). The unit of observation is an *industry–year* at the benchmark frequency.

Parsing to long format. For each benchmark sheet, we identify the commodity block (stopping before value-added and other summary lines) and reshape the commodity-by-industry matrix into a long table of intermediate purchases with fields:

{`Year`, `IndustryKey`, `IndustryName`, `CommodityCode`, `CommodityName`, `Value`}.

We retain only rows with finite values and non-empty identifiers.

AI-related commodity identification (ITSE proxy). We classify commodities as AI-related using keyword matching applied to `CommodityName`. We implement two sets:

- **Narrow** (services/data/compute): `data processing`, `hosting`, `cloud`, `internet`, `telecommunications`, `computer systems design`, `software`, `information services`, `data service`, `web`, `streaming`.
- **Broad** (narrow plus energy and hardware complements): narrow keywords plus `electric power`, `utilities`, `natural gas`, `petroleum`, `energy`, `computer and electronic products`, `semiconductor`, `electronic component`.

The broad definition is intended to better align with an ITSE-style bundle by incorporating energy and key hardware categories, albeit at a coarse keyword-based level.

Shares. For each industry–year in the benchmark Use tables, we compute:

$$\text{TotalIntermediate}_t \equiv \sum_c \text{IntermediatePurchases}_{c,t}, \tag{D.8}$$

$$\text{Minputs}_t \equiv \sum_{c \in \mathcal{C}_M} \text{IntermediatePurchases}_{c,t}, \tag{D.9}$$

$$\alpha^{\text{IO}} \equiv \frac{\text{Minputs}_t}{\text{TotalIntermediate}_t}, \tag{D.10}$$

where $\mathcal{C}_M$ is either the narrow or broad keyword-defined commodity set. Observations with non-positive or non-finite TotalIntermediate are excluded for the share computation; missing numerators are set to zero prior to division.

Pooled moments (2007, 2012, 2017). Pooling all available industry–year observations across benchmark years yields positive and economically meaningful shares for both the broad and narrow definitions, with the broad definition intended as the closer analogue to the ITSE mapping.

D.3 Reconciliation and interpretation

The KLEMS and IO approaches are complementary and not mechanically identical. The production-account (KLEMS-style) measure is compensation-based and naturally expressed as a share of total or variable cost, whereas the benchmark-year BEA Use-table measure is purchase-based and expressed as a share of total intermediate purchases. Despite these conceptual differences, the ITSE-motivated mappings produce overlapping empirical ranges and jointly support a moderate AI-input cost share. In the main text, we adopt the annual BEA–BLS production-account series as the primary calibration target due to its annual coverage over 2007–2023 and its close alignment with the cost-based objects in the model, while using benchmark-year IO estimates as an external validation check.

E Derivation of the AI-Augmented Real Marginal Cost

This appendix derives the reduced-form expression for (log) real marginal cost,

$$mc_t = \lambda x_t + \big(-\phi_A + \alpha\phi_{pM} + \phi_M\big) z_t, \tag{E.1}$$

starting from firms' cost minimization, the AI-induced productivity process, the AI-input price process, and the labor-market wedge induced by monopsony power.

E.1 Production technology and cost minimization

Each intermediate firm $i \in [0, 1]$ produces output using labor $L_t(i)$ and the AI-intensive input $M_t(i)$:

$$Y_t(i) = A_t\, L_t(i)^{1-\alpha} M_t(i)^{\alpha}, \qquad \alpha \in (0, 1). \tag{E.2}$$

Let $W_t(i)$ denote the nominal wage paid by firm i and $P_{M,t}$ the nominal price of the AI input. The firm minimizes nominal cost subject to Equation (E.2):

$$\min_{L_t(i),\, M_t(i)} \; W_t(i)L_t(i) + P_{M,t}M_t(i) \quad \text{s.t.} \quad Y_t(i) = A_t\, L_t(i)^{1-\alpha} M_t(i)^{\alpha}. \tag{E.3}$$

The Lagrangian is

$$\mathcal{L} = W_t(i)L_t(i) + P_{M,t}M_t(i) + \Lambda_t(i)\left[Y_t(i) - A_t\, L_t(i)^{1-\alpha} M_t(i)^{\alpha}\right]. \tag{E.4}$$

First-order conditions imply:

$$\frac{\partial \mathcal{L}}{\partial L_t(i)} = 0 \Rightarrow W_t(i) = \Lambda_t(i)\,(1-\alpha)\,A_t\, L_t(i)^{-\alpha} M_t(i)^{\alpha}, \tag{E.5}$$

$$\frac{\partial \mathcal{L}}{\partial M_t(i)} = 0 \Rightarrow P_{M,t} = \Lambda_t(i)\,\alpha\, A_t\, L_t(i)^{1-\alpha} M_t(i)^{\alpha-1}. \tag{E.6}$$

Dividing Equation (E.6) by Equation (E.5) yields the optimal input ratio:

$$\frac{M_t(i)}{L_t(i)} = \frac{\alpha}{1-\alpha}\, \frac{W_t(i)}{P_{M,t}}. \tag{E.7}$$

Under constant returns to scale, nominal marginal cost equals the Lagrange multiplier:

$$MC_t(i) = \Lambda_t(i). \tag{E.8}$$

Using Equation (E.5) and $Y_t(i) = A_t L_t(i)^{1-\alpha} M_t(i)^{\alpha}$, we obtain the standard CRS identity

$$MC_t(i) = \frac{1}{A_t} \left(\frac{W_t(i)}{1-\alpha}\right)^{1-\alpha} \left(\frac{P_{M,t}}{\alpha}\right)^{\alpha}. \tag{E.9}$$

Taking logs and defining lower-case variables as logs, $\log(\cdot)$, gives

$$\log MC_t(i) = (1-\alpha)\log W_t(i) + \alpha \log P_{M,t} - \log A_t + \text{const}. \tag{E.10}$$

E.2 From nominal to real marginal cost

Define real marginal cost as marginal cost divided by the aggregate price level:

$$mc_t(i) \equiv \log\left(\frac{MC_t(i)}{P_t}\right) = \log MC_t(i) - \log P_t. \tag{E.11}$$

Using Equation (E.10) and collecting terms,

$$mc_t(i) = (1-\alpha)\big(w_t(i) - p_t\big) + \alpha\big(p_{M,t} - p_t\big) - a_t + \text{const.}, \tag{E.12}$$

where $a_t \equiv \log A_t$, $w_t(i) \equiv \log W_t(i)$, $p_{M,t} \equiv \log P_{M,t}$, and $p_t \equiv \log P_t$.

Let $\hat{p}_{M,t}$ denote the *real* price of the AI input (log deviation from steady state),

$$\hat{p}_{M,t} \equiv \big(p_{M,t} - p_t\big) - \overline{\big(p_M - p\big)}. \tag{E.13}$$

Likewise, define the (log) real wage $\hat{w}_t(i) \equiv (w_t(i) - p_t) - \overline{(w-p)}$. Log-linearizing Equation (E.12) around a symmetric steady state and dropping constants yields

$$\widehat{mc}_t(i) = (1-\alpha)\hat{w}_t(i) + \alpha\hat{p}_{M,t} - a_t. \tag{E.14}$$

E.3 Introducing the output-gap component

In the baseline NK environment, log deviations in marginal cost co-move with the output gap x_t through the endogenous real wage (and, more generally, labor-market tightness and household marginal utility). Rather than re-deriving the full labor-supply block, we summarize these general-equilibrium forces with the standard reduced-form mapping:

$$(1-\alpha)\hat{w}_t = \lambda x_t, \qquad \lambda > 0, \tag{E.15}$$

where λ is the semi-elasticity of real marginal cost with respect to the output gap implied by preferences and labor-market structure.[5]

Aggregating across firms in a symmetric equilibrium, Equation (E.14) becomes

$$mc_t = \lambda x_t + \alpha\hat{p}_{M,t} - a_t, \tag{E.16}$$

where we have dropped hats on mc_t for notational convenience.

E.4 Allowing for an AI-induced monopsony wedge

With monopsony power, the firm faces an upward-sloping labor supply, so the marginal expenditure on labor exceeds the observed wage. Let the effective wage be

$$W_t^{\text{eff}}(i) = W_t(i)\left(1 + \frac{1}{\varphi_t}\right), \tag{E.17}$$

[5] In a textbook NK model with separable preferences and flexible wages, λ is proportional to $(\sigma+\varphi)$, where σ is the intertemporal elasticity parameter (inverse IES) and φ is the inverse Frisch elasticity. In richer labor-market environments, λ is a composite of wage-setting and labor-supply elasticities.

where φ_t is the wage elasticity of labor supply to the firm. We assume that higher AI intensity strengthens monopsony power by reducing the effective elasticity of labor supply:

$$\varphi_t = \bar{\varphi} - \eta_M z_t, \qquad \eta_M > 0. \tag{E.18}$$

A first-order log-linear approximation around $\bar{\varphi}$ implies that the log effective wage satisfies

$$\widehat{w}_t^{\text{eff}} \approx \widehat{w}_t + \phi_w z_t, \qquad \phi_w \equiv \frac{\eta_M}{\bar{\varphi}(\bar{\varphi}+1)} > 0. \tag{E.19}$$

Since the real wage enters (log) real marginal cost with coefficient $(1-\alpha)$ under the Cobb–Douglas production function, the corresponding monopsony contribution to marginal cost is scaled by the labor share. Define the reduced-form monopsony wedge in marginal cost as

$$\phi_M \;\equiv\; (1-\alpha)\phi_w \;=\; (1-\alpha)\frac{\eta_M}{\bar{\varphi}(\bar{\varphi}+1)} \;>\; 0. \tag{E.20}$$

Replacing $\hat{w}_t$ by $\hat{w}_t^{\text{eff}}$ in the cost block is therefore equivalent to adding $\phi_M z_t$ to marginal cost. Hence, Equation (E.16) becomes

$$mc_t = \lambda x_t + \alpha \hat{p}_{M,t} - a_t + \phi_M z_t. \tag{E.21}$$

E.5 AI processes for productivity and AI-input prices

AI affects productivity and the relative price of the AI input according to

$$a_t = \phi_A z_t, \qquad \phi_A > 0, \tag{E.22}$$

$$\hat{p}_{M,t} = \phi_{pM} z_t, \qquad \phi_{pM} > 0. \tag{E.23}$$

Finally, define α as the steady-state cost share of the AI input in marginal cost. Under Cobb–Douglas, $\alpha = \alpha$, but we keep $\alpha \in (0,1)$ as a general share parameter to allow for wedges between technological and expenditure shares (e.g. due to taxes, adjustment costs, or measurement differences). Replacing α by α in Equation (E.21) gives

$$mc_t = \lambda x_t + \alpha \hat{p}_{M,t} - a_t + \phi_M z_t. \tag{E.24}$$

Substituting equations (E.22) and (E.23) into Equation (E.24) yields

$$\begin{aligned} mc_t &= \lambda x_t + \alpha \phi_{pM} z_t - \phi_A z_t + \phi_M z_t \\ &= \lambda x_t + \big(-\phi_A + \alpha \phi_{pM} + \phi_M\big) z_t, \end{aligned} \tag{E.25}$$

which is the reduced-form expression reported in Equation (E.1).

F Solution and AI Shock Dynamics

This appendix derives the closed-form equilibrium response of the output gap and inflation to a persistent AI shock in the one-sector New Keynesian framework developed in Section 6. The derivations use the *Method of Undetermined Coefficients* and make explicit how the productivity, input-cost, market-power, and demand channels jointly determine the sign and magnitude of inflation responses.

F.1 Baseline system

Recall the IS curve,

$$x_t = \mathbb{E}_t x_{t+1} - \frac{1}{\sigma}\left(i_t - \mathbb{E}_t \pi_{t+1} - r_t^n\right) + \lambda_d d_t, \qquad d_t = \phi_D z_t, \tag{F.1}$$

the New Keynesian Phillips Curve,

$$\pi_t = \beta \mathbb{E}_t \pi_{t+1} + \kappa \lambda x_t + \kappa \Delta z_t, \qquad \Delta \equiv -\phi_A + \alpha \phi_{pM} + \phi_M + \phi_\mu, \tag{F.2}$$

and the Taylor rule,

$$i_t = r_t^n + \phi_\pi \pi_t + \phi_x x_t. \tag{F.3}$$

Artificial intelligence follows an AR(1) process,

$$z_t = \rho_z z_{t-1} + \varepsilon_t, \qquad |\rho_z| < 1, \tag{F.4}$$

so that $\mathbb{E}_t z_{t+1} = \rho_z z_t$.

F.2 Method of Undetermined Coefficients

To solve for the equilibrium response to z_t, conjecture linear laws of motion

$$x_t = A_x z_t, \qquad \pi_t = A_\pi z_t, \tag{F.5}$$

with coefficients (A_x, A_π) to be determined. Under Equation (F.4), the implied expectations are

$$\mathbb{E}_t x_{t+1} = A_x \mathbb{E}_t z_{t+1} = A_x \rho_z z_t, \qquad \mathbb{E}_t \pi_{t+1} = A_\pi \rho_z z_t. \tag{F.6}$$

Substitute the Taylor rule (F.3) into the IS curve (F.1). The natural rate r_t^n cancels out:

$$i_t - \mathbb{E}_t \pi_{t+1} - r_t^n = \phi_\pi \pi_t + \phi_x x_t - \mathbb{E}_t \pi_{t+1}.$$

Using (F.5)–(F.6) and dividing by z_t yields

$$\left(1 - \rho_z + \frac{\phi_x}{\sigma}\right) A_x + \frac{\phi_\pi - \rho_z}{\sigma} A_\pi = \lambda_d \phi_D. \tag{F.7}$$

Next, substitute (F.5)–(F.6) into the NKPC (F.2) and divide by z_t:

$$-\kappa \lambda A_x + (1 - \beta \rho_z) A_\pi = \kappa \Delta. \tag{F.8}$$

Equations (F.7)–(F.8) form a 2×2 linear system in (A_x, A_π). Define its determinant as

$$\Omega \equiv \left(1 - \rho_z + \frac{\phi_x}{\sigma}\right)(1 - \beta\rho_z) + \frac{\kappa\lambda}{\sigma}(\phi_\pi - \rho_z), \tag{F.9}$$

which is positive under standard stability and policy conditions.

F.3 Closed-form coefficients

Solving (F.7)–(F.8) by Cramer's rule yields

$$A_x = \frac{(1 - \beta\rho_z)\lambda_d\phi_D - \frac{\phi_\pi - \rho_z}{\sigma}\kappa\Delta}{\Omega}, \tag{F.10}$$

and

$$A_\pi = \frac{\kappa\lambda\,\lambda_d\phi_D + \kappa\Delta\left(1 - \rho_z + \frac{\phi_x}{\sigma}\right)}{\Omega}. \tag{F.11}$$

Therefore, the equilibrium responses are

$$x_t = A_x z_t, \qquad \pi_t = A_\pi z_t, \tag{F.12}$$

with (A_x, A_π) given by (F.10)–(F.11). Because the natural rate term cancels from the IS curve once the Taylor rule is substituted, r_t^n does not appear in the closed-form elasticities; its role is instead to pin down the level of the policy rate in (F.3).

F.4 Channel decomposition and interpretation

Using (F.11), the inflation response can be written as

$$\pi_t = \underbrace{\frac{\kappa\lambda}{\Omega}\lambda_d\phi_D}_{\text{Demand}}\, z_t + \underbrace{\frac{\kappa\,(1 - \rho_z + \phi_x/\sigma)}{\Omega}}_{\text{Policy \& persistence attenuation}} \underbrace{\Delta}_{\text{Net supply/market-power effect}} z_t, \tag{F.13}$$

where

$$\Delta = -\phi_A + \alpha\phi_{pM} + \phi_M + \phi_\mu$$

collects the four supply-side and market-power channels emphasized in Section 6: productivity (disinflationary), AI-input bottlenecks (inflationary), monopsony power (inflationary), and endogenous markups (inflationary). Equation (F.13) makes clear that AI is not generically a "pure productivity shock": even holding demand fixed, inflation can rise if the net cost and market-power forces dominate the productivity gain (i.e., if $\Delta > 0$). Conversely, when $\Delta < 0$, AI tends to be disinflationary through the supply block, and the overall sign depends on whether the positive demand component dominates.

Similarly, the output-gap response implied by (F.10) is

$$x_t = \left[\underbrace{\frac{(1 - \beta\rho_z)\lambda_d\phi_D}{\Omega}}_{\text{Demand}} + \underbrace{\frac{\kappa(\phi_\pi - \rho_z)}{\sigma\,\Omega}}_{\text{Policy \& persistence attenuation}} \underbrace{(-\Delta)}_{\text{Net supply/market-power effect}}\right] z_t, \tag{F.14}$$

highlighting that the demand channel pushes activity up, while the same net supply/market-power block that raises inflation when $\Delta > 0$ tends to depress activity (through tighter pricing conditions and policy feedback).

F.5 Special case: non-persistent AI shocks ($\rho_z = 0$)

When $\rho_z = 0$, the determinant simplifies to

$$\Omega\Big|_{\rho_z=0} = \left(1 + \frac{\phi_x}{\sigma}\right) + \frac{\kappa\lambda}{\sigma}\phi_\pi, \tag{F.15}$$

and the elasticities reduce to

$$A_x\Big|_{\rho_z=0} = \frac{\lambda_d \phi_D - \frac{\phi_\pi}{\sigma}\kappa\Delta}{\Omega\Big|_{\rho_z=0}}, \qquad A_\pi\Big|_{\rho_z=0} = \frac{\kappa\lambda\,\lambda_d\phi_D + \kappa\Delta\left(1 + \frac{\phi_x}{\sigma}\right)}{\Omega\Big|_{\rho_z=0}}. \tag{F.16}$$

This special case corresponds to purely impact effects, while (F.10)–(F.11) show how persistence reshapes the propagation and attenuates impact responses through ρ_z and the policy feedback embedded in Ω.

G Data and Measurement Details for the Supply-Side Decomposition

This appendix describes the empirical construction of the variables used in the decomposition of the AI supply-side term reported in Table 8. The goal is to map each theoretical component in

$$\Delta = -\phi_A + \alpha\phi_{p_M} + \phi_M + \phi_\mu \tag{G.1}$$

to a transparent empirical counterpart. The same AI-intensity variable and the same conditioning information set are used across all channel regressions, so that the resulting components are directly comparable.

G.1 Common AI-intensity variable and estimation approach

Let z_t denote the quarterly AI-intensity measure. The baseline channel estimates use the composite AI index that excludes data centers and combines AI-related investment, AI capital services, and AI patents. The series is standardized within the estimation sample, so each coefficient can be interpreted as the response of the relevant channel proxy to a one-standard-deviation increase in AI intensity.

For each channel variable Y_t^j, the empirical specification is

$$Y_t^j = a_j + b_j z_t + \Gamma_j' X_t + u_t^j, \tag{G.2}$$

where X_t includes the common macroeconomic controls used in the channel-estimation exercise. These controls include a global supply-chain pressure measure, real transfer payments deflated by the GDP deflator, an energy-price measure, Federal Reserve total assets, a tariff proxy, and the post-2020 inflation-surge dummy. Standard errors are computed using Newey–West corrections. The coefficient b_j is then mapped into the corresponding channel parameter: ϕ_A, ϕ_{p_M}, ϕ_M, or ϕ_μ.

G.2 Productivity channel

The productivity channel captures the effect of AI on productive efficiency. Since higher productivity lowers marginal cost, the contribution of this channel enters the decomposition as $-\phi_A$.

Two productivity measures are used. The first is labor productivity in the nonfarm business sector:

$$g_t^A = 100\Delta \log(OPHNFB_t), \tag{G.3}$$

where $OPHNFB_t$ is the nonfarm business output-per-hour index. The associated regression is

$$100\Delta \log(OPHNFB_t) = a_A + \phi_A z_t + \Gamma_A' X_t + u_t^A. \tag{G.4}$$

This produces the labor-productivity estimate reported in the table.

The second productivity measure is a residual TFP measure constructed from the multi-input marginal-cost accounting exercise. The starting point is a KLEMS-style input-price bundle that combines input-price growth for capital, labor, energy, materials, and services using input-cost shares:

$$\sum_{j\in\{K,L,E,M,S\}} s_t^j \Delta \log P_t^j, \tag{G.5}$$

where s_t^j is the cost share of input j and P_t^j is the corresponding input-price proxy. The residual productivity term is the productivity correction needed to convert this input-price bundle into a marginal-cost measure:

$$\Delta \log MC_t = \sum_{j\in\{K,L,E,M,S\}} s_t^j \Delta \log P_t^j - \Delta \log A_t. \tag{G.6}$$

The series $\Delta \log A_t$ obtained from this construction is used as the residual TFP measure. In the empirical implementation it is expressed as

$$g_t^{TFP} = 100\Delta \log A_t. \tag{G.7}$$

The corresponding productivity-channel coefficient is estimated from

$$g_t^{TFP} = a_A + \phi_A z_t + \Gamma_A' X_t + u_t^A. \tag{G.8}$$

G.3 AI-input-price channel

The AI-input-price channel captures the possibility that AI deployment raises the cost of complementary inputs used in AI-intensive production. These inputs include computing equipment, software, data-center infrastructure, energy, and related intermediate services. Let p_t^M denote the AI-input-price proxy. The channel coefficient is estimated from

$$p_t^M = a_M + \phi_{p_M} z_t + \Gamma_M' X_t + u_t^M. \tag{G.9}$$

The estimated response ϕ_{p_M} is then multiplied by the marginal-cost share of AI-related inputs, denoted α. Thus, the contribution of this channel is

$$\alpha\phi_{p_M}. \tag{G.10}$$

In the baseline decomposition, $\alpha = 0.10$. This gives an AI-input-price contribution of approximately 0.011.

G.4 Labor-market power and wage-markdown channel

The labor-market channel measures the extent to which AI is associated with a wedge between labor productivity and worker compensation. The empirical proxy is based on the ratio of labor productivity to real hourly compensation:

$$M_t = \log(OPHNFB_t) - \log(COMPRNFB_t), \tag{G.11}$$

where $OPHNFB_t$ is nonfarm business output per hour and $COMPRNFB_t$ is real hourly compensation in the nonfarm business sector. This series rises when output per hour increases relative to real compensation per hour, which is consistent with a larger wage-markdown or labor-market-power wedge.

The channel coefficient is estimated from

$$M_t = a_M + \phi_M z_t + \Gamma_M' X_t + u_t^M. \tag{G.12}$$

The resulting estimate enters the decomposition positively because a larger labor-market wedge raises the marginal-cost-relevant distortion in the model.

G.5 Product-market markup channel

The markup channel captures the possibility that AI increases product-market pricing power, improves price discrimination, or raises desired markups. Two markup proxies are used.

The first proxy is based on unit labor cost:

$$\mu_t^{ULC} = \log(IPDNBS_t) - \log(ULCNFB_t), \tag{G.13}$$

where $IPDNBS_t$ is the nonfarm business output-price deflator and $ULCNFB_t$ is nonfarm business unit labor cost. This proxy measures the part of output prices not explained by labor cost per unit of output. As a closely related check, one can also write the labor-cost component as compensation per unit of output:

$$\log(IPDNBS_t) - \log\left(\frac{COMPNFB_t}{OPHNFB_t}\right), \tag{G.14}$$

where $COMPNFB_t$ is hourly compensation and $OPHNFB_t$ is output per hour. This construction is conceptually close to the unit-labor-cost markup.

The second proxy uses a broader marginal-cost measure that incorporates multiple inputs. Marginal-cost growth is constructed from capital, labor, energy, materials, and services:

$$\Delta \log MC_t^{KLEMS} = \sum_{j\in\{K,L,E,M,S\}} s_t^j \Delta \log P_t^j - \Delta \log A_t. \tag{G.15}$$

The cost shares s_t^j are taken from the BEA–BLS KLEMS production-account data. The baseline broad marginal-cost measure uses time-varying annual KLEMS shares repeated within each quarter of the year. A fixed-share version, based on average input-cost shares over the sample, is also computed as a robustness check.

The broad markup proxy is then constructed as

$$\mu_t^{MC} = \log(IPDNBS_t) - \log(MC_t^{KLEMS}). \tag{G.16}$$

The table reports the version based on time-varying KLEMS shares. The fixed-share version produces very similar estimates and is summarized in the table caption.

For each markup proxy, the markup-channel coefficient is estimated from

$$\mu_t = a_\mu + \phi_\mu z_t + \Gamma_\mu' X_t + u_t^\mu. \tag{G.17}$$

G.6 Constructing the implied supply-side term

For each measurement combination, the implied supply-side AI term is calculated as

$$\widehat{\Delta} = -\widehat{\phi}_A + \widehat{\alpha\phi}_{p_M} + \widehat{\phi}_M + \widehat{\phi}_\mu. \tag{G.18}$$

The table reports point estimates, standard errors, and 90% confidence intervals for each component and for the resulting value of $\widehat{\Delta}$.

The standard error for $\widehat{\Delta}$ is computed by combining the component standard errors:

$$SE(\widehat{\Delta}) = \sqrt{SE(\widehat{\phi}_A)^2 + SE(\widehat{\alpha\phi}_{p_M})^2 + SE(\widehat{\phi}_M)^2 + SE(\widehat{\phi}_\mu)^2}. \tag{G.19}$$

This calculation treats the separately estimated channel components as independent. The resulting values of $\widehat{\Delta}$ are then compared with the direct NKPC estimate Δ_z. The comparison shows whether the aggregate supply-side AI coefficient obtained from the Phillips curve can be rationalized by independently measured productivity, input-price, labor-market, and markup channels.

G.7 Robustness using production-function-based TFP measures

As an additional robustness check, we also construct quarterly production-function-based TFP proxies and use them in place of the productivity measures described above. The purpose of this exercise is to verify whether the decomposition of the AI supply-side term is sensitive to the way productivity is measured. In particular, instead of using labor productivity or the residual productivity term from the marginal-cost accounting exercise, we construct a quarterly TFP proxy from output growth net of a KLEMS-weighted input-growth bundle.

The production-function TFP proxy is constructed as

$$g_t^{TFP} = g_t^Y - \sum_{j\in\{K,L,E,M,S\}} s_{y(t)}^j g_t^j, \tag{G.20}$$

where g_t^Y is quarterly output growth, g_t^j is the quarterly growth rate of input j, and $s_{y(t)}^j$ is the annual KLEMS cost share for the year containing quarter t. The annual cost shares are obtained from the BEA–BLS KLEMS production-account compensation data and repeated within the four quarters of each year. This procedure preserves genuine quarterly variation in output and input quantities while using the KLEMS accounts to discipline the relative importance of capital, labor, energy, materials, and services.

We construct two versions of the quarterly TFP proxy. The first uses nonfarm-business real value-added output as the output measure, while the second uses real GDP. In both cases, capital input is proxied by a perpetual-inventory capital-stock series constructed from real private nonresidential fixed investment. Labor input is measured using nonfarm-business hours. Energy, materials, and services inputs are proxied using quarterly series for energy production, materials excluding energy, and private services-producing gross output. These series are not meant to replace the benchmark productivity measures; rather, they provide a production-function-based robustness check using independently constructed quarterly input and output information.

For each production-function TFP proxy, we estimate the productivity-channel coefficient from the same conditional regression used in the main decomposition:

$$g_t^{TFP} = a_A + \phi_A z_t + \Gamma_A' X_t + u_t^A. \tag{G.21}$$

The resulting estimate of ϕ_A is then inserted into the decomposition

$$\widehat{\Delta} = -\widehat{\phi}_A + \widehat{\alpha\phi}_{p_M} + \widehat{\phi}_M + \widehat{\phi}_\mu, \tag{G.22}$$

leaving the other channel estimates unchanged. Tables G.1 and G.2 report the results. The first table uses the longer-sample AI measure that excludes data centers. The second table uses the AI measure that includes data centers and therefore starts in 2014Q1.

Approach	Statistic	$-\phi_A$	$\alpha\phi_{p_M}$	ϕ_M	ϕ_μ	Δ
PF TFP, business output; ULC markup	Estimate	-0.085	0.011	0.253	0.281	0.461
	SE	0.009	0.001	0.016	0.017	0.025
	90% CI	[-0.100, -0.071]	[0.010, 0.013]	[0.227, 0.280]	[0.254, 0.308]	[0.420, 0.501]
PF TFP, business output; broad MC markup	Estimate	-0.085	0.011	0.253	0.376	0.556
	SE	0.009	0.001	0.016	0.010	0.021
	90% CI	[-0.100, -0.071]	[0.010, 0.013]	[0.227, 0.280]	[0.359, 0.393]	[0.521, 0.591]
PF TFP, real GDP; ULC markup	Estimate	-0.059	0.011	0.253	0.281	0.487
	SE	0.008	0.001	0.016	0.017	0.024
	90% CI	[-0.072, -0.047]	[0.010, 0.013]	[0.227, 0.280]	[0.254, 0.308]	[0.446, 0.527]
PF TFP, real GDP; broad MC markup	Estimate	-0.059	0.011	0.253	0.376	0.582
	SE	0.008	0.001	0.016	0.010	0.021
	90% CI	[-0.072, -0.047]	[0.010, 0.013]	[0.227, 0.280]	[0.359, 0.393]	[0.548, 0.616]

Table G.1: Decomposition of the AI supply-side inflation term using production-function-based TFP measures over the longer sample that excludes data centers from the AI index. The table reports the components of $\Delta = -\phi_A + \alpha\phi_{p_M} + \phi_M + \phi_\mu$, where $-\phi_A$ is the productivity component, $\alpha\phi_{p_M}$ is the AI-input-price component, ϕ_M is the wage-markdown/monopsony component, and ϕ_μ is the product-market markup component. Standard errors are reported below point estimates, and 90% confidence intervals are reported below standard errors. "PF TFP" denotes production-function-based TFP. The business-output version uses nonfarm-business real value-added output, while the real-GDP version uses real GDP. For the broad marginal-cost markup, the table reports the changing-share KLEMS estimate; the corresponding fixed-share estimate is very similar, with $\phi_\mu = 0.379$ (SE 0.010).

Approach	Statistic	$-\phi_A$	$\alpha\phi_{p_M}$	ϕ_M	ϕ_μ	Δ
PF TFP, business output; ULC markup	Estimate	-0.090	0.021	0.450	0.439	0.821
	SE	0.018	0.003	0.037	0.033	0.053
	90% CI	[-0.120, -0.061]	[0.017, 0.026]	[0.390, 0.511]	[0.385, 0.493]	[0.734, 0.907]
PF TFP, business output; broad MC markup	Estimate	-0.090	0.021	0.450	0.287	0.668
	SE	0.018	0.003	0.037	0.030	0.051
	90% CI	[-0.120, -0.061]	[0.017, 0.026]	[0.390, 0.511]	[0.237, 0.336]	[0.585, 0.752]
PF TFP, real GDP; ULC markup	Estimate	-0.108	0.021	0.450	0.439	0.802
	SE	0.016	0.003	0.037	0.033	0.052
	90% CI	[-0.135, -0.082]	[0.017, 0.026]	[0.390, 0.511]	[0.385, 0.493]	[0.717, 0.888]
PF TFP, real GDP; broad MC markup	Estimate	-0.108	0.021	0.450	0.287	0.650
	SE	0.016	0.003	0.037	0.030	0.050
	90% CI	[-0.135, -0.082]	[0.017, 0.026]	[0.390, 0.511]	[0.237, 0.336]	[0.567, 0.733]

Table G.2: Decomposition of the AI supply-side inflation term using production-function-based TFP measures and the AI index that includes data centers over the shorter sample 2014Q1–2025Q3. The table reports the components of $\Delta = -\phi_A + \alpha\phi_{p_M} + \phi_M + \phi_\mu$, where $-\phi_A$ is the productivity component, $\alpha\phi_{p_M}$ is the AI-input-price component, ϕ_M is the wage-markdown/monopsony component, and ϕ_μ is the product-market markup component. Standard errors are reported below point estimates, and 90% confidence intervals are reported below standard errors. "PF TFP" denotes production-function-based TFP. The business-output version uses nonfarm-business real value-added output, while the real-GDP version uses real GDP. For the broad marginal-cost markup, the table reports the changing-share KLEMS estimate; the corresponding fixed-share estimate is very similar, with $\phi_\mu = 0.279$ (SE 0.031). The directly estimated NKPC value over this shorter data-center sample is $\Delta_z = 1.072$.

H A nonlinear two-sector New Keynesian model with AI, robot substitution, monopsony, and endogenous markup dynamics

This section presents a nonlinear two-sector New Keynesian model designed to study how shocks to artificial intelligence (AI) propagate to real activity and, especially, to inflation. The model preserves the core mechanisms emphasized in the one-sector framework, but extends them to an environment with sectoral heterogeneity and relative-price reallocation. In particular, AI affects the economy through productivity, through the relative efficiency of robot services versus labor, through endogenous AI investment, through a bottleneck input whose scarcity can raise costs even when AI also improves efficiency, through firms' wage-setting power in labor markets, and through endogenous desired markups in product markets. The two-sector structure is useful because it allows these mechanisms to operate with different intensities across sectors and introduces an additional reallocation margin through changes in relative prices.

The economy is populated by households and monopolistically competitive firms. Households consume a CES aggregate of sectoral final goods, supply labor, save in one-period nominal bonds, and own the claims associated with robot capital and AI capability in both sectors. Firms in each sector face Calvo price rigidity and therefore adjust prices only intermittently. Monetary policy follows a standard Taylor rule. The resulting equilibrium is a nonlinear system of equations in levels, suitable for quantitative analysis of the inflationary consequences of AI shocks.

Two additional features are particularly important for the paper. First, the labor market is not perfectly competitive. Firms face an upward-sloping labor supply schedule, so the wage they pay affects the quantity of labor they can hire. This gives rise to a monopsony wedge between the wage and the marginal expenditure on labor, and AI can affect that wedge by changing effective labor-market power. Second, desired markups are not imposed as an ad hoc function of AI. Instead, they are generated endogenously through a deep-habits demand structure. AI can then affect pricing power by altering customer retention and the elasticity of demand faced by firms. These two mechanisms are important because they allow AI to affect inflation not only through direct costs and demand pressure, but also through changes in labor-market power and product-market power.

H.1 Households

A representative household chooses aggregate consumption, sectoral labor supply, and bond holdings to maximize expected lifetime utility:

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\frac{(C_t - hC_{t-1})^{1-\sigma}}{1-\sigma} - \frac{\chi}{1+\psi}\left(N_{H,t}^{1+\psi} + N_{L,t}^{1+\psi} \right) \right], \qquad \beta \in (0,1),\ h \in [0,1). \tag{H.1}$$

The parameter h captures external habit formation in aggregate consumption. When $h = 0$, habits are shut down and utility becomes standard CRRA in current consumption. Keeping habits in the baseline is useful because they introduce inertia in aggregate demand and make the dynamic response of inflation more realistic.

Aggregate consumption is a CES composite of the two sectoral final goods:

$$C_t = \left[\omega^{1/\eta} C_{H,t}^{\frac{\eta-1}{\eta}} + (1-\omega)^{1/\eta} C_{L,t}^{\frac{\eta-1}{\eta}} \right]^{\frac{\eta}{\eta-1}}, \qquad \eta > 0, \quad \omega \in (0,1), \tag{H.2}$$

where $C_{H,t}$ and $C_{L,t}$ denote consumption of the goods produced in sectors H and L, respectively. Cost minimization implies the sectoral demand functions

$$C_{H,t} = \omega \left(\frac{P_{H,t}}{P_t}\right)^{-\eta} C_t, \qquad C_{L,t} = (1-\omega)\left(\frac{P_{L,t}}{P_t}\right)^{-\eta} C_t, \tag{H.3}$$

together with the aggregate price index

$$P_t = \left[\omega P_{H,t}^{1-\eta} + (1-\omega) P_{L,t}^{1-\eta}\right]^{\frac{1}{1-\eta}}. \tag{H.4}$$

This CES structure implies that relative price movements redistribute demand across sectors, which adds an additional transmission channel for AI shocks.

The household faces the nominal budget constraint

$$P_t C_t + B_t = W_{H,t} N_{H,t} + W_{L,t} N_{L,t} + \Pi_t^f + \Pi_t^k + R_{t-1} B_{t-1}, \tag{H.5}$$

where B_t denotes one-period nominal bond holdings chosen at time t, R_t is the gross nominal interest rate between t and $t+1$, $W_{i,t}$ is the sectoral nominal wage, Π_t^f denotes profits rebated by firms, and Π_t^k denotes profits or returns associated with ownership of robot capital and AI capability.

The marginal utility of aggregate consumption is

$$U_{C,t} = (C_t - hC_{t-1})^{-\sigma} - \beta h (C_{t+1} - hC_t)^{-\sigma}. \tag{H.6}$$

Letting $\Lambda_t \equiv U_{C,t}/P_t$ denote the marginal utility of nominal income, the household's Euler equation for the one-period bond is

$$1 = \beta \mathbb{E}_t \left[\frac{\Lambda_{t+1}}{\Lambda_t} R_t\right]. \tag{H.7}$$

This condition governs intertemporal substitution in the standard way: higher expected real returns depress current consumption and therefore shift aggregate demand.

Sectoral labor supply satisfies

$$\chi N_{i,t}^{\psi} = \frac{W_{i,t}}{P_t} U_{C,t}, \qquad i \in \{H, L\}. \tag{H.8}$$

This specification is consistent with the quantitative model, in which the household supplies labor separately to the two sectors while the marginal utility of consumption is common across them.

H.2 Labor aggregation and monopsony

To introduce monopsony in a parsimonious but micro-founded way, each sector employs a continuum of differentiated labor services $N_{i,t}(j)$, aggregated through

$$N_{i,t} = \left(\int_0^1 N_{i,t}(j)^{\frac{\varphi_{i,t}-1}{\varphi_{i,t}}} dj\right)^{\frac{\varphi_{i,t}}{\varphi_{i,t}-1}}, \qquad \varphi_{i,t} > 1, \tag{H.9}$$

for $i \in \{H, L\}$. Cost minimization on the part of firms implies demand for labor variety j in sector i:

$$N_{i,t}(j) = \left(\frac{W_{i,t}(j)}{W_{i,t}}\right)^{-\varphi_{i,t}} N_{i,t}, \tag{H.10}$$

where $W_{i,t}(j)$ is the wage paid for type-j labor and $W_{i,t}$ is the corresponding sectoral wage index.

This structure implies that firms in each sector face an upward-sloping labor supply curve. Hiring more labor requires raising the wage on the inframarginal workforce as well. As a result, the relevant object for production decisions is not simply the wage, but the *marginal expenditure on labor*:

$$ME_{i,t}(j) = W_{i,t}(j)\left(1 + \frac{1}{\varphi_{i,t}}\right). \tag{H.11}$$

The associated monopsony wedge is

$$\mu^N_{i,t} \equiv \frac{ME_{i,t}(j)}{W_{i,t}(j)} = 1 + \frac{1}{\varphi_{i,t}}. \tag{H.12}$$

A lower labor-supply elasticity $\varphi_{i,t}$ implies greater labor-market power and a larger gap between the wage and the true marginal labor cost.

AI affects monopsony power through sector-specific labor-supply elasticities:

$$\varphi_{i,t} = \bar{\varphi}_i \exp(-\psi_{\varphi,i} z_{i,t}), \qquad z_{i,t} \equiv \log\left(\frac{Z_{i,t}}{\bar{Z}_i}\right), \qquad i \in \{H, L\}. \tag{H.13}$$

When $\psi_{\varphi,i} > 0$, a rise in AI capability in sector i lowers the effective labor-supply elasticity and raises firms' labor-market power in that sector. Economically, this can be interpreted as AI improving monitoring, scheduling, screening, or matching technologies in ways that reduce workers' outside options and make labor supply to each firm less elastic. The result is a higher marginal expenditure on labor and, potentially, more inflationary pressure even if measured wages do not rise one-for-one.

H.3 Production and AI technology

There is a representative final-good producer in each sector that aggregates a continuum of differentiated intermediate varieties into the sectoral final good. For each sector $i \in \{H, L\}$, the final-good aggregator is standard Dixit–Stiglitz:

$$Y_{i,t} = \left(\int_0^1 Y_{i,t}(j)^{\frac{\varepsilon_i - 1}{\varepsilon_i}} dj\right)^{\frac{\varepsilon_i}{\varepsilon_i - 1}}, \qquad \varepsilon_i > 1. \tag{H.14}$$

In the baseline constant-elasticity case, this implies

$$Y_{i,t}(j) = \left(\frac{P_{i,t}(j)}{P_{i,t}}\right)^{-\varepsilon_i} Y_{i,t}, \tag{H.15}$$

with sectoral price index

$$P_{i,t} = \left(\int_0^1 P_{i,t}(j)^{1-\varepsilon_i} dj\right)^{\frac{1}{1-\varepsilon_i}}. \tag{H.16}$$

As below, this demand system will be modified using deep habits in order to generate endogenous desired markups.

Each intermediate producer j in sector i operates the production technology

$$Y_{i,t}(j) = A_{i,t}\, V_{i,t}(j)^{1-\alpha_i} M_{i,t}(j)^{\alpha_i}, \qquad \alpha_i \in (0,1), \tag{H.17}$$

where $A_{i,t}$ is sector-specific total factor productivity, $V_{i,t}(j)$ is value added, and $M_{i,t}(j)$ is a bottleneck input that captures scarce AI-related infrastructure or intermediate services, such as computing capacity, data-center services, specialized chips, or energy-intensive AI-processing needs.

Value added is itself a CES aggregate of effective labor and robot services:

$$V_{i,t}(j) = \left[(1-\nu_i)\big(\Xi_{i,t}L_{i,t}(j)\big)^{\frac{\rho_i-1}{\rho_i}} + \nu_i\big(\Gamma_{i,t}R_{i,t}(j)\big)^{\frac{\rho_i-1}{\rho_i}}\right]^{\frac{\rho_i}{\rho_i-1}}, \qquad \rho_i > 0. \tag{H.18}$$

Here $L_{i,t}(j)$ is labor employed by firm j in sector i, $R_{i,t}(j)$ is robot services, ν_i is the sector-specific share parameter, and ρ_i is the elasticity of substitution between effective labor and effective robot services. The terms $\Xi_{i,t}$ and $\Gamma_{i,t}$ are labor-augmenting and robot-augmenting efficiency shifters, respectively. A rise in AI capability can therefore increase overall efficiency while also changing the relative productivity of labor and robots, thereby altering the cost-minimizing input mix and the evolution of marginal cost in each sector.

AI capability enters through sector-specific objects:

$$A_{i,t} = \bar{A}_i \exp(\phi_{A,i} z_{i,t}) \exp(\varepsilon^A_{i,t}), \tag{H.19}$$

$$\Gamma_{i,t} = \bar{\Gamma}_i \exp(\phi_{\Gamma,i} z_{i,t}), \qquad \Xi_{i,t} = \bar{\Xi}_i \exp(\phi_{\Xi,i} z_{i,t}), \tag{H.20}$$

where $Z_{i,t}$ is the AI capability stock in sector i.

Unlike the earlier direct-shock formulation, the quantitative model does not treat innovations as exogenous shocks to the AI stock itself. Instead, sectoral AI disturbances are modeled as *persistent investment-efficiency shocks* that operate through endogenous AI investment. This distinction is important. A direct shock to $Z_{i,t}$ would make the sector appear more technologically advanced without requiring current resources to be devoted to adoption. By contrast, an investment-efficiency shock raises the return to AI investment and induces firms to expand AI adoption endogenously, preserving the demand-and-investment channel that is central to the paper.

The bottleneck input has a sector-specific endogenous real price. In sector i, the reduced-form pricing equation is

$$p_{M,i,t} \equiv \frac{P_{M,i,t}}{P_{i,t}} = \bar{p}_{M,i}\left(\frac{D_{M,i,t}}{\bar{D}_{M,i}}\right)^{\upsilon_i}, \qquad \upsilon_i > 0, \tag{H.21}$$

where effective bottleneck demand in sector i is a weighted combination of production use, AI-investment use, and robot-investment use:

$$D_{M,i,t} = M_{i,t} + \omega_{Z,i} I^Z_{i,t} + \omega_{R,i} I^R_{i,t}. \tag{H.22}$$

This specification matches the quantitative model more closely than a formulation based only on production demand for $M_{i,t}$. It allows the scarcity price of AI-related infrastructure to rise not only because current production becomes more bottleneck-intensive, but also because AI and robot deployment directly absorb compute- or infrastructure-intensive inputs.

Cost minimization implies a nominal marginal cost $MC_{i,t}(j)$ for firms in sector i. Since the exact expression is lengthy under the nested technology, it is convenient to summarize the real marginal cost as the unit-cost function associated with (H.17)–(H.18):

$$\frac{MC_{i,t}(j)}{P_{i,t}} = \mathcal{MC}_i\left(\frac{ME_{i,t}(j)}{P_{i,t}},\, \frac{r^R_{i,t}}{P_{i,t}},\, p_{M,i,t},\, A_{i,t},\, \Xi_{i,t},\, \Gamma_{i,t}\right). \tag{H.23}$$

The important point is that the labor-market block enters marginal cost through the monopsony wedge, while the bottleneck block enters through both production needs and investment-related infrastructure demand.

H.4 Robot capital and AI capability accumulation

Robot services in sector i are generated from a sector-specific stock of robot capital, $K^R_{i,t}$, according to

$$R_{i,t} = u^R_{i,t} K^R_{i,t}, \tag{H.24}$$

where $u^R_{i,t}$ is a utilization margin. To keep the model parsimonious, one may set $u^R_{i,t} = 1$ in the quantitative implementation.

Robot capital in each sector evolves according to

$$K^R_{i,t+1} = (1-\delta_{R,i})K^R_{i,t} + I^R_{i,t}\left[1 - \frac{\phi_{R,i}}{2}\left(\frac{I^R_{i,t}}{I^R_{i,t-1}} - 1\right)^2\right]. \tag{H.25}$$

The adjustment-cost term slows down robot adoption and prevents unrealistically abrupt jumps in the capital stock after an AI shock.

AI capability in sector i evolves according to

$$Z_{i,t+1} = (1-\delta_{Z,i})Z_{i,t} + I^Z_{i,t}\left[1 - \frac{\phi_{Z,i}}{2}\left(\frac{I^Z_{i,t}}{I^Z_{i,t-1}} - 1\right)^2\right]. \tag{H.26}$$

This law of motion turns AI into a genuine capital-like stock rather than an exogenous index. Firms or capital owners must devote resources to building AI capability in each sector, which creates an endogenous demand channel: AI investment absorbs output today in exchange for productivity or cost advantages tomorrow.

Consistent with the one-sector model, the exogenous AI shock is not a shock to (H.26) itself. Instead, the disturbance is a *persistent investment-efficiency shock* that shifts the effective incentive to invest in AI. Let $\mathcal{E}^Z_{i,t}$ denote this sector-specific investment-efficiency shifter, with

$$\log \mathcal{E}^Z_{i,t} = \rho_{E,i} \log \mathcal{E}^Z_{i,t-1} + \varepsilon^E_{i,t}, \qquad i \in \{H, L\}. \tag{H.27}$$

A favorable innovation $\varepsilon^E_{i,t}$ raises the payoff to current AI investment and therefore increases sectoral AI adoption endogenously. This is the two-sector analogue of the one-sector investment-efficiency specification and is the appropriate interpretation of the quantitative model's sectoral AI shock processes.

To close this block, introduce sector-specific shadow values $q^R_{i,t}$ and $q^Z_{i,t}$ for robot capital and

AI capability. Their first-order conditions are

$$q_{i,t}^{R} = \beta \mathbb{E}_t \left[\frac{\Lambda_{t+1}}{\Lambda_t} \left(\frac{r_{i,t+1}^{R}}{P_{t+1}} + (1 - \delta_{R,i}) q_{i,t+1}^{R} \right) \right], \tag{H.28}$$

and

$$q_{i,t}^{Z} = \beta \mathbb{E}_t \left[\frac{\Lambda_{t+1}}{\Lambda_t} \left(\frac{\Pi_{i,t+1}^{Z}}{P_{t+1}} + (1 - \delta_{Z,i}) q_{i,t+1}^{Z} \right) \right]. \tag{H.29}$$

Here $\Pi_{i,t}^{Z}$ represents the payoff to AI capability in sector i, which can be interpreted as the marginal contribution of $Z_{i,t}$ to profits through (H.19), (H.20), the sectoral production block, the monopsony block, and the endogenous-markup block introduced below.

The exact investment Euler equations with adjustment costs can be placed in an appendix. In the main text, the key intuition is sufficient: AI and robot accumulation are forward-looking and therefore sensitive to expected future profitability, inflation, and interest rates. The important point for the present paper is that AI shocks operate through the incentive to invest in AI capability, not through an exogenous jump in the stock itself.

H.5 Endogenous markups through deep habits

To introduce endogenous markups without imposing an ad hoc reduced-form markup process, the model adopts a deep-habits demand structure separately within each sector. Consumers form variety-specific consumption habits, so current demand depends not only on current prices and total demand, but also on the stock of past purchases of each variety. This makes the elasticity of demand faced by a firm state dependent and therefore makes desired markups endogenous.

Let $S_{i,t}(j)$ denote the habit stock attached to variety j in sector i:

$$S_{i,t}(j) = (1 - \rho_{S,i}) C_{i,t}(j) + \rho_{S,i} S_{i,t-1}(j), \qquad \rho_{S,i} \in (0,1). \tag{H.30}$$

Consumption of the sectoral final good is then generated by the habit-adjusted aggregator

$$C_{i,t} = \left(\int_0^1 \left[C_{i,t}(j) - \varkappa_{i,t} S_{i,t-1}(j) \right]^{\frac{\varepsilon_i - 1}{\varepsilon_i}} dj \right)^{\frac{\varepsilon_i}{\varepsilon_i - 1}}, \tag{H.31}$$

where $\varkappa_{i,t} \in [0,1)$ governs the strength of customer retention in sector i. When $\varkappa_{i,t} = 0$, the model collapses to the standard Dixit–Stiglitz case with constant elasticity and a constant desired markup in that sector.

The resulting firm-level demand curve is no longer characterized by a constant elasticity. Instead, firms that have accumulated a large habit stock face more captive demand and therefore a lower effective demand elasticity. Denote that effective elasticity in sector i by $\varepsilon_{i,t}^{d}(j)$, with associated desired markup

$$\mu_{i,t}^{p}(j) \equiv \frac{\varepsilon_{i,t}^{d}(j)}{\varepsilon_{i,t}^{d}(j) - 1}. \tag{H.32}$$

The important point is that $\mu_{i,t}^{p}(j)$ is endogenous. It varies over time with the state of demand, habit stocks, and the retention parameter in each sector.

AI affects this product-market channel through the sectoral retention technology:

$$\varkappa_{i,t} = \bar{\varkappa}_i \exp(\psi_{\varkappa,i} z_{i,t}), \tag{H.33}$$

with $\psi_{\varkappa,i} > 0$ implying that greater AI capability in sector i strengthens customer retention, personalization, targeted pricing, or product lock-in. Economically, this means that AI can increase firms' desired markups even without any exogenous markup shock.

H.6 Price setting and inflation

Intermediate firms in each sector are monopolistically competitive and face Calvo price rigidity. In sector i, a firm can reset its price with probability $1-\theta_i$; otherwise, it keeps its price unchanged.

Let $P^*_{i,t}$ denote the common reset price chosen by firms that can reoptimize in sector i. The sectoral price index then evolves according to

$$P_{i,t}^{1-\varepsilon_i} = (1-\theta_i)(P^*_{i,t})^{1-\varepsilon_i} + \theta_i P_{i,t-1}^{1-\varepsilon_i}. \tag{H.34}$$

Sectoral inflation is defined by

$$\Pi_{i,t} \equiv \frac{P_{i,t}}{P_{i,t-1}}, \qquad i \in \{H, L\}, \tag{H.35}$$

and aggregate inflation is the inflation rate associated with the aggregate price index (H.4).

A firm that resets at time t in sector i chooses $P^*_{i,t}$ to maximize the expected discounted value of profits while its price remains in effect:

$$\max_{P^*_{i,t}} \mathbb{E}_t \sum_{k=0}^{\infty} (\beta\theta_i)^k \frac{\Lambda_{t+k}}{\Lambda_t} \left[\left(\frac{P^*_{i,t}}{P_{i,t+k}} \right) Y_{i,t+k|t} - \left(\frac{MC_{i,t+k}}{P_{i,t+k}} \right) Y_{i,t+k|t} \right], \tag{H.36}$$

where

$$Y_{i,t+k|t} = \mathcal{D}_i \left(\frac{P^*_{i,t}}{P_{i,t+k}},\, Y_{i,t+k},\, S_{i,t+k-1},\, \varkappa_{i,t+k} \right) \tag{H.37}$$

is the demand faced by a firm in sector i that last reset at time t.

The associated first-order condition implies a nonlinear reset-price formula of the form

$$P^*_{i,t} = \mu^{p,*}_{i,t} \frac{\mathbb{E}_t \sum_{k=0}^{\infty} (\beta\theta_i)^k \Lambda_{t+k}\, \Omega_{i,t+k|t}\, MC_{i,t+k} Y_{i,t+k|t}}{\mathbb{E}_t \sum_{k=0}^{\infty} (\beta\theta_i)^k \Lambda_{t+k}\, \Omega_{i,t+k|t}\, Y_{i,t+k|t}}, \tag{H.38}$$

where $\mu^{p,*}_{i,t}$ is the endogenous desired markup for a re-optimizing firm in sector i and $\Omega_{i,t+k|t}$ collects the terms induced by the deep-habits demand schedule.

The inflation mechanism in the model can now be stated more completely. AI affects sectoral inflation through sectoral marginal cost and through sectoral desired markups. A positive AI shock may lower marginal cost in a given sector if productivity rises quickly or if robot substitution reduces the effective cost of production. But the same shock may raise inflation if it increases bottleneck scarcity, strengthens labor-market power through monopsony, or raises firms' pricing power through deep habits and endogenous customer retention. Since firms adjust prices only gradually, the timing and persistence of these forces matter for the path of inflation in each sector and, through aggregation, for overall inflation.

H.7 Monetary policy

Monetary policy follows a standard nonlinear Taylor rule in gross terms:

$$\log R_t = (1 - \rho_R) \log \bar{R} + \rho_R \log R_{t-1} + \phi_\pi \log \left(\frac{\Pi_t}{\bar{\Pi}} \right) + \phi_y \log \left(\frac{Y_t}{\bar{Y}} \right) + \varepsilon_t^R. \tag{H.39}$$

The central bank raises the nominal rate when inflation exceeds target or when output rises above its steady-state level.

H.8 Resource constraint and market clearing

The aggregate resource constraint is

$$Y_t = C_t + \sum_{i \in \{H,L\}} \left(I_{i,t}^R + I_{i,t}^Z + \mathcal{AC}_{i,t}^R + \mathcal{AC}_{i,t}^Z \right), \tag{H.40}$$

where the sector-specific adjustment-cost terms are

$$\mathcal{AC}_{i,t}^R = \frac{\phi_{R,i}}{2} \left(\frac{I_{i,t}^R}{I_{i,t-1}^R} - 1 \right)^2 I_{i,t}^R, \qquad \mathcal{AC}_{i,t}^Z = \frac{\phi_{Z,i}}{2} \left(\frac{I_{i,t}^Z}{I_{i,t-1}^Z} - 1 \right)^2 I_{i,t}^Z. \tag{H.41}$$

Sectoral goods-market clearing requires

$$Y_{i,t} = C_{i,t} + I_{i,t}^R + I_{i,t}^Z + \mathcal{AC}_{i,t}^R + \mathcal{AC}_{i,t}^Z, \qquad i \in \{H, L\}. \tag{H.42}$$

Labor market clearing within each sector requires

$$N_{i,t} = \int_0^1 L_{i,t}(j)\, dj, \qquad i \in \{H, L\}, \tag{H.43}$$

and sectoral robot-service market clearing implies

$$R_{i,t} = u_{i,t}^R K_{i,t}^R, \qquad i \in \{H, L\}. \tag{H.44}$$

The sectoral bottleneck-input markets clear at

$$M_{i,t} = \int_0^1 M_{i,t}(j)\, dj, \qquad i \in \{H, L\}. \tag{H.45}$$

These conditions complete the model's real side. They also clarify why AI shocks need not be uniformly disinflationary. Higher AI capability pushes on both sides of the resource constraint: it raises future productive capacity, but it also stimulates current investment in AI and robots. At the same time, it can tighten sector-specific bottleneck markets, alter labor-market power, strengthen product-market power, and reallocate demand across sectors through relative prices.

H.9 AI shocks and the transmission to inflation

The model is built so that an AI shock can be interpreted in a disciplined way. A natural benchmark is a positive innovation to sector-specific AI investment efficiency, (H.27), or to expected sector-specific AI profitability through the payoff to AI capability. Such a shock raises the expected return to AI investment in one or both sectors, stimulates AI accumulation, and through (H.19) and (H.20) improves production possibilities. However, it also raises demand for bottleneck inputs, can alter the composition of factor demand toward robot services, can change firms' effective power in labor markets, can affect pricing power in product markets, and can shift relative prices in ways that reallocate expenditure across sectors.

Six broad forces then determine the inflation response. First, there is a *productivity channel*. Higher $A_{i,t}$, $\Xi_{i,t}$, or $\Gamma_{i,t}$ lowers unit costs and tends to reduce inflation, especially if these gains occur rapidly. Second, there is a *demand-and-investment channel*. Since AI and robot accumulation are endogenous in each sector, a favorable AI shock raises current investment demand. This tends to increase output and can push inflation upward, particularly when monetary policy is gradual. Third, there is a *scarcity-cost channel*. If the AI shock raises demand for the bottleneck input faster than sectoral capacity expands, then (H.21) implies a higher relative price of that input. This raises marginal cost directly and can generate inflation even in the presence of productivity growth. Fourth, there is a *monopsony channel*. If AI lowers the labor-supply elasticity faced by firms, then the marginal expenditure on labor rises relative to the wage. This increases effective labor cost and therefore marginal cost, even if measured wage inflation alone appears muted. Fifth, there is an *endogenous-markup channel*. If AI strengthens retention and lowers the effective elasticity of demand through deep habits, then firms optimally choose higher desired markups. Inflation can then rise even without a proportionate increase in physical production costs. Sixth, there is a *reallocation channel*. Because sectors differ in their exposure to AI, productivity gains, bottleneck pressures, and pricing behavior, AI shocks can change relative prices and shift expenditure across sectors. This redistribution can amplify or attenuate aggregate inflation depending on which sector experiences the stronger cost and markup effects.

The model is therefore well suited to study the empirical question of whether AI has recently acted as a disinflationary force, an inflationary force, or both at different horizons. In the short run, inflation may rise if the demand, scarcity, monopsony, or markup channels dominate, or if reallocation shifts demand toward the sector with stronger inflationary pressure. Over longer horizons, inflation may fall if productivity gains diffuse broadly enough across sectors.

H.10 Equilibrium

Given initial conditions and exogenous shock processes, a competitive equilibrium is a sequence

$$\{C_t, Y_t, P_t, \Pi_t, R_t, B_t, \Lambda_t\}_{t=0}^{\infty}$$

together with sector-specific sequences

$$\left\{Y_{i,t}, P_{i,t}, \Pi_{i,t}, C_{i,t}, N_{i,t}, L_{i,t}, R_{i,t}, K_{i,t}^R, I_{i,t}^R, Z_{i,t}, I_{i,t}^Z, q_{i,t}^R, q_{i,t}^Z, MC_{i,t}, \varkappa_{i,t}, S_{i,t}, \mathcal{E}_{i,t}^Z\right\}_{t=0}^{\infty}, \qquad i \in \{H, L\},$$

such that households maximize utility subject to their budget constraint, labor aggregation and firm-level labor demand satisfy the upward-sloping labor supply schedule and the associated monopsony wedge in each sector, final-good firms in each sector aggregate differentiated varieties competitively, intermediate firms in each sector minimize cost and set prices optimally under Calvo frictions taking into account the state-dependent demand system induced by deep habits, robot capital and AI capability in each sector evolve according to their laws of motion and satisfy their optimality conditions, the monetary policy rule holds, and all goods, labor, bond, robot-service, and bottleneck-input markets clear.

The resulting equilibrium system is nonlinear in levels. Conceptually, it keeps the economic channels transparent: AI affects inflation through marginal cost, expenditure composition, endogenous scarcity, labor-market power, product-market power, and cross-sector reallocation. Quantitatively, it permits the analysis of large shocks and nonlinear transitions, which are especially relevant when studying rapid technological change.

H.10.1 A closed-form expression for the response of inflation to Z_t and the role of ρ

To make the role of the elasticity of substitution between labor and robot services transparent, it is useful to work with a stripped-down log-linear version of the one-sector model. The goal is not to reproduce every feature of the full quantitative system, but rather to derive a closed-form expression that shows explicitly how ρ enters the inflation response. The key simplifying step is to retain the productivity block, the labor-supply block, the New Keynesian pricing block, and the demand side, while shutting down the bottleneck, monopsony, and endogenous-markup wedges. The investment channel is kept in reduced form through its effect on aggregate demand.

The relevant linearized equations are

$$\pi_t = \beta\,\mathbb{E}_t\pi_{t+1} + \kappa\, mc_t, \tag{H.46}$$

$$y_t = \mathbb{E}_t y_{t+1} - \frac{1}{\sigma}\Big(R_t - \mathbb{E}_t\pi_{t+1}\Big) + d\, z_t, \tag{H.47}$$

$$R_t = \phi_\pi \pi_t + \phi_y y_t, \tag{H.48}$$

$$z_t = \rho_z z_{t-1} + \varepsilon_t^z, \qquad |\rho_z| < 1, \tag{H.49}$$

where y_t is output, π_t is inflation, mc_t is real marginal cost, R_t is the nominal interest rate, and z_t is the log deviation of the AI stock from steady state. The coefficient $d > 0$ captures, in reduced form, the demand effect associated with AI adoption and investment. In the full model, this term arises because a higher expected return to AI raises current expenditure on adoption and accumulation. Here it is introduced directly in order to keep the algebra transparent.

The role of ρ comes from the production block. Abstracting from the bottleneck input and writing the value-added block in log-linear form yields

$$y_t = a_t + (1-\nu)(\xi_t + n_t) + \nu(\gamma_t + k_{R,t}), \tag{H.50}$$

where

$$a_t = \phi_A z_t, \qquad \xi_t = \phi_\Xi z_t, \qquad \gamma_t = \phi_\Gamma z_t. \tag{H.51}$$

Thus, AI affects output through neutral productivity, labor-augmenting efficiency, and robot-augmenting efficiency. The robot stock $k_{R,t}$ is treated as predetermined on impact.

The linearized labor-cost condition in the one-sector model, after shutting down monopsony, is

$$w_t = mc_t + y_t + \frac{1-\rho}{\rho} v_t + \frac{\rho - 1}{\rho}\xi_t - \frac{1}{\rho} n_t, \tag{H.52}$$

while labor supply is

$$\varphi n_t = w_t + \lambda_t. \tag{H.53}$$

Since $v_t = y_t - a_t$, one can combine (H.50)–(H.53) and solve for marginal cost as a function of output, the AI state, and predetermined capital. After substitution and simplification,

$$mc_t = a(\rho)\, y_t - \lambda_t - b(\rho)\, z_t - c(\rho)\, k_{R,t}, \tag{H.54}$$

with

$$a(\rho) = \frac{\varphi + 1/\rho}{1-\nu}, \tag{H.55}$$

$$b(\rho) = \frac{\varphi + 1 - \nu + \nu/\rho}{1-\nu}\phi_A + (\varphi+1)\phi_\Xi + \frac{\nu(\varphi + 1/\rho)}{1-\nu}\phi_\Gamma, \tag{H.56}$$

$$c(\rho) = \frac{\nu(\varphi + 1/\rho)}{1-\nu}. \tag{H.57}$$

Equation (H.54) is the key intermediate result. It shows that the elasticity of substitution ρ affects marginal cost through the coefficient $a(\rho)$ multiplying output and through the coefficient $b(\rho)$ multiplying the AI stock. In particular,

$$\frac{\partial a(\rho)}{\partial \rho} = -\frac{1}{(1-\nu)\rho^2} < 0. \tag{H.58}$$

Hence a lower ρ raises the sensitivity of marginal cost to output. This is the formal expression of the idea that when labor and robot services are more complementary, a demand expansion translates more strongly into marginal-cost pressure.

To isolate the effect of ρ on inflation, it is convenient to absorb predetermined terms such as λ_t and $k_{R,t}$ into the state vector and focus on the reduced-form relation

$$mc_t \approx a(\rho)\, y_t - b(\rho)\, z_t. \tag{H.59}$$

The term $-b(\rho)z_t$ is the direct cost-reducing effect of AI operating through productivity and efficiency. The term $a(\rho)y_t$ is the endogenous marginal-cost effect of higher activity. The inflation response will therefore depend on the relative strength of a disinflationary supply effect and an inflationary demand effect.

We now solve the system by the method of undetermined coefficients. Since z_t follows the AR(1) process (H.49), conjecture linear solutions of the form

$$y_t = y_z z_t, \qquad \pi_t = \pi_z z_t. \tag{H.60}$$

It follows that

$$\mathbb{E}_t y_{t+1} = y_z \rho_z z_t, \qquad \mathbb{E}_t \pi_{t+1} = \pi_z \rho_z z_t. \tag{H.61}$$

Substituting (H.60) and (H.61) into the IS equation (H.47) gives

$$y_z z_t = \rho_z y_z z_t - \frac{1}{\sigma}\Big(\phi_\pi \pi_z z_t + \phi_y y_z z_t - \rho_z \pi_z z_t\Big) + d z_t. \tag{H.62}$$

Dividing by z_t and rearranging,

$$D\, y_z = d - F\, \pi_z, \tag{H.63}$$

where, for notational convenience,

$$D \equiv 1 - \rho_z + \frac{\phi_y}{\sigma}, \qquad F \equiv \frac{\phi_\pi - \rho_z}{\sigma}. \tag{H.64}$$

Hence

$$y_z = \frac{d - F\pi_z}{D}. \tag{H.65}$$

Next substitute (H.60), (H.61), and (H.59) into the NKPC (H.46):

$$\pi_z z_t = \beta\rho_z \pi_z z_t + \kappa\Big(a(\rho)y_z z_t - b(\rho)z_t\Big). \tag{H.66}$$

Dividing by z_t gives

$$G\,\pi_z = \kappa\Big(a(\rho)y_z - b(\rho)\Big), \qquad G \equiv 1-\beta\rho_z. \tag{H.67}$$

Substituting (H.65) into (H.67) yields

$$G\pi_z = \kappa\left[a(\rho)\frac{d-F\pi_z}{D} - b(\rho)\right]. \tag{H.68}$$

Collecting the terms in π_z then gives the closed-form coefficient

$$\boxed{\pi_z = \frac{\kappa\,[a(\rho)d - b(\rho)D]}{GD+\kappa a(\rho)F}}. \tag{H.69}$$

Therefore,

$$\boxed{\pi_t = \frac{\kappa\,[a(\rho)d - b(\rho)D]}{GD+\kappa a(\rho)F}\,z_t}. \tag{H.70}$$

Equation (H.70) makes the role of ρ explicit. The elasticity of substitution matters through $a(\rho)$ and $b(\rho)$. The term $b(\rho)$ captures the direct supply-side cost reduction from AI. The term $a(\rho)d$ captures the inflationary effect of AI operating through the demand-and-investment channel. Since $a(\rho)$ is decreasing in ρ, a lower elasticity of substitution raises the sensitivity of marginal cost to output. This means that, for a given demand effect d, lower ρ amplifies the inflationary component of the transmission mechanism.

The sign of the inflation response depends on the numerator in (H.69). Inflation rises if and only if

$$a(\rho)d > b(\rho)D. \tag{H.71}$$

This condition has a natural interpretation. The left-hand side is the demand-driven marginal-cost effect of AI adoption; the right-hand side is the direct cost-reducing supply effect, adjusted for monetary-policy and state persistence through D. When ρ falls, the coefficient $a(\rho)$ rises, so the demand-driven component becomes more powerful. This is precisely the sense in which complementarity between labor and robot services can make AI shocks more inflationary in the short run.

Several limiting cases help interpret the result. If the investment channel is shut down, so that $d=0$, then (H.69) becomes

$$\pi_z^{\text{supply}} = -\frac{\kappa\, b(\rho)\, D}{GD+\kappa a(\rho)F} < 0, \tag{H.72}$$

provided $b(\rho) > 0$. In that case, a higher AI stock acts as a standard supply shock and lowers inflation. By contrast, if the direct productivity effect is weak relative to the demand effect, then $a(\rho)d$ dominates $b(\rho)D$, and inflation rises. In that regime, a lower ρ strengthens the inflation response because it magnifies the marginal-cost consequences of the demand expansion triggered by AI investment.

The closed-form solution therefore clarifies the logic behind the numerical impulse responses. The elasticity of substitution ρ does not matter only because it governs how easily firms replace labor with robots. It also determines how strongly any AI-induced expansion in expenditure is transmitted into marginal cost and inflation. When labor and automation are more complementary, the output term in marginal cost becomes more sensitive, and the inflationary effect of the demand-and-investment channel becomes larger. That is why, in quantitative exercises driven by AI investment-efficiency shocks, lower ρ can generate a stronger short-run increase in inflation even though AI also raises productive efficiency.

H.11 A two-sector model with Endogenous AI

The two-sector model retains the main economic structure of the one-sector framework while introducing sectoral heterogeneity in AI exposure, production structure, pricing, and investment dynamics. In particular, the economy is divided into a high-AI-intensive sector, denoted H, and a low-AI-intensive sector, denoted L. As in the one-sector model, households consume, supply labor, and save in nominal bonds; firms are monopolistically competitive and face Calvo price rigidity; robot capital and AI capability are accumulated endogenously; bottleneck inputs raise production costs when scarce; labor markets feature monopsony; and desired markups vary endogenously through deep habits. The difference is that these mechanisms now operate with different intensities across sectors, and relative-price movements reallocate expenditure between them. The full nonlinear framework is reported in Appendix H. In the main text, it is sufficient to present the key equations that define the additional two-sector structure and that are most relevant for the transmission of AI shocks to inflation.

Households consume a CES aggregate of the two sectoral final goods,

$$C_t = \left[\omega^{1/\eta} C_{H,t}^{\frac{\eta-1}{\eta}} + (1-\omega)^{1/\eta} C_{L,t}^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}}, \qquad \eta > 0, \quad \omega \in (0,1), \tag{H.73}$$

where ω denotes the expenditure weight of the high-AI-intensive sector and η is the elasticity of substitution across sectoral goods. Cost minimization implies the sectoral demand schedules

$$C_{H,t} = \omega \left(\frac{P_{H,t}}{P_t}\right)^{-\eta} C_t, \qquad C_{L,t} = (1-\omega) \left(\frac{P_{L,t}}{P_t}\right)^{-\eta} C_t, \tag{H.74}$$

with the aggregate price index

$$P_t = \left[\omega P_{H,t}^{1-\eta} + (1-\omega) P_{L,t}^{1-\eta}\right]^{\frac{1}{1-\eta}}. \tag{H.75}$$

This aggregation block is important because it implies that sectoral relative-price movements redistribute demand across sectors, creating a reallocation margin that is absent from the one-sector model.

Aggregate inflation is determined by the sectoral inflation rates and the expenditure shares. In the nonlinear model, the sectoral price indices evolve separately, and aggregate inflation is tied to the aggregate price index above. For interpretation, the key relationship is that aggregate inflation is a weighted combination of sectoral inflation outcomes,

$$\Pi_t \equiv \frac{P_t}{P_{t-1}}, \qquad \Pi_{H,t} \equiv \frac{P_{H,t}}{P_{H,t-1}}, \qquad \Pi_{L,t} \equiv \frac{P_{L,t}}{P_{L,t-1}}, \tag{H.76}$$

so that changes in sectoral inflation feed into aggregate inflation according to the evolving sectoral composition of expenditure. Closely related is the sectoral relative price,

$$S_t \equiv \frac{P_{H,t}}{P_{L,t}}, \tag{H.77}$$

which governs expenditure switching between the high- and low-AI-intensive sectors. This relative-price margin becomes quantitatively important whenever AI shocks affect the two sectors asymmetrically.

Production in each sector follows the same basic logic as in the one-sector model, but with sector-specific parameters. For each sector $i \in \{H, L\}$,

$$Y_{i,t} = A_{i,t} \, V_{i,t}^{1-\alpha_i} M_{i,t}^{\alpha_i}, \tag{H.78}$$

where $A_{i,t}$ is sector-specific total factor productivity, $V_{i,t}$ is value added, and $M_{i,t}$ is the bottleneck input. Value added is produced by combining effective labor and robot services through a CES technology,

$$V_{i,t} = \left[(1-\nu_i)\big(\Xi_{i,t} N_{i,t}\big)^{\frac{\rho_i - 1}{\rho_i}} + \nu_i \big(\Gamma_{i,t} R_{i,t}\big)^{\frac{\rho_i - 1}{\rho_i}} \right]^{\frac{\rho_i}{\rho_i - 1}}. \tag{H.79}$$

As in the one-sector environment, AI capability affects production through three margins: a neutral productivity component, a robot-augmenting component, and a labor-augmenting component. Sectoral heterogeneity enters because these elasticities can differ between H and L, allowing AI to raise productivity more strongly in the high-AI-intensive sector than in the low-AI-intensive one.

A central difference relative to a simpler multisector setup is that AI remains an endogenous capital-like stock in each sector. Sector-specific AI capability evolves according to

$$Z_{i,t+1} = (1-\delta_{Z,i}) Z_{i,t} + I_{i,t}^Z \left[1 - \frac{\phi_{Z,i}}{2} \left(\frac{I_{i,t}^Z}{I_{i,t-1}^Z} - 1 \right)^2 \right], \qquad i \in \{H, L\}, \tag{H.80}$$

so that current AI investment raises future sectoral AI capability subject to adjustment costs. The key shocks in the quantitative analysis are sector-specific AI investment-efficiency shocks, which make investment in AI more productive in one sector or the other. This preserves the demand-and-investment channel that is central to the paper: a favorable AI shock raises the incentive to adopt AI, stimulates current investment spending, and only gradually builds future productive capacity. Robot capital evolves analogously at the sectoral level, so the economy features joint transitions in AI capability, robot capital, and production structure.

Nominal rigidity is modeled separately within each sector. Intermediate firms in sector i face Calvo price adjustment, so sectoral inflation is determined by a sector-specific New Keynesian Phillips mechanism in which inflation depends on expected future inflation, real marginal cost, and the desired markup. Since both marginal cost and markups respond endogenously to AI through bottlenecks, monopsony, and deep habits, sectoral inflation can react differently across sectors even after a common aggregate shock. Aggregate inflation then reflects both these sectoral inflation responses and the reallocation of expenditure across sectors.

The two-sector structure therefore adds one important layer to the one-sector model. In the one-sector model, AI affects inflation through productivity, investment demand, bottleneck scarcity, labor-market power, and product-market power. In the two-sector model, all of these channels remain active, but they now operate with different strengths across sectors and are supplemented by a reallocation channel. A positive AI shock concentrated in the high-AI-intensive sector, for example, can raise investment demand and bottleneck scarcity there more strongly than in the low-AI-intensive sector, while also changing the relative price of sectoral goods and shifting expenditure across sectors. Aggregate inflation is then shaped not only by within-sector cost and markup pressures, but also by the endogenous redistribution of demand toward the sector experiencing the stronger inflationary or disinflationary effects. This makes the two-sector framework a useful extension for studying how uneven AI adoption across industries translates into aggregate inflation dynamics.

H.11.1 Parameterization

The two-sector model is designed as a disciplined extension of the one-sector benchmark rather than an independent calibration exercise. Sector-specific parameters are chosen so that their expenditure-weighted averages (with $\omega_H = 0.20$) reproduce the one-sector calibration, ensuring that the aggregate implications of the two-sector model are comparable to the benchmark. All preference, nominal-friction, capital accumulation, and monetary-policy parameters are held fixed at their one-sector values: $\beta = 0.99$, $\sigma = 1.50$, $h = 0.65$, $\chi = 2.40$, $\varphi = 1.00$, $\theta = 0.733$, $\varepsilon = 6.0$, $\delta_R = 0.025$, $\delta_Z = 0.015$, $\phi_R = 4.0$, $\phi_Z = 5.0$, and the Taylor-rule coefficients $\rho_R = 0.65$, $\phi_\pi = 0.875/(1 - \rho_R) = 2.50$, and $\phi_y = 0.04374/(1 - \rho_R) = 0.125$. The persistence of the sector-specific AI investment-efficiency shocks is set to $\rho_{\varepsilon Z} = 0.786$ in both sectors, matching the one-sector benchmark. The expenditure weight assigned to the high-AI-intensive sector is $\omega_H = 0.20$, capturing an economy in which the most AI-intensive activities remain a minority share of aggregate expenditure while being large enough to affect aggregate inflation dynamics. The Armington elasticity of substitution between the two sectoral composites in the final-good aggregator is set to $\eta_s = 2.00$, implying moderate expenditure switching in response to relative-price movements.

The output elasticity of the AI-infrastructure bottleneck input is $\alpha_H = 0.14$ in the high-AI sector and $\alpha_L = 0.09$ in the low-AI sector, with a weighted average of 0.10 as in the one-sector model. The robot-services share in the CES value-added nest is $\nu_H = 0.36$ and $\nu_L = 0.285$, yielding a weighted average of exactly 0.30, consistent with the US capital income share once utilisation is accounted for (**?**). The elasticity of substitution between effective labor and robot services is $\rho_H = 0.90$ in the high-AI sector and $\rho_L = 1.025$ in the low-AI sector, with a weighted average of 1.00. The high-AI sector operates in the gross-complements regime ($\rho_H < 1$): an increase in robot efficiency raises the marginal product of effective labor rather than displacing it, consistent with capital-labor substitution elasticities in the range 0.5–0.7 reported by Chirinko (2008) and Oberfield and Raval (2021) for capital-intensive industries. The low-AI sector exhibits mild gross substitutability ($\rho_L > 1$), reflecting routine-task-intensive industries in which automation more directly replaces workers, consistent with the evidence in **?** that labor-share displacement is concentrated in industries undergoing rapid automation. The weighted average of $\rho = 1.00$ corresponds to Cobb-Douglas and is consistent with the broad employment neutrality of automation documented at the aggregate level.

The three AI-augmentation elasticities are calibrated jointly in each sector so that the AI investment Euler equation holds at the non-stochastic steady state. In the high-AI sector, $(\phi_{A,H}, \phi_{\Gamma,H}, \phi_{\Xi,H}) = (0.03867, 0.03095, 0.00774)$; in the low-AI sector, $(\phi_{A,L}, \phi_{\Gamma,L}, \phi_{\Xi,L}) = (0.02256, 0.01805, 0.00452)$. Across both sectors the ratio $\phi_\Gamma : \phi_A \approx 0.80$ is maintained, preserving the empirical finding that AI most directly amplifies machine-process effectiveness rather than neutral total factor productivity (Brynjolfsson et al. (2017)). The labor-augmenting elasticity ϕ_Ξ is set to approximately one-fifth of ϕ_A in each sector, reflecting that AI raises worker productivity more slowly than it amplifies automation capital.

The steady-state real price of the AI-infrastructure input is $\bar{p}_{M,H} = 0.060$ in the high-AI sector and $\bar{p}_{M,L} = 0.0475$ in the low-AI sector, with a weighted average of 0.050 matching the one-sector benchmark. The scarcity elasticity of the bottleneck price is $\upsilon_H = 0.80$ and $\upsilon_L = 0.55$: the steeper supply curve in the high-AI sector reflects tighter semiconductor and data-centre capacity constraints in AI-intensive industries. Both sectors normalise steady-state bottleneck capacity to unity ($\bar{M}_{\text{ss},H} = \bar{M}_{\text{ss},L} = 1.0$), so that $p_M = \bar{p}_M$ exactly at the non-stochastic steady state and deviations are interpretable as pure scarcity premia. The AI- and robot-investment scale factors are $(\omega_{Z,H}, \omega_{R,H}) = (1.20, 1.10)$ in the high-AI sector and $(\omega_{Z,L}, \omega_{R,L}) = (0.95, 0.975)$ in the low-AI sector, with weighted averages of unity by construction.

The steady-state labor-supply elasticity faced by the representative firm is $\bar{\varphi}_{w,H} = 3.00$ in the high-AI sector and $\bar{\varphi}_{w,L} = 4.25$ in the low-AI sector. The implied steady-state monopsony wedges are $\mu^N_H = 1 + 1/\bar{\varphi}_{w,H} = 1.33$ and $\mu^N_L = 1 + 1/\bar{\varphi}_{w,L} = 1.24$, so the high-AI sector exhibits greater monopsony power, reflecting higher employer concentration and platform-like labor-market dynamics in AI-intensive industries Dube et al. (2020) The weighted average of $\bar{\varphi}_w = 4.00$ reproduces the one-sector benchmark. The AI sensitivity of the labor-supply elasticity is $\psi_{\theta,H} = 0.40$ and $\psi_{\theta,L} = 0.275$, with a weighted average of 0.30 matching the one-sector value.

The steady-state customer-retention coefficient is $\bar{\kappa}_H = 0.50$ in the high-AI sector and $\bar{\kappa}_L = 0.375$ in the low-AI sector. These imply steady-state effective demand elasticities of $\varepsilon_{d,H} = \varepsilon(1 - \bar{\kappa}_H) = 3.0$ and $\varepsilon_{d,L} = \varepsilon(1 - \bar{\kappa}_L) = 3.75$, and corresponding desired markups of $\mu^p_H = 1.50$ and $\mu^p_L = 1.36$. The weighted average of $\bar{\kappa} = 0.40$ is within the range $[0.30, 0.60]$ calibrated by **?** to match the cyclicality of markups and the slope of the empirical Phillips curve. The AI sensitivity of customer retention is $\psi_{\kappa,H} = 0.40$ and $\psi_{\kappa,L} = 0.275$, with a weighted average of 0.30, consistent with the one-sector parameterization. maximising markup variation at business-cycle frequencies.

H.11.2 Impulse responses in the two-sector model

Figures H.1 and H.2 show the impulse responses of the two-sector model to a one-standard-deviation AI investment-efficiency shock hitting, respectively, the high-AI-intensive and the low-AI-intensive sector. Each figure compares the restricted specification — in which only the productivity and capital-accumulation channels are active — to the full model that also allows bottleneck scarcity, monopsony, and endogenous markups to respond. Taken together, the responses confirm the central message of the paper: once AI adoption is embedded in a richer dynamic general-equilibrium environment, inflation is not governed solely by the direct productivity effects of the technology.

Shock to the high-AI sector. In the restricted specification, high-AI output (Y_H) rises sharply on impact — reaching approximately 0.25 percent above baseline within the first two quarters — and remains positive for roughly fifteen quarters before reverting to trend (Figure H.1). High-AI inflation (π_H) is muted throughout, peaking at only around 0.05 percent and decaying quickly, consistent with a clean productivity boom that lowers marginal cost without triggering significant price pressure. Low-AI output (Y_L) responds modestly and negatively on impact, reflecting the initial crowding-out of investment from the unshocked sector, visible in AI investment in the low-AI sector (I^Z_L) falling by up to -2 percent in the first few quarters. Tobin's q for AI in the high-AI sector (q^Z_H) falls by about -7 percent on impact under the restricted model, reflecting the standard asset-price response to a positive investment-efficiency shock. The relative price (s) is essentially flat, indicating that in the absence of the additional channels there is little expenditure reallocation between sectors.

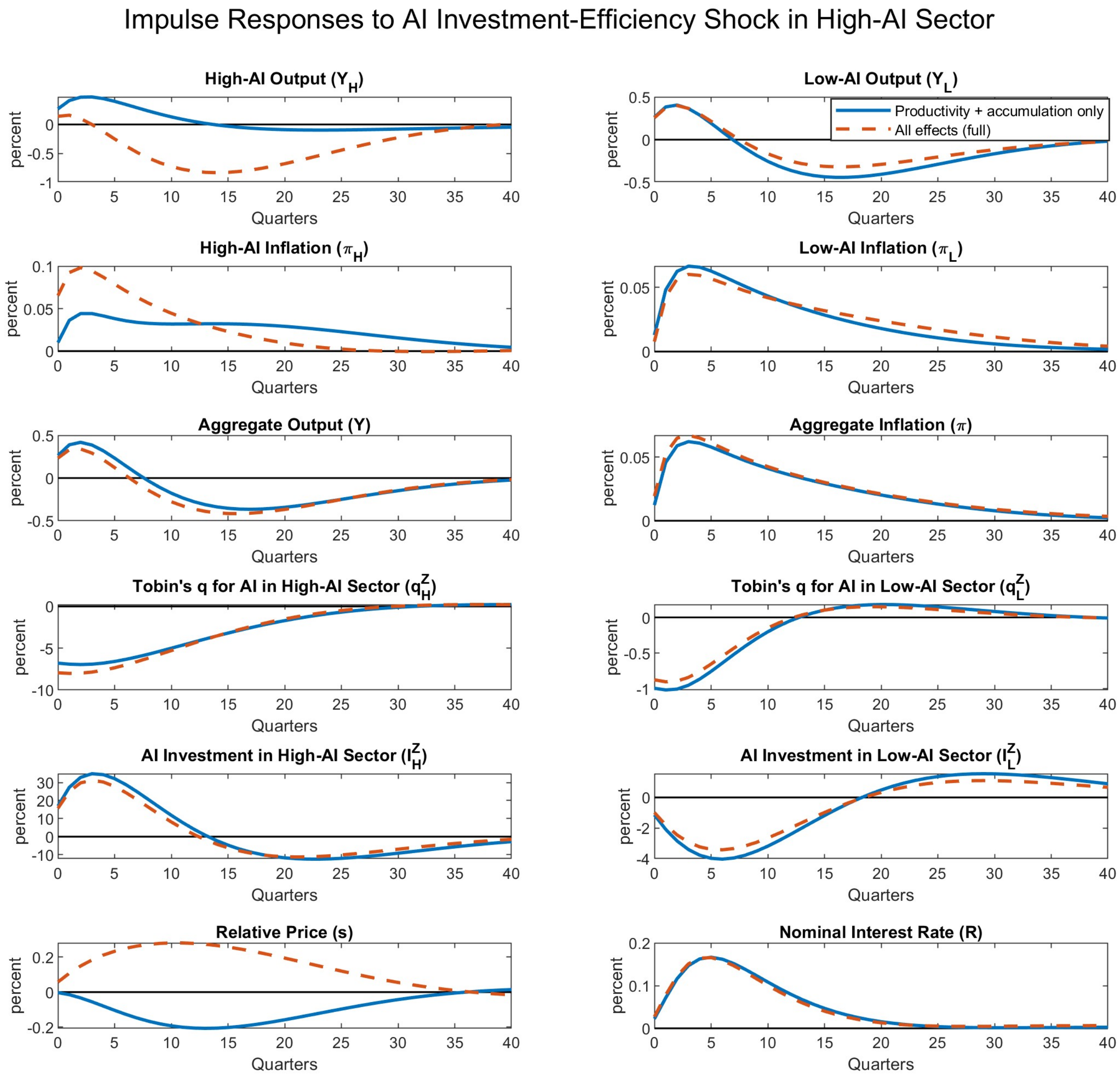


Figure H.1: Impulse responses in the two-sector DSGE model following an AI investment-efficiency shock in the high AI-intensive sector.

Once the full set of channels is activated, the picture changes substantially. High-AI output is visibly weaker at all horizons and turns negative by around quarter five, falling to nearly -0.5 percent by quarter ten before recovering — a pattern that is absent in the restricted case. High-AI inflation rises more than twice as strongly on impact, reaching approximately 0.10 percent, and is considerably more persistent, remaining above zero for over twenty quarters. This divergence indicates that the inflationary forces introduced by the full model — bottleneck scarcity driving up $p_{M,H}$, rising monopsony power compressing effective labor supply, and strengthening customer-retention markups through the deep-habits channel — more than offset the direct cost-reducing effect of higher AI efficiency in the sector most exposed to the shock.

Low-AI inflation (π_L) also rises noticeably in the full model relative to the restricted case, despite the shock originating elsewhere, reflecting the spillover of higher aggregate marginal cost through the common consumption and monetary-policy block. The relative price (s) now moves appreciably upward under the full model — rising to approximately 0.2 percent at its peak — consistent with a reallocation of expenditure toward the low-AI sector as the high-AI sector faces higher costs. This relative-price adjustment is absent in the restricted model, demonstrating that the additional channels are the source of the expenditure-switching dynamics. The nominal interest rate (R) rises modestly under the full model as the central bank responds to the higher aggregate inflation, whereas the restricted model produces only a negligible monetary-policy response.

Shock to the low-AI sector. When the efficiency shock hits the low-AI sector, the qualitative results are broadly symmetric but quantitatively different, reflecting the lower AI intensity and weaker bottleneck and markup channels in that sector (Figure H.2). Under the restricted model, low-AI output (Y_L) rises on impact to around 0.75 percent and decays smoothly, while low-AI inflation (π_L) peaks at approximately 0.15 percent — somewhat larger than the high-AI sectoral inflation response to the equivalent shock in Figure H.1, consistent with the higher output response in the less constrained sector. AI investment in the low-AI sector (I_L^Z) surges by approximately 25 percent on impact under the restricted model and decays over roughly ten quarters, while AI investment in the high-AI sector (I_H^Z) falls by about 5 percent, reflecting the cross-sector crowding-out of investment demand.

In the full model, the low-AI output response is more muted at intermediate horizons and somewhat more negative at longer horizons relative to the restricted case, while low-AI inflation peaks higher and is more persistent, consistent with the bottleneck and markup channels adding inflationary pressure even in the less AI-intensive sector. High-AI output (Y_H) falls below the restricted-model path in the full specification, turning negative around quarter three and reaching approximately -0.75 percent at its trough — a more pronounced negative spill-in than was visible for the low-AI sector when the shock originated in the high-AI sector. High-AI inflation (π_H) rises more strongly in the full model, peaking at roughly 0.2 percent, well above the 0.05 percent peak in the restricted specification. This cross-sectoral inflation transmission reflects the general-equilibrium interactions created by common monetary policy, common consumption, and relative-price adjustment. The relative price (s) moves in the opposite direction to the high-AI shock case — falling on impact and turning negative by around quarter three under the full model — indicating that the low-AI shock raises the relative price of the low-AI good, shifting expenditure toward the high-AI sector and thereby transmitting cost pressure upward into the more AI-intensive block. The nominal interest rate response is somewhat larger and more persistent than in the high-AI shock case, consistent with the larger peak inflation response at the aggregate level when the shock originates in the broader, higher-weight sector.

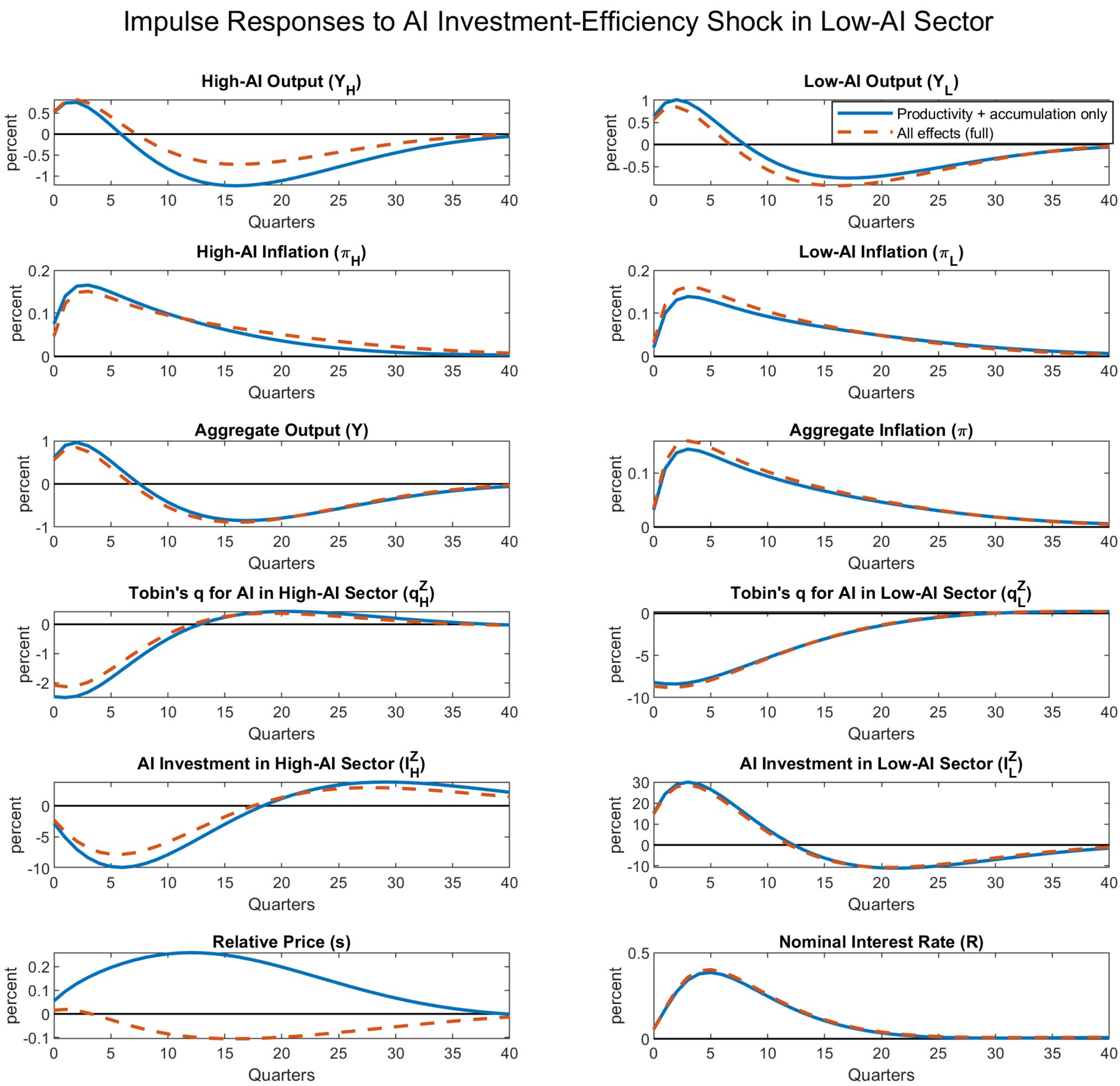


Figure H.2: Impulse responses in the two-sector DSGE model following an AI investment-efficiency shock in the high AI-intensive sector.

The two-sector impulse responses deliver two complementary findings. First, they provide robustness for the one-sector model by confirming that the main qualitative message, namely AI shocks need not be disinflationary on impact once additional channels beyond pure productivity are active, survives in a richer environment that allows for heterogeneous sectoral exposure and relative-price adjustment. Second, the two-sector structure sharpens the interpretation of that result by revealing that aggregate inflation can mask substantial sectoral heterogeneity: the sector directly hit by the shock and the sector affected through spillovers can exhibit meaningfully different inflation dynamics, and the relative-price adjustment is a key transmission mechanism that the one-sector model cannot show explicitly. The crowding-out of investment across sectors and the resulting asymmetric bottleneck and markup responses are the proximate source of this heterogeneity, and they operate more powerfully the larger the sectoral asymmetries in AI intensity, bottleneck exposure, monopsony, and customer-retention pricing.